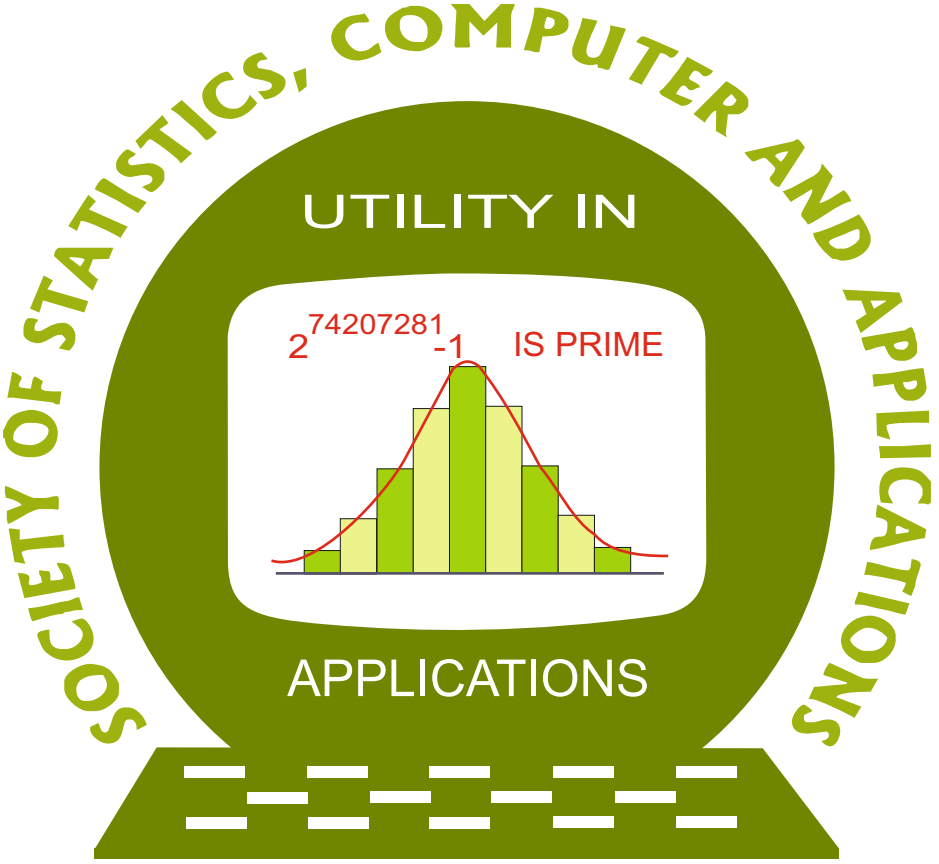
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Volume 20, No. 1, 2022 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

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

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury

Council Members

Anil Kumar Yadav, B. Re. Victor Babu, Manish Sharma, Manisha Pal, Piyush Kant Rai
Praggya Das, Rajni Jain, Rakhi Singh, Ranjit Kumar Paul, Raosaheb V. Latpate
S.A. Mir, Spam Sobita, V. Srinivasa Rao, V.M. Chacko, Vishal Deo

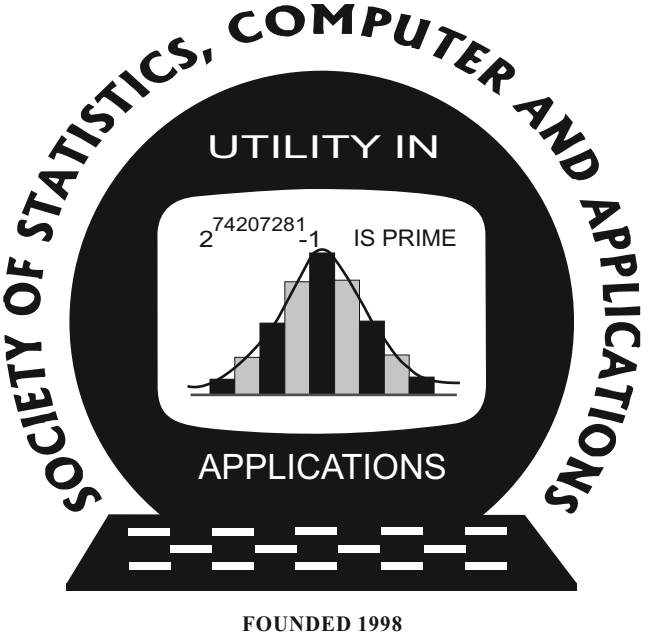
Ex-Officio Members (By Designation)

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

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

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

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

Volume 20, No. 1, 2022 (New Series)

Statistics and Applications

Volume 20, No. 1, 2022 (New Series)

Nripes Kumar Mandal, Flat No. 5, 141/2B, South Sinthee Road, Kolkata-700050; mandalnk2001@yahoo.co.in
P. Venkatesan, Professor Computational Biology SRIHER, Chennai, Adviser, CMRF, Chennai; venkaticmr@gmail.com
Ramana V. Davuluri, Department of Biomedical Informatics, Stony Brook University School of Medicine, Health Science Center Level 3, Room 043 Stony Brook, NY 11794-8322, USA; ramana.davuluri@stonybrookmedicine.edu; ramana.davuluri@gmail.com
S. Ejaz Ahmed, Faculty of Mathematics and Science, Mathematics and Statistics, Brock University, ON L2S 3A1, Canada; sahmed5@brocku.ca
Sanjay Chaudhuri, Department of Statistics and Applied Probability, National University of Singapore, Singapore -117546; stasc@nus.edu.sg
Sat N. Gupta, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC -27412, USA; sngupta@uncg.edu
Saumyadipta Pyne, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
T.V. Ramanathan; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
Tapio Nummi, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
Tathagata Bandyopadhyay, Indian Institute of Management Ahmedabad, Gujarat; tathagata.bandyopadhyay@gmail.com, tathagata@iima.ac.in
Tirupati Rao Padi, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
V. Ramasubramanian, ICAR-IASRI, Library Avenue, PUSA, New Delhi – 110012; ram.vaidhyanathan@gmail.com

Corrigendum

With apology, we wish to share that in the following volumes and issues of Statistics and Applications, the ISSN Number mentioned on the title page of each individual paper is 2452-7395 (online). The correct ISSN Number, though, is 2454-7395 (online). The cover page of these issues does mention the correct ISSN Number as 2454-7295 (online). It is requested to please read the ISSN Number on each individual paper in these issues as 2454-7395 (online) in place of 2452-73975 (online)

1. Vol. 18, No. 2 (Special Issue), 2020 (New Series)
2. Vol. 18, No. 1, 2020 (New Series)
3. Vol. 17, No. 2, 2019 (New Series)
4. Vol. 17, No. 1, 2019 (New Series)
5. Vol. 16, No. 1, 2018 (New Series)
6. Vol. 15, Nos. 1 & 2, 2017 (New Series)
7. Vol. 14, Nos. 1 & 2, 2016 (New Series)

CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 20, No. 1, 2022 (New Series)

Statistics and Applications {ISSN 2454-7395 (Online)}
Volume 20, No. 1, 2022 (New Series), pp 1-15

Markov-Switching GARCH and Mixture of GARCH-type Models for Accuracy in Forecasting

Godfrey Joseph Saqware[1] and Ismail B.[2]
[1]*Deparment of Statistics, Mangalore University, Karnataka, India*
[2]*Department of Statistics, Yenepoya* (*Deemed to be University*), *Mangalore, Karnataka, India*

Received:15 May 2020; Revised: 07 January 2021; Accepted: 18 January 2021

Abstract

The stock markets all over the world have been experiencing fluctuations. These fluctuations are due to some political and administrative decisions. For example, in Tanzania, structural transformations in the economic sectors have been happening time after time, which resulted in fluctuations in the stock market. In this paper, the stock market's volatility was modelled using Markov-Switching GARCH (MS GARCH) and the mixture of GARCH type models. The Bayesian Information Criterion (BIC) was employed to get the best GARCH type models with respective conditional distributions. The GARCH (1, 1) with skewed normal distribution, EGARCH (1, 1) with student's *t*-distribution and Glosten, Jagannathan and Runkle-GARCH (GJR GARCH) (1, 1) with generalized error distribution selected for further analysis. The study found that the three-state heterogeneous regime MS GARCH and Mixture of the selected GARCH type models provide the best fit and the dynamic feedback between components for the DSEI All-share stock data. The Bayesian Markov Chain Monte Carlo (MCMC) method resulted in an acceptance rate of 28.7%, which lies between 20% and 50% as the requirement of the rule of thumb. The different sample sizes employed on the Bayesian MCMC technique have also proven the fitted model's powerfulness since all acceptance sampler rate falls within the range. Furthermore, the forecasting results for the next 30, 60, 90, and 120 days have shown a continuous fluctuation in the DSEI All-share Stock Index.

Key words: MS GARCH; GARCH; GJR GARCH; Bayesian MCMC; DES.

1. Introduction

The global economy has been experiencing fluctuations in response to policy directives. Stock Market performance is also affected by the economic and other related instabilities. Stochastic models play essential roles in the forecasting stock market volatility. The famous symmetric models such as Autoregressive Heteroscedasticity (ARCH) (Engle, 1982) and Generalized ARCH (Bollerslev, 1986); and asymmetric models namely Exponential GARCH (Nelson, 1991), Threshold GARCH (Glosten *et al*., 1993), GARCH-M (Hamilton, 1994) and Fractionally Integrated Generalized Autoregressive Conditional Heteroscedastic (FIGARCH) (Baillie *et al*.,1996) were extended from Generalized ARCH model to capture asymmetric characteristics in the stock market. The complexity and uncertainty of the financial time series have resulted in the continuous modification of the GARCH-type model. To handle volatility prediction in the stock market. The best way forward to the question is to allow the GARCH model parameters to vary over time by considering the regime-switching. A single regime is

Corresponding Author: Godfrey Joseph Saqware
Email: godjose70@yahoo.com

inflexible; thus, it's crucial to incorporate regime-switching. The Markov-Switching GARCH (MS GARCH) model is a new approach introduced in more than one decade. It enables a quick adaption to the unconditional volatility variations (Oseifua and Korkpoe, 2018).

This paper will be the first in incorporating the heterogeneous regimes-switching model to the DSEI All-share Index log-returns. The annotated article is distributed into four major sections. The first section covers the introduction; the second section will cover the literature reviews on modelling volatility based on Markov-Switching GARCH models. Part three will formally lay out the materials and methods employed in the study. Section four covers empirical analysis and discussion. Finally, section five concludes the paper.

2. Related Work

The MS GARCH models' history goes back to introducing the mixed normal distribution that was combined with the GARCH-type structure (MN-GARCH) that captures conditional variance and the dynamic feedback between the components(Haas et al., 2004). Recent studies have shown that volatility predictions using GARCH type models failed to capture the stock market volatility's actual variation due to the regime changes and volatility dynamics (Korkpoe and Kawor, 2018). The MS GARCH models are flexible alternatives to GARCH models with fixed parameters. Bayesian inference estimate based on data augmentation has solved the path dependence problem. Furthermore, the model is useful for capturing changes in the dynamics and volatilities in the financial market (Bauwens *et al*., 2014). Based on this perspective, the effective and efficient prediction of the market volatility has been crucial for smooth economic growth.

Moreover, the era of fast-growing technology and computer applications resulted in gaps in the modelling and forecasting volatility. The MS GARCH models with regime-switching have shown the best forecasting performances based on the management perspective compared to forecasting based on a single regime (Ardia *et al*., 2016). The MS GARCH model provides a better evaluation of volatility by imposing the higher volatility component in each state, which results in the dynamic structure regime that reacts to the various species of shocks (Alemohammad *et al*., 2016). The MS GARCH with the two-regime has exhibited the best in-sample performance with an inverted leverage effect in low and high volatility regimes and their volatility dynamics (Ardia *et al*., 2019). The regime-switching models revealed a better volatility forecast than the constant-variance or a single-regime GARCH (Bibi and Ghezal, 2018). Thus, the earlier researchers have tried to model volatility without defining clearly the process of obtaining conditional distributions. The study involves selecting the conditional distribution and applying the three-state heterogeneous MS GARCH and the Mixture of GARCH-type models to the stock data.

3. Materials and Methods

3.1. The Markov-switching GARCH models

The method allows the regime-switching in the conditional variance process. I*f* $\mathrm{I}_{t-1} \equiv \{r_{t-1},\, i > 0\}$ is the information set denoted by I_{t-1} for the observation up to $t-1$. The general Markov Switching specification is given by

$$r_t \setminus (s_t = k, \mathrm{I}_{t-1}) \sim D(0, h_{t,k}, \xi_k) \tag{1}$$

where $D(0, h_{t,k}, \xi_k)$ refers to a continuous distribution with mean zero, $h_{t,k}$ is the time-varying variance and ξ_k is the additional shape parameter and k is the number of regimes. The Stochastic variable s_t defined under the discrete space $\{1,2,...,K\}$ characterizes the Markov-Switching GARCH Models.

3.1.1. Markov-switching ARCH model

The ARCH model (Engle, 1982) that incorporates k the regime can be written as:-

$$h_{k,t} = \alpha_{0,k} + \alpha_{1,k} r_{t-1}^2 \; ; k = 1,2,3,...,K \tag{2}$$

The $\alpha_{0,k}, \alpha_{1,k}, \beta_k \geq 0$ is required for the positivity while in each regime for the covariance-stationarity $\alpha_{1,k} < 1$.

3.1.2. Markov-switching GARCH model

The GARCH model (Bollerslev, 1986) that incorporates the k regimes Markov-Switching is given by

$$h_{k,t} = \alpha_{0,k} + \alpha_{1,k} r_{t-1}^2 + \beta_k h_{k,t-1}; k = 1,2,3,...,K \tag{3}$$

The $\alpha_{0,k}, \alpha_{1,k}, \beta_k \geq 0$ is required for the positivity while in each regime for the covariance-stationarity $\alpha_{1,k} + \beta_k < 1$.

3.1.3. Markov-switching EGARCH model

The Exponential GARCH Model (Nelson, 1991) that incorporates the k regimes is given by

$$h_{k,t} = \alpha_{0,k} + \alpha_{1,k}\left(\left|\eta_{k,t-1}\right| - E\left(\left|\eta_{k,t-1}\right|\right)\right) + \alpha_{2,k}\eta_{k,t-1} + \beta_k \ln(h_{k,t-1}) \; ; \quad k = 1,2,3,...,K \tag{4}$$

The model considers the leverage effects, where the past negative values influence conditional volatility compared to the previous positive values. The covariance stationary in each regime to be achieved requires $\beta_k < 1$.

3.1.4. Markov-switching GJR GARCH model

The GJR GARCH model (Glosten *et al*., 1993) captures as well the asymmetric conditional volatility. The GJR GARCH model that incorporates the k regimes Markov-Switching is expressed by

$$h_{k,t} = \alpha_{0,k} + \left(\alpha_{1,k} + \alpha_{2,k}\tau\left\{r_{t-1} < 0\right\}\right) r_{t-1}^2 + \beta_k h_{k,t-1} \;;\; k = 1,2,3,...,K \tag{5}$$

To ensure positivity, $\alpha_{0,k}, \alpha_{1,k} > 0$ and $\alpha_{2,k}, \beta_k \geq 0$, whereas for the covariance stationarity

$$\alpha_{1,k} + \alpha_{2,k} E\left[\eta_{k,t}^2 \tau\left\{\eta_{k,t} < 0\right\}\right] + \beta_k < 1 \tag{6}$$

3.1.5. Markov-switching TGARCH model

The Threshold GARCH model of Zakoian (1994) included conditional volatility as the dependent variable instead of the conditional variance. The model in (7) incorporates Markov-Switching.

$$h_{k,t}^{1/2} \equiv \alpha_{0,k} + \left(\alpha_{1,k}\tau\left\{r_{t-1} \geq 0\right\} - \alpha_{2,k}\tau\left\{r_{t-1} < 0\right\}\right) r_{t-1} + \beta_k h_{k,t-1}^{1/2} \quad ;k = 1,2,3,...,K \tag{7}$$

To ensure positivity, $\alpha_{0,k}, \alpha_{1,k} > 0$; $\alpha_{2,k} > 0$ and $\beta_k \geq 0$, while for the covariance stationarity

$$\alpha_{1,k}^2 + \beta_k^2 - 2\beta_k\left(\alpha_{1,k} + \alpha_{2,k}\right) E\left[\eta_{k,t}\tau\left\{\eta_{k,t} < 0\right\}\right] - \left(\alpha_{1,k}^2 - \alpha_{2,k}^2\right) E\left[\eta_{k,t}^2 \tau\left\{\eta_{k,t} < 0\right\}\right] < 1 \tag{8}$$

3.2. Conditional distributions

The specification of the model to be completed requires conditional distributions. The commonly used conditional distributions are Normal distribution, Student's *t*-distribution and the generalized error distribution.

Normal distribution: The probability distribution function for the normal distribution is given by

$$f_N(r) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\eta^2}, \qquad \eta \in \sim \tag{9}$$

Student's *t* distribution: The probability distribution function for the Student's *t*-distribution is given by

$$f_S(\eta;V) \equiv \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{(v+1)\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{\eta^2}{(v-2)}\right)^{-\left(\frac{v+1}{2}\right)}, \qquad \eta \in \sim \tag{10}$$

The $\Gamma(\bullet)$ is the Gamma function, and $v > 2$ for the existence of the second moment.

Generalized error distribution: The probability distribution function for the generalized error distribution (GED) is given by

$$f_{GED}(\eta;V)\equiv\frac{Ve^{-\frac{1}{2}\left|\frac{\eta}{\lambda}\right|^{v}}}{\lambda 2^{\left(1+\frac{1}{v}\right)}\Gamma\left(1/v\right)},\quad \lambda\equiv\left(\frac{\Gamma\left(1/v\right)}{4^{1/v}\Gamma\left(3/v\right)}\right)^{1/2},\quad \eta\in\sim\ \ and\ \ v>0 \tag{11}$$

Skewed distributions: Recently, the unimodal standardized distributions introduced skewness in estimating the EGARCH, GJR GARCH, and TGARCH models (Trottier and Ardia, 2016).

3.3. Model estimation

The estimation of the MS GARCH and the Mixture of the GARCH type models can either based on the Bayesian Markov Monte Carlo (MCMC) or Maximum Likelihood (ML) methods. The two methods require evaluation under the maximum likelihood function.

3.3.1. The maximum likelihood method

Let $\psi=\left(\theta_1,\xi_1,...,\theta_K,\xi_K,\mathrm{P}\right)$ be the vector of the model parameters whose likelihood function is given by

$$L\left(\psi\,|\,\mathrm{I}_T\right)\equiv\prod_{t=1}^{T}f\left(r_t\,|\,\psi,\mathrm{I}_{t-1}\right) \tag{12}$$

where $f\left(r_t\,|\,\psi,\mathrm{I}_{t-1}\right)$ refers to the density function of r_t given the past observations, I_{t-1} is the information set and ψ the model parameters.

The MS GARCH model for the conditional density of r_t is given by

$$f\left(r_t\,|\,\psi,\mathrm{I}_{t-1}\right)=\sum_{i=1}^{K}\sum_{j=1}^{K}p_{i,j}z_{i,t-1}f_D\left(r_t\,|\,s_t=j,\psi,\mathrm{I}_{t-1}\right) \tag{13}$$

where $z_{i,t-1}\equiv P\left(s_{t-1}=i\,|\,\psi,\mathrm{I}_{t-1}\right)$ refers to the filter probability of state i and time $t-1$.

Moreover, for the Mixture of the GARCH type models, the conditional density function for r_t becomes

$$f\left(r_t\,|\,\psi,\mathrm{I}_{t-1}\right)=\sum_{i=1}^{K}w_j f_D\left(r_t\,|\,s_t=j,\psi,\mathrm{I}_{t-1}\right) \tag{14}$$

Combining the two, the conditional density of the r_t in-state or component $s_t=k$ given ψ and I_{t-1} is denoted by $f_D\left(r_t\,|\,s_t=k,\psi,\mathrm{I}_{t-1}\right)$.

3.3.2. Bayesian Markov Monte Carlo (MCMC) method

The MCMC estimation requires a combination of the likelihood with a truncated prior $f(\psi)$ to building kernel of the posterior distribution $f\left(\psi|\mathrm{I}_T\right)$. The unknown form of posterior distribution should be approximated based on the simulation techniques. The random-walk Metropolis sampler generates draws of the posterior distribution (Vihola, 2012).

Furthermore, for the Bayesian estimation, the likelihood function is combined with a prior $f(\psi)$ in building kernel of the posterior distribution $f\left(\psi|\mathrm{I}_T\right)$. The build prior based on the independent diffuse priors is done as follows

$$
\begin{aligned}
&f(\boldsymbol{\psi}) = f\left(\boldsymbol{\theta}_1,\xi_1\right)\ldots f\left(\boldsymbol{\theta}_K,\xi_K\right) f(\mathbf{P}) \\
&f\left(\boldsymbol{\theta}_k,\boldsymbol{\xi}_k\right) \propto f\left(\boldsymbol{\theta}_k\right) f\left(\boldsymbol{\xi}_k\right) \mathrm{I}\left\{\left(\boldsymbol{\theta}_k,\boldsymbol{\xi}_k\right) \in CSC_k\right\} \quad \left(k=1,2,3,\ldots,K\right) \\
&f\left(\boldsymbol{\theta}_k\right) \propto f_N\left(\boldsymbol{\theta}_k;\boldsymbol{\mu}\boldsymbol{\theta}_k,\ diag\left(\boldsymbol{\sigma}^{\mathbf{2}}\boldsymbol{\theta}_{\mathbf{k}}\right)\right) \mathrm{I}\left\{\xi_{k,1} \in PC_k\right\} \quad \left(k=1,2,3,\ldots,K\right) \\
&f\left(\boldsymbol{\xi}_k\right) \propto f_N\left(\boldsymbol{\xi}_k;\mu\boldsymbol{\xi}_k,\ diag\left(\sigma^2\boldsymbol{\xi}_k\right)\right) \mathrm{I}\left\{\xi_{k,1}>0,\ \xi_{k,2}>2\right\} \quad \left(k=1,2,3,\ldots,K\right) \\
&f(\mathbf{P}) \propto \prod_{i=1}^{K}\left(\prod_{j=1}^{K} p_{i,j}\right) \mathrm{I}\left\{0<p_{i,i}<1\right\}
\end{aligned}
\tag{15}
$$

where CSC_k and PC_k denotes the covariance stationarity and the positivity conditions for the regime k, respectively.

3.4. Data specification

The website hosted at https://www.investing.com/indices/tanzania-all-share is the source of the data for this investigation. The data contain information such as stock day open, low, high and close prices. The dataset ranged from 15/08/2009 to 20/1/2020 with a total of 2067 observations. Tanzania DSEI All-share Index has the market capitalization-weighted index with 1000 base reference. The index includes all stocks listed at the Dar Es Salaam Stock Exchange categorized into Commercial Banks, Cement Companies, Tanzania Breweries companies, Tanzania Cigarette Company and Liquefied Natural Gas and Oil companies.

4. Empirical Results and Discussions

4.1. Descriptive statistics for the log-returns

The summary statistics reported in Table 1 below shows that the mean log-return is the positive value of 0.0004 and a standard deviation of 0.0175. The most significant price drop is –32.09%, and the largest price increase is 32.81%. Data exhibit a positive skewness and a very large excess kurtosis. The Jarque-Bera test for normality has shown that the data is not normally distributed ($p < 0.05$). The suitable distributions for this kind of data are the skewed conditional distributions.

Table 1: Summary statistics

Statistic	Mean	Min	Max	SD	Skewness	Kurtosis	JB	JB(*p*-value)
Value	0.0004	–0.321	0.3281	0.0175	0.3847	168.563	3647640	2.20E-16

The computations of the log-returns for the DSEI All-share closing price is given by

$r_t = \log P_t - \log P_{t-1}$, where r_t is the daily log-return, while P_t *and* P_{t-1} are the stock prices for time t *and* $t-1$ respectively.

4.2. Time series of the DSEI all-share stock index

The sharp decline in the DSEI All-share stock Index at different periods was a result of various factors. The drop observed almost every year since DSEI All-share Stock Market started its operation in August 2011. The Central Bank of Tanzania (BOT) merged some banks because of bankruptcy in 2018/2019. Moreover, the closure and liquidation non-performing banks aimed at stabilizing the banking system. The current President of the United Republic of Tanzania Hon. Dr John Pombe Magufuli has tried to support the economy; nevertheless, some companies failed to survive since he came into power in October 2015. The log-returns exhibited a continuous and frequent period of high and low volatility since the stock market started operation in August 2011. Figure 1 below shows a time series plot for DSEI All-share Stock Index.

Figure 1: Time series plot of the DSEI series

The plots of the DSEI All-share Index log-returns have also revealed a presence of volatility clustering. The prolonged stock instability observed between 2013 and 2018. The fluctuation has resulted in the collapse and closure of different companies and merging of the key players in the DSEI All-share Stock Market. Figure 2 below shows the plotting of the log-returns series for the DSEI All-share Index.

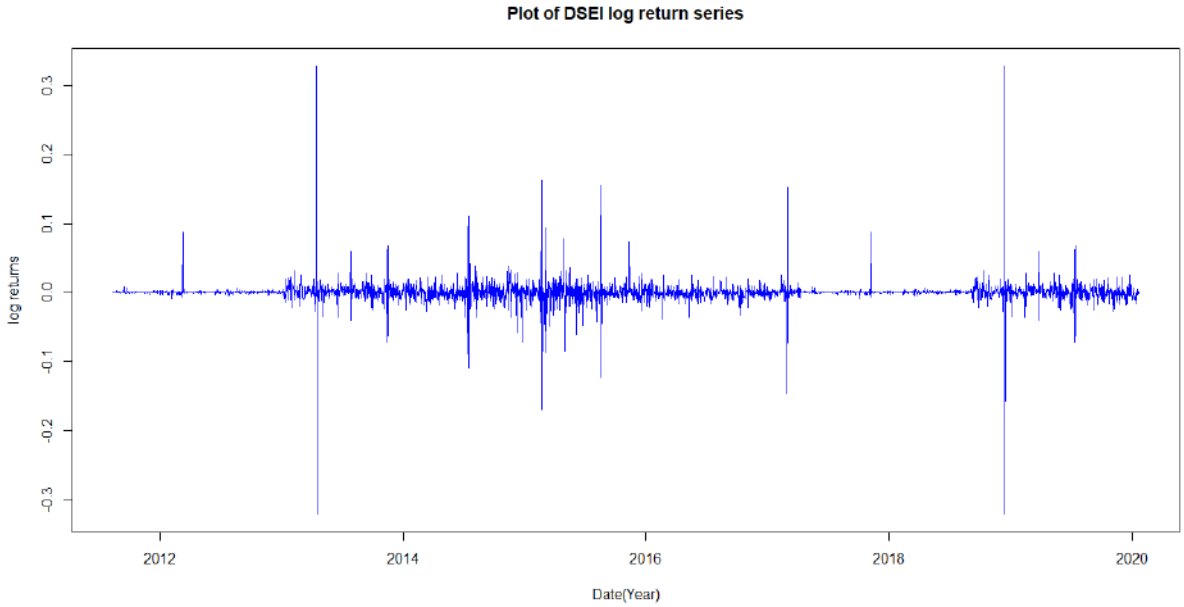


Figure 2: Plot of the DSEI log-return series

The Augmented Dickey-Fuller test confirmed that the return series is stationary since the p-value is less than 5% (p-value = 0.01). Moreover, the GARCH model building has been done by first confirming the (G) ARCH effects in the stock data. The ARCH-LM test gave a

χ^2 = 1097.3 with 10 degrees of freedom and p-value less than 5%. Therefore, we conclude the presence of (G) ARCH in the log-return series.

4.3. Model estimation and selection

This paper extended the study by Haas (2004) with the Markov-Switching GARCH (MSGARCH) and mixture GARCH type model, incorporating three heterogeneous state regimes and conditional distributions. The selected GARCH type models based on the minimum Bayesian Information Criterion (BIC) are; GARCH (1, 1) with the skewed student-t-distribution (sstd), EGARCH (1, 1) with the skewed generalized distribution (sged) and GJR GARCH (1, 1) with the skewed generalized distribution (sged). The study by (Catania et al., 2018) has come up with regimes without telling where they made the GARCH-type models selection with the conditional distributions. Still, in this paper, we made some initial effort before proceeding with model estimation. Table 2 below shows the result of the model selection based on the BIC.

Table 2: BIC values for conditional distributions

Model	GARCH (1,1)			EGARCH (1,1)			GJR GARCH (1,1)		
Distribution	snorm	sstd	sged	snorm	sstd	sged	snorm	sstd	sged
BIC	–6.3832	**–7.8578**	–4.1269	–6.974	–7.9305	**–8.2764**	–6.3403	–7.8551	**–8.097**

4.3.1. Model estimation based on maximum likelihood (ML) method

The estimated parameters depict the difference in the volatility process from one regime to another. The difference in negative past reactions levels of unconditional volatility of $\alpha_{(2,1)} \approx 0.00$, $\alpha_{(2,2)} \approx -0.0148$ and $\alpha_{(3,2)} \approx 0.7213$ for the three-state heterogeneous regimes. The volatility persistence for the model reports $\alpha_{(1,1)} + \frac{1}{2}\alpha_{(2,1)} + \beta_1 \approx 0.904$, $\alpha_{(1,2)} + \frac{1}{2}\alpha_{(2,2)} + \beta_2 \approx 1.0866$ and $\alpha_{(1,3)} + \frac{1}{2}\alpha_{(2,3)} + \beta_3 \approx 0.9990$ in three-states, respectively. The result implies that the first regime characterized by low unconditional volatility, a strong volatility reaction to the past negative log-returns, and the low volatility process persistence. The second and third regimes are characterized by high unconditional volatility, weak volatility reaction to the past negative log-returns, and high volatility. The market participants can categorize regime one as "tranquil market condition" compared to regimes two and three, which has the "turbulent market condition". Table 3 shows the estimated model summary based on the ML technique.

Table 3: Estimated model summary for ML technique

K	Parameter	Estimate	*t* value	*P*-value
1	$\alpha_{(0,1)}$	0.0000	1.0000E+8	<1e-16
	$\alpha_{(1,1)}$	0.1092	5.43394E+21	<1e-16
	β_1	0.7948	6.91272E+22	<1e-16
	nu_1	3.7560	1.26907E+22	<1e-16
	xi_1	0.9784	6.05698E+21	<1e-16
2	$\alpha_{(0,2)}$	–0.4566	–3.67276E+20	<1e-16
	$\alpha_{(1,2)}$	0.1421	3.03187E+20	<1e-16
	$\alpha_{(2,2)}$	–0.0148	–9.8235E+19	<1e-16
	β_2	0.9519	3.19745E+22	<1e-16
	nu_2	1.2262	3.75161E+23	<1e-16
	xi_2	0.9855	1.41481E+22	<1e-16
3	$\alpha_{(0,3)}$	0.000	3.05331E+21	<1e-16
	$\alpha_{(1,3)}$	0.2527	2.74911E+22	<1e-16
	$\alpha_{(2,3)}$	0.7213	3.26518E+24	<1e-16
	β_3	0.3856	3.64538E+26	<1e-16
	nu_3	0.7000	7.4042E+36	<1e-16
	xi_3	0.9998	1.0202E+8	<1e-16

The stable probabilities of being in the three states are about 32.15%, 63.97%, and 3.88% respectively. The results indicate that the likelihood of being in the three states differs. The unconditional probabilities reports; 4.57%, 40.41% and 20.97 for state 1, state 2, and state 3 respectively. Thus, this implies a high unconditional probability in state two compared to the rest of the states. Moreover, all the three states' smooth probabilities are closer to one; this evidence a sharp increase in the volatility process. Table 4 below shows the stable probabilities, unconditional volatility, and smooth probabilities for the three states.

Table 4: Results for the three states probabilities and unconditional volatility

State	1	2	3
Stable Probabilities	0.3215	0.6397	0.0388
Unconditional Volatility	0.0457	0.4041	0.2097
Smooth Probability	0.9675	0.9958	0.9994

4.3.2. Model estimation based on Bayesian Markov Chain Monte Carlo (MCMC) method

The estimation model based on Maximum likelihood seems to be not more powerful than the Bayesian MCMC method due to a stuck in the local maximum which may result in unreliable estimates (Billio and Cavicchioli, 2017; Das and Yoo, 2004). The adaptive MCMC estimation always based on the posterior distribution. The MCMC sampler requires decomposition and the Eigenvalue computations, therefore, it largely depends on the Linear Algebra library (Vihola, 2012). The proposed three-state heterogeneous regimes MS GARCH and the Mixture of GARCH type models such as GARCH, EGARCH and GJR GARCH model with skewed normal, Student's *t* and generalized error conditional distributions respectively, has used 10000 iterations, 5000 burn-in phase and ten thinning factors in the estimation of the Bayesian Markov Monte Carlo (MCMC) estimation. The acceptance rate of 28.5% was obtained in the model estimation. The acceptance rate lied within 20%-50% 'rule of thumb' as recommended (Chib and Greenberg, 1995; Roberts and Rosenthal, 2009).

The number of independent and identically distributed sample draws from the posterior distribution is required for the relative numerical efficiency (RNE). This determines how quickly the convergence of the algorithm occurs. The checking of the MCMC sampling scheme for the output quality is necessary (Geweke, 1992; Korkpoe and Kawor, 2018). In the proposed model, we found the values of RNE relatively low (<1), which are considered better for the fast convergence of the MCMC chains. Table 5 below shows the summary of parameter estimated for the three states heterogeneous regimes for the MS GARCH and Mixture of GARCH type models using the Bayesian MCMC method.

The increase of the number of the MCMC draws say 15000, 20000, 30000, 50000, 100000, 500000 and 1000000 for the estimation three state heterogeneous regimes MS GARCH and the Mixture of GARCH type models resulted into the same range of the acceptance rate of (20-50) %. The acceptance rate reveals the consistency of the estimated model. The best model is usually based on the minimum Deviance Information Criterion (Spiegelhalter *et al*., 2002). Moreover, at least 4000 burn-in phase is recommended for the model estimation (Raftery and Lewis, 1992). The thin of every tenth minimizes the posterior draws autocorrelations. The high autocorrelations can result in bias and Monte Carlo standard errors. The number of researchers has raised concern on the appropriate number of thinning, but the thinning number of 10L sounds good (Link and Eaton, 2012; Owen, 2017). Table 6 shows the estimated model summary for the different MCMC sample draws.

Table 5: Estimated model summary for the Bayesian MCMC technique

K	Parameter	Mean	SD	SE	TSSE	RNE
1	$\alpha_{(1,1)}$	0.3509	0.0325	0.001	0.0086	0.0142
	β_1	0.5768	0.024	0.0008	0.0065	0.0138
	nu_1	2.2511	0.0431	0.0014	0.0153	0.0079
	xi_1	0.992	0.0325	0.001	0.003	0.1180
2	$\alpha_{(0,2)}$	–0.0351	0.0427	0.0013	0.0223	0.0036
	$\alpha_{(1,2)}$	0.4177	0.0391	0.0012	0.0056	0.0496
	$\alpha_{(2,2)}$	–0.0983	0.0201	0.0006	0.0019	0.1128
	β_2	0.9955	0.0042	0.0001	0.0024	0.0031
	nu_2	0.7337	0.0273	0.0009	0.0174	0.0025
	xi_2	0.9999	0.0005	0.0000	0.0000	0.3027
3	$\alpha_{(1,3)}$	0.1737	0.0461	0.0015	0.0260	0.0031
	$\alpha_{(2,3)}$	0.0021	0.0018	0.0001	0.0008	0.0048
	β_3	0.7816	0.0194	0.0006	0.0100	0.0038
	nu_3	0.7000	0.0000	0.0000	0.0000	0.0033
	xi_3	14.1021	6.3244	0.200	3.9098	0.0026

Table 6: Estimated model summary for different MCMC sample draws

nithin	nburn	nmcmc	Acceptance Rate	DIC
10	5000	15000	27.6%	–22609.639
10	5000	20000	28.3%	–10986.542
10	5000	25000	27.5%	–23339.681
10	5000	30000	28.0%	–21846.993
10	5000	50000	27.4%	–23881.170
10	5000	100000	27.2%	–22595.525
10	5000	500000	26.1%	–19464.122
10	5000	1000000	25.7%	–11562.360

4.3.3. Forecasting of the conditional volatility based on Bayesian MCMC estimated model

The prediction based on 30, 60, 90 and 120 days ahead has shown fluctuations in the DSEI All-share Index log-returns. The study identified the number of future periods of high conditional volatility; March for the 60 days forecasts in (b), April for the 90 days forecasts in (c), and May for the 120 days forecasts in (d) for four months in the year 2020. The results still show instability in the stock market for the next four months. Figure 3 below shows stock volatility for the next four months.

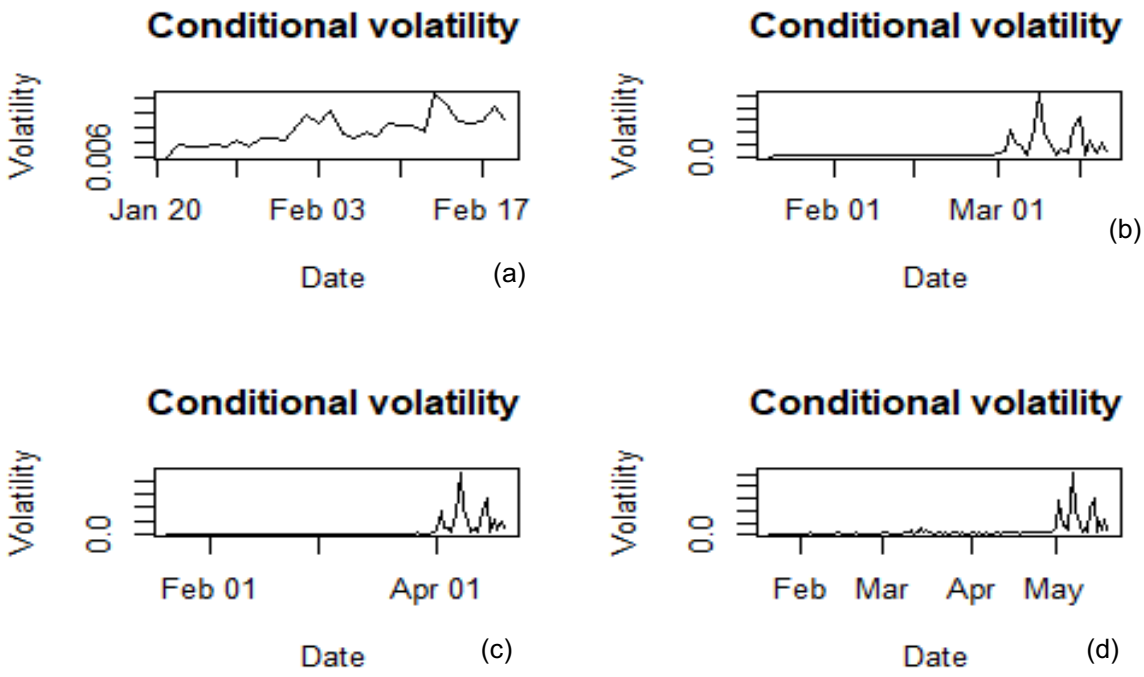


Figure 3: Conditional volatility forecasting

4. Conclusion

The stock market volatility will continue to be topical in finance since traders and investors observe historical data trends for future investments. The insertion of the regime changes become indispensable to model volatility in the stock market. The changing economic condition has caused persistent fluctuations in the stock market across regimes. The study acts as a benchmark for the countries to adopt the best trading policies and strategies to buffer downside. The Central Bank of Tanzania (BOT) reported a decline of shares, trading, market capitalization and underperformance of the Dar Es Salaam Stock Exchange (BOT, 2018).

Moreover, stockbrokers like the five social security funds were joined into Public Service Social Security Fund (PSSSF) that serves public-sector employees and the National Social Security Fund (NSSF) for the private-sector employees and self-employed persons in 2018/2019. The situation has disturbed the performance of the stock market. Eventually, the

country's business environment's control and regulation become inevitable for the stock markets' growth.

References

Alemohammad, N., Rezakhah, S. and Alizadeh, S. H. (2016). Markov switching component GARCH model: Stability and forecasting. *Communications in Statistics-Theory and Methods*, **45**(**15**), 4332–4348.

Ardia, D., Bluteau, K., Boudt, K., Catania, L. and Trottier, D. A. (2016). Markov-switching GARCH models in R: The MS GARCH package. *Journal of Statistical Software*, **91**(**4**), 1-38.

Ardia, D., Bluteau, K. and Rüede, M. (2019). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, **29**, 266–271. https://doi.org/10.1016/j.frl.2018.08.009

Baillie, R. T., Bollerslev, T. and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **74**(**1**), 3–30.

Bauwens, L., Dufays, A. and Rombouts, J. V. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*, **178**, 508–522.

Bibi, A. and Ghezal, A. (2018). Markov-Switching BILINEAR- GARCH models: Structure and estimation. *Communications in Statistics-Theory and Methods*, **47**(**2**), 307–323.

Billio, M. and Cavicchioli, M. (2017). Markov Switching GARCH Models: Filtering, Approximations and Duality. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance* (pp. 59–72). Springer.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**(**3**), 307–327.

BOT (2018). *Tanzania Financial Stability Report*. Dar es Salaam, Tanzania.

Catania, L., Ardia, D., Bluteau, K., Boudt, K. and Trottier, D. A. (2018). Markov–switching GARCH Models in R: The MS GARCH Package. *Journal of Statistical Software*.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, **49**(**4**), 327–335.

Das, D. and Yoo, B. H. (2004). A Bayesian MCMC algorithm for Markov switching

GARCH models. *Econometric Society* 2004 *Far Eastern Meetings* 451.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, **50**(**4**), 987–1007.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, **4**, 641–649.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, **48**(**5**), 1779–1801.

Haas, M. (2004). Mixed Normal Conditional Heteroskedasticity. *Journal of Financial Econometrics*, **2**(**2**), 211–250. https://doi.org/10.1093/jjfinec/nbh009

Haas, Markus, Mittnik, S. and Paolella, M. S. (2004). Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics*, **2**(**2**), 211–250.

Hamilton, J. D. (1994). *Time Series Analysis* (Vol. 2). Princeton University Press, Princeton, NJ.

Korkpoe, C. H. and Kawor, S. (2018). A Rollercoaster Ride through the Equity Markets—Evidence from the Ghana Stock Exchange. *African Finance Journal*, **7**(**3**), 1-21.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, **3**(**1**), 112–115.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, **59**(**2**), 347–370.

Oseifuah, E. K. and Korkpoe, C. H. (2018). Regime changes in the South African Rand exchange rate against the Dollar. *Academy of Accounting and Financial Studies* Journal, **22**(**3**), 13.

Owen, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *Journal of Computational and Graphical Statistics*, **26**(**3**), 738–744.

Raftery, A. E. and Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: Comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Statistical Science*, **7**(**4**), 493–497.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**(**2**), 349–367.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), **64**(**4**), 583–639.

Trottier, D. A. and Ardia, D. (2016). Moments of standardized Fernandez–Steel skewed distributions: Applications to the estimation of GARCH-type models. *Finance Research Letters*, **18**, 311–316.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, **22(5)**, 997–1008.

Zakoian, J. M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, **18(5)**, 931-955.

Statistics and Applications {ISSN 2454-7395(online)}
Volume 20, No. 1, 2022 (New Series), pp 17–31

Adjusted Design Effect Model for Individual Variables in Survey Data

Mohammed Quazi and Yan Lu
Department of Mathematics and Statistics
University of New Mexico, Albuquerque, New Mexico, USA

Received: 25 July 2020; Revised: 06 December 2020; Accepted: 28 January 2021

Abstract

In this study we have extended longitudinal generalized variance functions (LGVF) for grouped variables variance estimation to individual variables variance estimation. Effect of survey design and change of population over time are incorporated into modeling to estimate variance of a survey statistic. Six such LGVF models are studied and results produced span over 15 years of Current Population Survey's (CPS) March Supplement data from socio-economic category. In addition to this, variables grouped together are also studied. 18 binary variables are considered. Simulation shows that individual variable variance estimation outperforms grouped variable variance estimation.

Key words: Longitudinal generalized variance function; Adjusted design effect model; Individual and grouped variable variance estimation.

MSC: 62D05

1. Introduction

Cost and labor involved in computing the estimated variances for thousands of estimates every year could be saved if computation could be simplified using generalized variance functions (GVF). This paper extends the results of Zhang, Cheng and Lu (2019) to individual variables variance estimation for a large-scale, complex survey data. Individual variables variance estimation will help to narrow the focus to only the variables of interest in survey data whereas grouped estimation brings in unwanted variability into the model and makes it harder to estimate individual parameters with high accuracy. Literature review in the area shows that the relative variance (relvar) of a survey statistic is a function of the population total. This idea is supported by Johnson and King (1987), Valliant (1987) and McIllece (2016). As for the GVF, Wolter (2007) discusses the application of GVF to estimate variance of a survey statistic. This method of variance estimation has been in use for Current Population Survey (CPS) data by US Census Bureau as well. Sampling error of GVF estimators for Current Employment Survey (CES) is evaluated by Cho, Eltinge, Gershunskaya and Huff (2002).

Corresponding Author: Mohammed Quazi
Email: mquazi@unm.edu

We use the March Supplement CPS data for the state of New Mexico (NM) from the socio-economic category from year 2003 to 2017. Figure 1 shows the sampling scheme used in this study. New Mexico households from same neighborhood in the data are divided into Ultimate Sampling Units (USUs) where each USU contains four households. We believe households in the same neighborhood have similar economic statuses. 100 USUs are randomly picked from each year's data. Four individuals from each USU are then sampled using probability proportional to size (PPS) method. The simulation process is explained in detail in section 4. Unemployment compensation, education benefits and disability payments are three of the eighteen variables studied – all the variables are shown in table 1 on page 23.

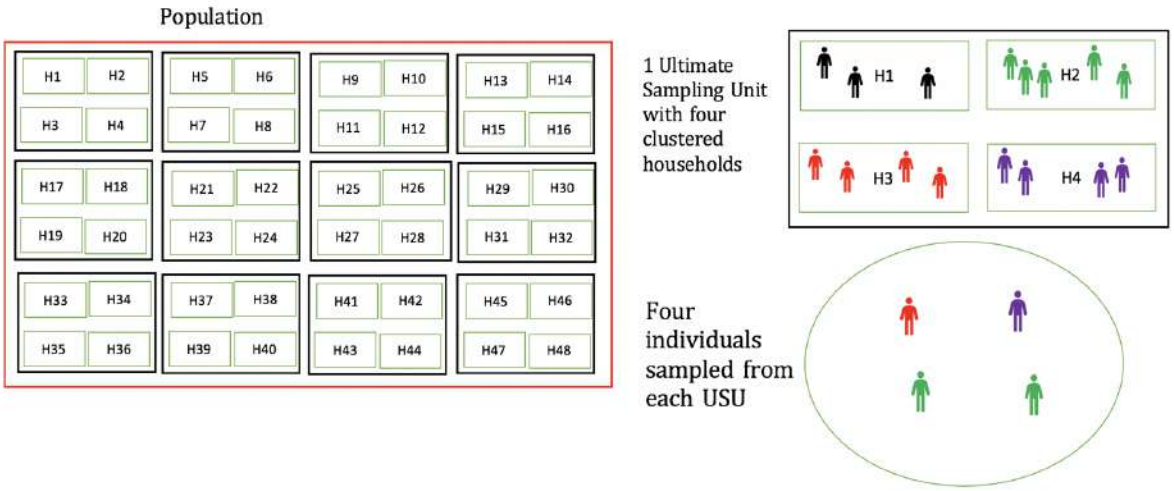


Figure 1: Sampling scheme employed in this study. USUs are black rectangles within the population. Oval shows individuals sampled using PPS method.

In the next section, section 2, longitudinal generalized variance functions (LGVFs) are briefly explained. Longitudinal adjusted design effect model (LADE) is discussed in section 3. Simulation process is discussed in section 4. Section 5, the final section, provides a succinct conclusion.

2. Longitudinal Generalized Variance Functions

We begin this section by providing a brief description of GVFs. Parameter of interest in this study is the proportion of people who benefit from the respective categories in table 1 on page 23. Let $\hat{p}$ be the estimated proportion of people benefiting for each such category. Let $\hat{T}$ be the estimated total number of people in the state of NM. T is estimated by $Horvitz-Thompson$ estimator

$$\hat{T}_t = \sum_h \sum_{i \in S_{th}} \left[\frac{M_{th} M_{thi} \bar{y}_{thi}}{n_{th} M_{thi}} \right] \quad , \tag{1}$$

where M_{th} is the total number of people sampled by CPS in stratum h for year t, technically, it is the total number of people sampled by CPS in NM for year t. For an instance, $M_{2010h} =$

2108, with a maximum number of people sampled from a single, unique household ("h_seq") being 20. M_{thi} is the number of secondary sampling units (SSU) in i^{th} primary sampling unit (PSU) for year t within stratum h. M_{thi} equals 4 for our study as shown in Figure 1. n_{th} is the number of PSUs picked in the sample within stratum h, this quantity equals 100 in this study. $\bar{y}_{thi}$ is the average of responses for year t and i^{th} PSU within stratum h. $\hat{T}_{thi} = M_{thi}\bar{y}_{thi}$ is the estimated total number of responses for year t and i^{th} PSU within stratum h.

From the formulation in Zhang *et al.* (2019), $\text{var}(\hat{T})$ can be estimated by using GVF

$$\widehat{\text{var}}(\hat{T}) = \hat{a}\hat{T}^2 + \hat{b}\hat{T}, \tag{2}$$

which after incorporating the time or the population effect gives us the regression model as in Zhang *et al.* (2019)

$$\text{relvar}(\hat{p}) = \hat{v}_{tv} = \hat{a} + \hat{b} \cdot \frac{e_t}{\hat{T}} \tag{3}$$

where

$$\text{relvar}(\hat{p}) = \frac{\text{Var}(\hat{p})}{[\text{E}(\hat{p})]^2}, \tag{4}$$

$\hat{a}$ and $\hat{b}$ are estimated linear regression coefficients.

$e_t = M_t/\bar{M}$ takes into account the effect of change of population in NM for year t, where M_t is the population total for the state of NM reported by U.S Census Bureau for year t, $\bar{M}$ is the average population total over 15 years, $\hat{v}_{tv}$ is the response variable, and $e_t/\hat{T}$ forms the predictor variable. The need to incorporate the population effect can be explained using Figure 2 where the change in population of NM over 2003-2018 is shown. Equation (3) is a LGVF model. This model spans over multiple years of data. This is a generalization of GVFs over time. Interested reader can refer to Zhang *et al.* (2019) for more detailed description of this model.

To evaluate $\widehat{\text{var}}(\hat{T})$ in equation (2), we have made use of the estimator mentioned by Royall (1986),

$$\widehat{\text{var}}(\hat{T}_t) = \sum_h n_{th}(n_{th} - 1)^{-1} \sum_{S_{th}} \gamma_{thi}^2 r_{thi}^2, \tag{5}$$

where

$$\gamma_{thi} = M_{th}(n_{th}M_{thi})^{-1} \quad \text{and} \quad r_{thi} = \hat{T}_{thi} - \left(\sum_{S_{th}} \gamma_{thj} \cdot \hat{T}_{thj}/M_{th}\right) M_{thi}.$$

3. Longitudinal Adjusted Design Effect Model

In this section, we discuss incorporating design effects in LGVFs. We introduce the design effect d_{tv} and the adjusted design effect $f_{tv} = d_{tv}/\bar{d}_t$. $\bar{d}_t$ is the average of design effects

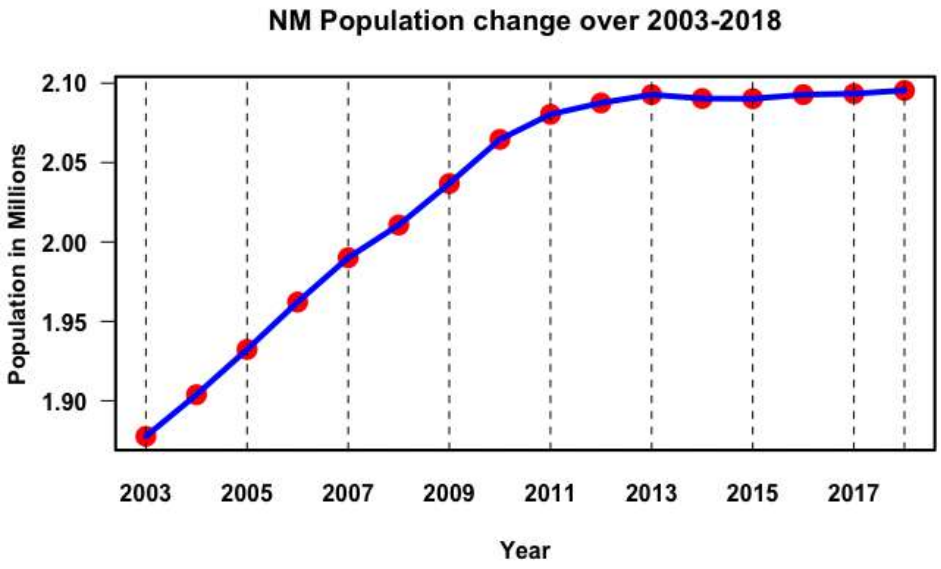


Figure 2: NM's population change over time.

for year t. $V = 18$ is the number of variables considered in the model and $\tau = 15$ is the number of years over which the model is spanned. $\boldsymbol{\theta} = (a, b)'$ be the LGVF coefficients which need to be estimated. $e_t = M_t/\bar{M}$ follows from previous section. Hence, for grouped variable case, we have $(V \times \tau)$ observations for regression to estimate a and b, whereas for individual variables case we have $(1 \times \tau)$ observations to estimate a and b. Let $a_{tv} = a = -\bar{d}_v/m$, and $b_{tv} = b = \bar{M}\bar{d}_v/m$ be the coefficients, from equation (2) we have

$$\begin{aligned}\widehat{\text{var}}(\hat{T}_{tv}) &= \frac{-\bar{d}_t}{m}\frac{d_{tv}}{\bar{d}_t}\hat{T}_{tv}^2 + \frac{\bar{M}\bar{d}_t}{m}\frac{M_t}{\bar{M}}\frac{d_{tv}}{\bar{d}_t}\hat{T}_{tv} \\ &= a_{tv}f_{tv}\hat{T}_{tv}^2 + b_{tv}e_t f_{tv}\hat{T}_{tv} \quad .\end{aligned}$$

The relative variance of $\hat{p}$ could be estimated by υ_{tv} for $t = 1, 2, \cdots, 15$ and $v = 1, 2, \cdots, 18$. We get

$$\upsilon_{tv}^* = \frac{\upsilon_{tv}}{f_{tv}} = a_{tv} + b_{tv} \cdot \frac{e_t}{\hat{T}_{tv}} \quad . \tag{6}$$

Equation (6), the LADE model, is applied for grouped variable variance estimation, and for individual variables estimation with $v = 1$. Properties of the estimators are reported in Zhang *et al.* (2019).

4. Results

We present the results for individual variables variance estimation and grouped variables variance estimation using LGVF and LADE models in this section.

4.1. Simulation

The data from 2003-2017 is considered to be the population for this study. Within each year, the households are assigned a USU after arranging the households in the increasing order of "h_seq". Then, 4 households are combined in that order to form a USU, 4 individuals are sampled from each USU. For years 2005, 2011 and 2014, one USU is dropped from each year because the USU contained 3 individuals. Table 2 on page 23 shows the number of USUs for each year. Following steps explain the simulation procedure in detail:

1. 100 USUs are picked based on PPS of the USUs which is about 50% sampling rate for each USU. 500 such random samples of size 100 USUs are picked for each year.

2. Estimates for relative variance, $\upsilon_{tv} = \widehat{\text{var}}(\hat{T}_{tv})/\hat{T}_{tv}$ are calculated using equations (1) and (5). Population totals are also calculated for year t. The population adjustment e_t is recorded as well. NM population totals (M_t) for years 2010-2017 are obtained from the US Census Bureau *factfinder.census.gov* (2020) and for years 2003-2009 from *countryeconomy.com* (2020). M_t is shown in Table 2 on page 23.

3. Ordinary least squares (OLS) regression model: LGVF1 model

$$\upsilon_{tv} = a_{tv} + b_{tv} \cdot \frac{e_t}{\hat{T}_{tv}} \tag{7}$$

   is applied and fits along with the coefficient estimates are recorded. Weighted least squares (WLS) regression – LGVF2 is applied with $weights = 1/\upsilon_{tv}$, and LGVF3 with $weights$ estimated from regressing *residuals* from OLS (LGVF1) onto $e_t/\hat{T}_{tv}$ is also applied.

4. The Adjusted design effect $f_{tv} = d_{tv}/\bar{d}_t$ is recorded and $\upsilon^*_{tv} = \upsilon_{tv}/f_{tv}$ is calculated.

5. OLS regression: LADE1 model

$$\upsilon^*_{tv} = a_{tv} + b_{tv} \cdot \frac{e_t}{\hat{T}_{tv}} \tag{8}$$

   is applied and fits along with the estimated coefficients are recorded. WLS regression – LADE2 is applied with $weights = 1/\upsilon^*_{tv}$, and LADE3 with $weights$ estimated from regressing *residuals* from OLS (LADE1) onto $e_t/\hat{T}_{tv}$ is also applied.

6. All the LGVF and LADE models are applied for grouped variables with $V \times \tau = 18 \times 15 = 270$ observations, and also for individual variables with $1 \times \tau = 1 \times 15 = 15$ observations.

7. This process is repeated for all the $R = 500$ samples picked in step (1).

8. Results along with the formulas used to calculate mean squared error (MSE), mean squared prediction error (MSPE), and Bias^2 are shown in Table 3 - 7 on page 24 - 30 in appendix.

5. Conclusion

Implications of obtained results are explained in this section. First, from grouped variable analysis – Table 3-5 on page 24-26, LADE models beat their LGVF counterparts when sum of MSEs, MSPEs or Bias^2 is considered but not necessarily for each variable, for example in Table 3 and 4, MSE and MSPE for variable 1 LADE1 model is higher than that of LGVF1 model. Similar conclusion can be drawn from Figure 3-4.

Second, from individual variable analysis – Table 6-7, LADE models outperform their LGVF counterparts when sum of MSEs or Bias^2 is considered and also for each variable when MSE is considered. Similar conclusion can be drawn from Figure 5.

Lastly, from overall analysis, LADE3 model – WLS regression where *weights* are estimated by using the *residuals* from OLS (LADE1), looks most promising out of all the models.

All in all, this paper has extended the results found in literature to individual variable variance estimation and proven that this application produces smaller error than the existing grouped variables variance estimation for longitudinal survey data. The idea of applying LGVFs for variance estimation in survey data is strengthened from the results obtained in this paper.

Acknowledgments

We thank the Chair Editor for his valuable advice and guidance.

References

Cho, M., Eltinge, J., Gershunskaya, J. and Huff, L. (2002). Evaluation of generalized variance function estimators for the U.S. current employment survey. *U.S. Bureau of Labor Statistics*.

countryeconomy.com. (2020). Retrieved from `http://www.countryeconomy.com/`

factfinder.census.gov. (2020). Retrieved from `http://www.factfinder.census.gov/`

Johnson, E. G. and King, B. F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics*, **3**, 235-250.

McIllece, J. (2016). Calculating generalized variance functions with a single series model in the current population survey. *U.S. Bureau of Labor Statistics*.

Royall, R. M. (1986). The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of the American Statistical Association*, **81(393)**, 119-123.

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, **82(398)**, 499-508.

Wolter, K. M. (2007). *Introduction to Variance Estimation. 2nd Edition.* New York: Spring-Verlag.

Zhang, G., Cheng, Y. and Lu, Y. (2019). Generalised variance functions for longitudinal survey data. *Statistical Theory and Related Fields*, **3(2)**, 150-157.

APPENDIX: Tables

Table 1: Eighteen variables studied – binary questions

SL. NO.	Variable	SL. NO.	Variable
1	Own business or self-employment	10	Survivors' payments
2	Unemployment compensation	11	Retirement payments
3	Social security benefits	12	Interest payments
4	Veterans' benefits	13	Dividend payments
5	Disability payments	14	Rental payments
6	Farm self-employment	15	Education benefits
7	Supplemental security benefits	16	Child support payments
8	Worker's compensation	17	Financial assistance payments
9	Public assistance/welfare benets	18	Other income payments

Table 2: Number of USUs for each year. One USU is dropped from 2005, 2011, and 2014.

Year	USUs	$\boldsymbol{Population\ total = M_t}$
2003	253	1877574
2004	247	1903808
2005	232	1932274
2006	214	1962137
2007	220	1990070
2008	205	2010662
2009	208	2036802
2010	193	2064588
2011	184	2080395
2012	186	2087549
2013	184	2092792
2014	192	2090342
2015	318	2090211
2016	384	2092789
2017	371	2093395

In the following tables, second-last row is the sum, and last row is the mean of the respective column. $r = \{1, 2, \cdots, R = 500\}$, implies the number of iterations.

(1) Simulation-Grouped variables results:

Table 3: Grouped-MSE

Grouped-MSE						
SL. NO.	**LGVF1**	**LGVF2**	**LGVF3**	**LADE1**	**LADE2**	**LADE3**
1	0.00979473	0.00070947	0.00260306	0.01186915	0.00016289	0.00169219
2	0.02570878	0.06140138	0.037641	0.01771583	0.02574147	0.01973988
3	0.00512692	0.01052768	0.00302777	0.00677246	0.00292996	0.00213194
4	0.10071434	0.33833742	0.33316984	0.07326416	0.22924619	0.19022522
5	0.01381009	0.00008473	0.00389565	0.01558064	0.00001434	0.00228918
6	0.00340038	0.00601356	0.00163116	0.0040311	0.00125022	0.00092687
7	0.01309319	0.06647825	0.05722773	0.01129848	0.03332084	0.02568329
8	0.00646665	0.0149201	0.00374728	0.00466677	0.00234624	0.00162239
9	0.03551605	0.09590948	0.07396578	0.0310229	0.05307642	0.04046547
10	0.05594735	0.23740761	0.20623049	0.043161	0.16183356	0.11629299
11	0.0099287	0.0006615	0.00267133	0.0102696	0.0001471	0.00144438
12	0.01524628	0.00002822	0.00436762	0.02292429	0.00000854	0.00346307
13	0.01025905	0.00056497	0.00265315	0.01552496	0.00011266	0.00227115
14	0.00619912	0.00223324	0.00166028	0.00774532	0.00049188	0.0011493
15	0.00743651	0.00149724	0.00182384	0.01175148	0.00030338	0.00169697
16	0.00415483	0.00408964	0.00115219	0.00819017	0.00076848	0.00127823
17	0.06939598	0.40039895	0.45240352	0.05689342	0.29200676	0.24633193
18	0.10164471	0.53987857	0.61302082	0.07296997	0.39209662	0.32619876
$\sum_{v=1}^{18}\left\{\frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}}{15}\right\}$	0.49384366	1.78114201	1.80289251	0.4256517	1.19585755	0.98490321
$\frac{\sum_{v=1}^{18}\left\{\frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}}{15}\right\}}{18}$	0.02743576	0.09895233	0.1001607	0.02364732	0.06643653	0.05471685

Table 4: Grouped-MSPE

Grouped-MSPE						
SL. NO.	**LGVF1**	**LGVF2**	**LGVF3**	**LADE1**	**LADE2**	**LADE3**
1	0.14692099	0.010642	0.03904588	0.17803722	0.00244339	0.0253828
2	0.38563175	0.92102077	0.564615	0.26573739	0.38612198	0.29609827
3	0.07690383	0.15791527	0.04541662	0.10158697	0.0439494	0.03197908
4	1.51071516	5.07506124	4.99754763	1.09896245	3.43869281	2.85337825
5	0.20715132	0.00127092	0.05843473	0.23370953	0.00021511	0.03433766
6	0.05100575	0.09020342	0.02446746	0.06046653	0.01875324	0.01390303
7	0.19639785	0.99717382	0.85841597	0.16947713	0.49981262	0.38524942
8	0.09699981	0.22380151	0.05620918	0.0700015	0.03519358	0.0243359
9	0.53274068	1.43864213	1.10948665	0.46534351	0.79614627	0.60698199
10	0.83921018	3.56111412	3.09345728	0.64741494	2.42750345	1.74439489
11	0.14893043	0.0099225	0.04006995	0.15404396	0.00220643	0.02166576
12	0.22869413	0.00042323	0.06551435	0.34386428	0.00012817	0.05194609
13	0.15388573	0.00847451	0.03979731	0.23287445	0.00168997	0.03406732
14	0.09298682	0.03349853	0.02490425	0.11617975	0.00737819	0.0172395
15	0.11154768	0.02245857	0.02735761	0.17627213	0.00455069	0.02545455
16	0.06232243	0.06134462	0.01728288	0.12285255	0.01152714	0.01917341
17	1.04093967	6.00598423	6.78605287	0.85340126	4.38010134	3.69497891
18	1.52467068	8.09817858	9.19531234	1.09454957	5.88144932	4.89298145
$\sum_{v=1}^{18}\left\{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}\right\}$	7.40765489	26.71713	27.043388	6.38477512	17.9378631	14.7735483
$\frac{\sum_{v=1}^{18}\left\{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}\right\}}{18}$	0.41153638	1.484285	1.50241044	0.35470973	0.99654795	0.82075268

Table 5: Grouped-Bias^2

Grouped-Bias^2						
SL. NO.	**LGVF1**	**LGVF2**	**LGVF3**	**LADE1**	**LADE2**	**LADE3**
1	0.00932572	0.00043432	0.00197376	0.01064828	0.00004937	0.00101139
2	0.0062715	0.02416079	0.00228669	0.00334595	0.00323062	0.00079391
3	0.00081157	0.0059821	0.00000064	0.00229332	0.00113499	0.0000174
4	0.04073135	0.04129099	0.00143959	0.02801676	0.00076304	0.00020865
5	0.01337792	0.00002794	0.00321029	0.01409359	0.00000051	0.00147906
6	0.00186233	0.0039306	0.00020954	0.0024186	0.0005942	0.00010469
7	0.00058671	0.01136091	0.00012126	0.00003427	0.00089336	0.00001651
8	0.00003579	0.00792993	0.00001012	0.0001096	0.00115992	0.00000763
9	0.00900689	0.02377147	0.0010449	0.00887119	0.00356649	0.00150128
10	0.01459668	0.02012355	0.0000033	0.01174225	0.00015466	0.00000029
11	0.00948235	0.00039525	0.00207171	0.00913495	0.00005537	0.00084008
12	0.01478368	0.00000039	0.0036159	0.02086139	0.00000409	0.00227514
13	0.00981552	0.00037491	0.00202855	0.01382798	0.00003019	0.0013744
14	0.0055339	0.00149276	0.00094956	0.00660325	0.00020156	0.00052133
15	0.00690787	0.00105097	0.0012046	0.01021451	0.0001159	0.00091916
16	0.00321939	0.0029361	0.0003016	0.00660708	0.00034792	0.00049029
17	0.01583157	0.01323211	0.00255808	0.01319929	0.0005553	0.00105202
18	0.03075134	0.01860175	0.00295764	0.0225499	0.00221982	0.00244322
$\sum_{v=1}^{18} \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)})}{R}\right\}}{15} - \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(v_{tv}^{(r)})}{R}\right\}}{15}$	0.19293208	0.17709684	0.02598773	0.18457216	0.01507731	0.01505645
$\frac{\sum_{v=1}^{18} \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)})}{R}\right\}}{15} - \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(v_{tv}^{(r)})}{R}\right\}}{15}}{18}$	0.01071845	0.00983871	0.00144376	0.01025401	0.00083763	0.00083647

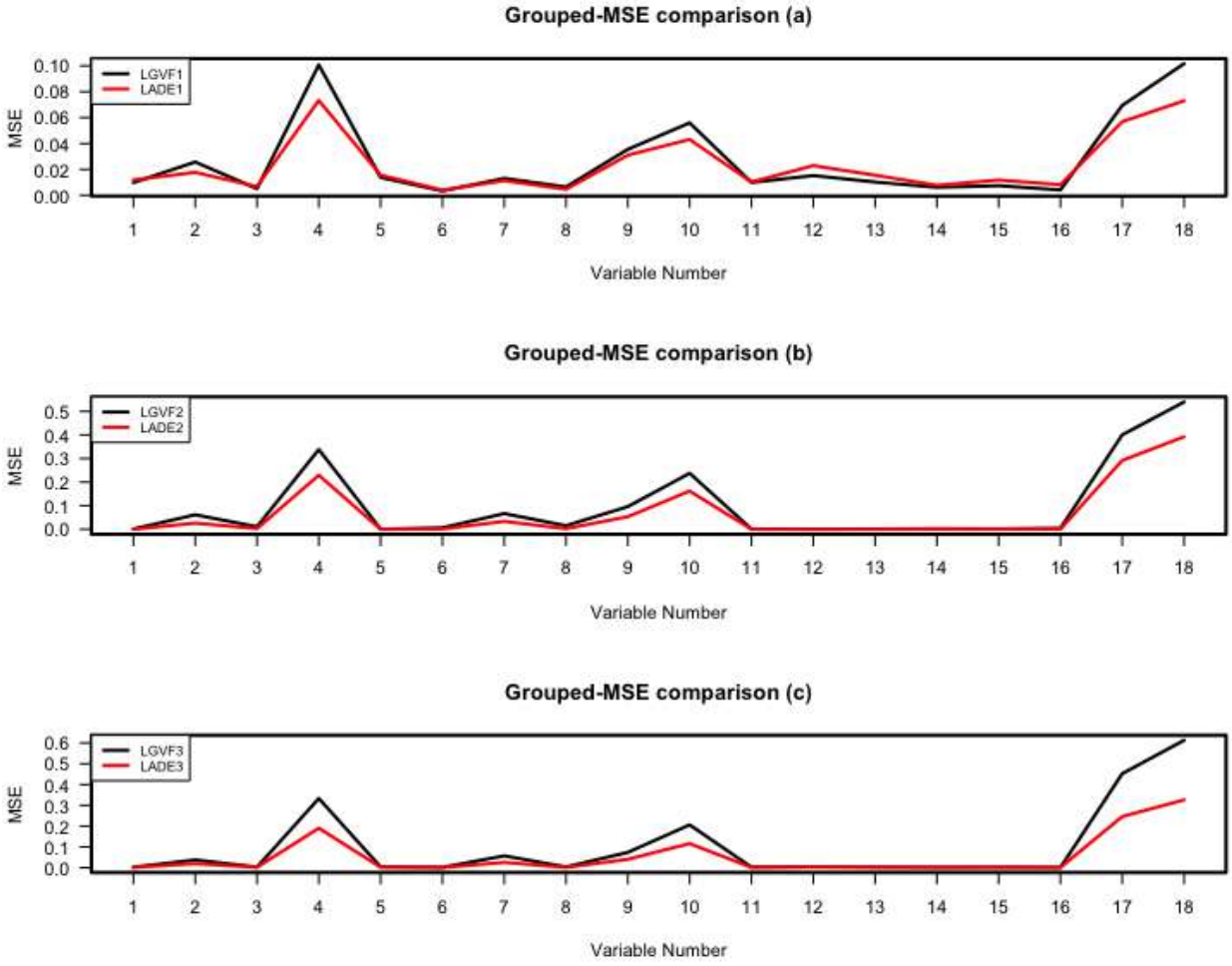


Figure 3: Grouped-MSEs for all variables – LADE models outperform LGVFs for most of the variables, but not all variables when each variable is compared.

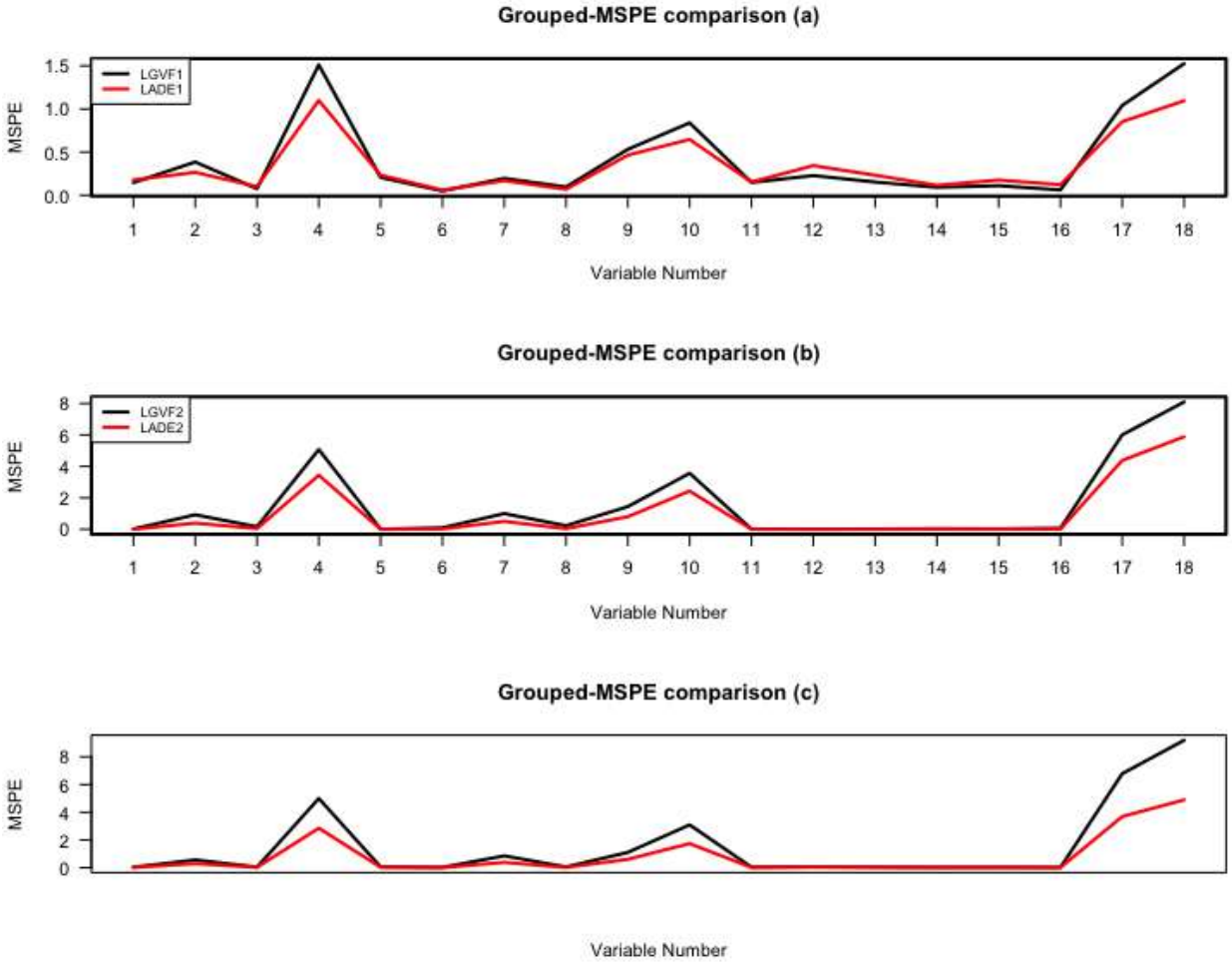


Figure 4: Grouped-MSPEs over all variables – LADE models outperform LGVFs for most of the variables, but not all the variables when each variable is compared.

(2) Simulation-Individual variables results:

Table 6: Individual-MSE

Individual-MSE						
SL. NO.	**LGVF1**	**LGVF2**	**LGVF3**	**LADE1**	**LADE2**	**LADE3**
1	0.00013186	0.00036702	0.00013588	7.5904E-05	8.5258E-05	7.6991E-05
2	0.00911577	0.03348467	0.00975517	0.00475778	0.00640601	0.00490773
3	0.00209517	0.00611522	0.00223322	0.00138177	0.00168737	0.00142106
4	0.03552953	0.10924189	0.03668525	0.03219963	0.03809822	0.03219542
5	1.1899E-05	6.5473E-05	1.2203E-05	6.9033E-06	7.3247E-06	6.9367E-06
6	0.00098323	0.00373517	0.00110428	0.00056473	0.00064215	0.00056823
7	0.00384158	0.02781115	0.00407711	0.00225242	0.00275966	0.00225011
8	0.00221485	0.01097709	0.00238354	0.00097313	0.00125433	0.00097575
9	0.01099218	0.04044535	0.01133325	0.00670696	0.00894584	0.00675416
10	0.02237646	0.0797136	0.02348702	0.01589717	0.02046451	0.01587004
11	0.0001145	0.00049466	0.00011709	7.3763E-05	8.5823E-05	7.4661E-05
12	2.9673E-06	1.9645E-05	3.0933E-06	1.0193E-06	1.058E-06	0.00000104
13	7.9252E-05	0.00029655	8.3895E-05	6.3977E-05	6.9632E-05	6.4492E-05
14	0.00036913	0.0014005	0.00038955	0.00023347	0.00026231	0.00023636
15	0.00020298	0.00083692	0.00020594	0.00016128	0.0001765	0.00016184
16	0.00052875	0.00221343	0.00054544	0.00036051	0.00040721	0.00036441
17	0.02863593	0.07480319	0.02905497	0.02817858	0.03538639	0.02789071
18	0.0402084	0.10769188	0.04082121	0.0377467	0.04157593	0.03749089
$\sum_{v=1}^{18}\left\{\frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}}{15}\right\}$	0.15743444	0.49971342	0.16242812	0.13163568	0.15831552	0.13131083
$\frac{\sum_{v=1}^{18}\left\{\frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)}-v_{tv}^{(r)})^2}{R}\right\}}{15}\right\}}{18}$	0.00874636	0.02776186	0.00902378	0.00731309	0.00879531	0.00729505

Table 7: Individual-Bias^2

Individual-Bias^2						
SL. NO.	**LGVF1**	**LGVF2**	**LGVF3**	**LADE1**	**LADE2**	**LADE3**
1	0.00000000	0.00007995	0.00000000	0.00000001	0.00000765	0.00000002
2	0.00000000	0.00781023	0.00000119	0.00000282	0.00062654	0.00000073
3	0.00000000	0.00148825	0.00000013	0.00000006	0.00022416	0.00000001
4	0.00000000	0.03136575	0.00000606	0.00039485	0.00357391	0.00032367
5	0.00000000	0.00001776	0.00000000	0.00000001	0.00000063	0.00000001
6	0.00000000	0.00080929	0.00000009	0.00000092	0.00006541	0.00000068
7	0.00000000	0.00396117	0.00000016	0.00000796	0.00027182	0.00000763
8	0.00000000	0.00252299	0.00000046	0.00000268	0.00016183	0.00000210
9	0.00000000	0.01072011	0.00000056	0.00000655	0.00110930	0.00000481
10	0.00000000	0.01717795	0.00000137	0.00008476	0.00138620	0.00007080
11	0.00000000	0.00015332	0.00000000	0.00000001	0.00001154	0.00000001
12	0.00000000	0.00000501	0.00000000	0.00000000	0.00000005	0.00000000
13	0.00000000	0.00006927	0.00000002	0.00000014	0.00000739	0.00000013
14	0.00000000	0.00037697	0.00000004	0.00000008	0.00002618	0.00000006
15	0.00000000	0.00022242	0.00000000	0.00000016	0.00002066	0.00000012
16	0.00000000	0.00057432	0.00000003	0.00000019	0.00004734	0.00000006
17	0.00000000	0.01985565	0.00000033	0.00023859	0.00379817	0.00021831
18	0.00000000	0.03093950	0.00000188	0.00042347	0.00433762	0.00034524
$\sum_{v=1}^{18} \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)})}{R}\right\}}{15} - \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(v_{tv}^{(r)})}{R}\right\}}{15}$	0.00000000	0.12814991	0.00001232	0.00116326	0.01567640	0.00097439
$\frac{\sum_{v=1}^{18} \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(\hat{v}_{tv}^{(r)})}{R}\right\}}{15} - \frac{\sum_{t=1}^{15}\left\{\frac{\sum_{r=1}^{500}(v_{tv}^{(r)})}{R}\right\}}{15}}{18}$	0.00000000	0.00711944	0.00000068	0.00006463	0.00087091	0.00005413

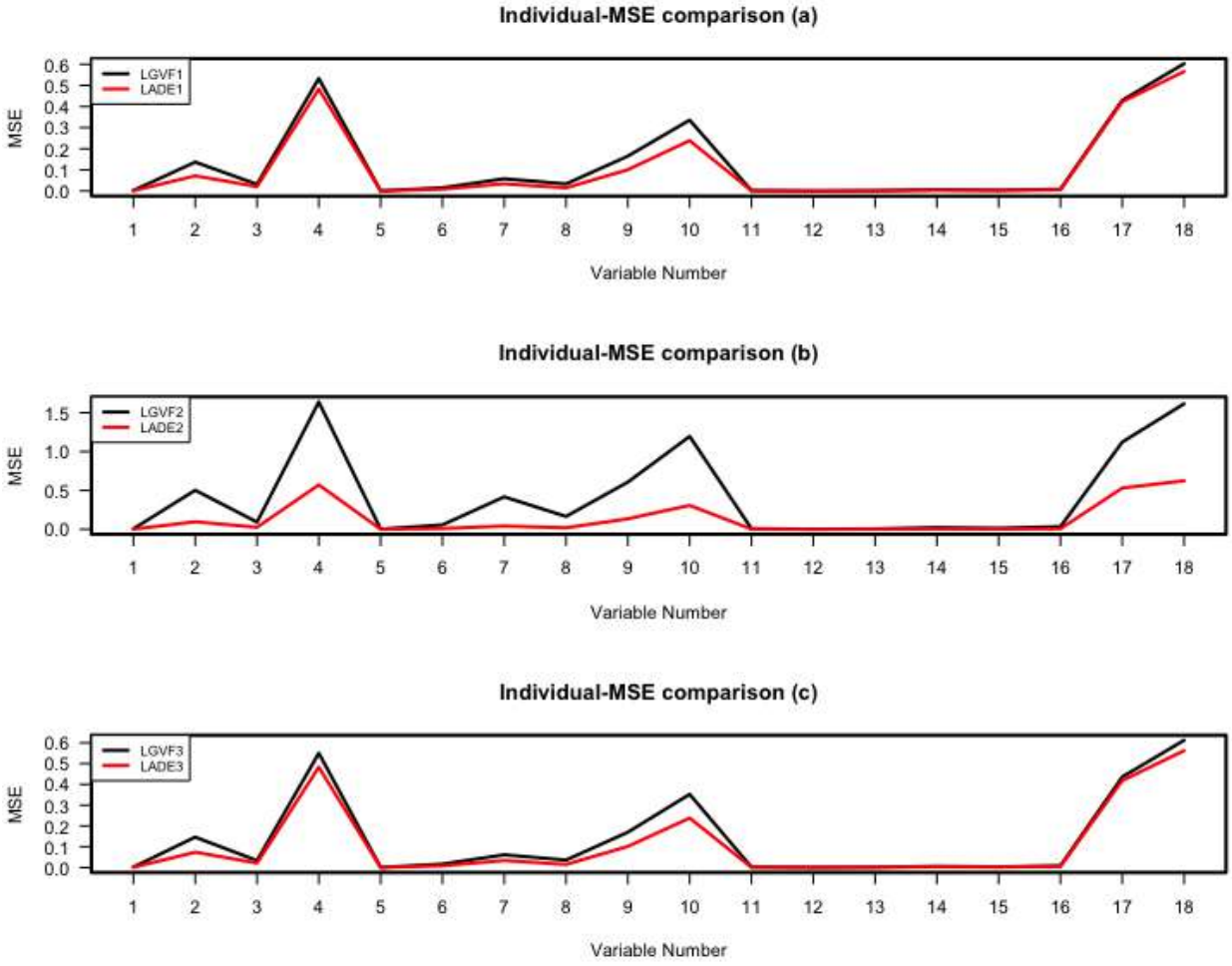


Figure 5: Individual-MSEs – LADE models outperform LGVFs for all of the variables when all the variables or each variable is compared.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 33-49

Selective Harvesting of Tobacco Leaves: An Approach Based on Texture Features

P. B. Mallikarjuna[1] and D. S. Guru[2]
[1]*Department of Computer Science & Engineering, JSS Academy of Technical Education, Bengaluru-560060, Karnataka, India.*
[2]*Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore-570006, Karnataka, India.*

Received: 17 June 2020; Revised: 24 December 2020; Accepted: 01 February 2021

Abstract

A texture-based model for classification of tobacco leaves for the purpose of selective harvesting is proposed in this paper. It relies on texture features which are extracted by various texture models to represent the roughness of leaves. Extracted texture features are fused using concatenation rule. Discriminative texture features are then selected by employing wrapper feature selection methods. Finally, *K-NN* classifier is adapted for the purpose of classification. An extensive experimentation has been conducted on our own dataset to evaluate the performance of the proposed model. The experimental results reveal that the proposed model has achieved the best classification accuracy.

Key words: Tobacco leaves; Harvesting; Texture features; Feature selection; Classification.

1. Introduction

Precision agriculture is an integrated crop management system that attempts to match the type and quantity of inputs with the actual crop requirements for small areas within a farm field (Srinivasan, 2001). The potential of precision agriculture in terms of economic and environmental benefits could be visualized through reduced use of water, fertilizers, herbicides and pesticides in addition to the cost farm equipments. Instead of managing an entire field based upon some hypothetical average conditions, a precision agriculture approach recognizes site-specific differences within the field and adjusts accordingly the actions of management (Goovaerts, 2000). The objectives of the precision agriculture are profit maximization, rationalization of agriculture input and environmental damage reduction, by restricting the agriculture practices to the site demands. These objective could be achieved by adapting some site specific practices such as application of agrochemicals, right time harvesting, and grading of crops. Human intervention in these practices raises many disadvantages such as wrong diagnosis of diseases in crops, wrong quality analysis of crops, man power, labor cost and time consuming. Therefore, we need to automate these practices to increase efficiency and speed using computer vision (*CV*) algorithmic models.

Requirement of Precision agriculture at different stages of plant growth is shown in Figure 1.

Corresponding Author: P B Mallikarjuna
Email: putaplar@gmail.com

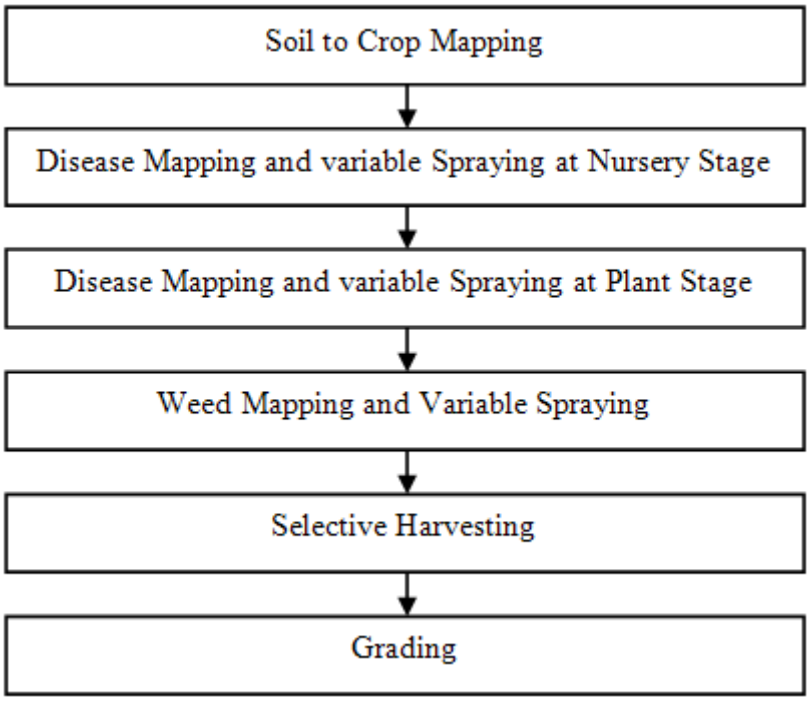


Figure 1: Requirement of precision agriculture at different stages of plant growth

Tobacco is a commercial crop in many countries like China, India, Brazil, United States, European Union, Zimbabwe, Indonesia, Malawi and Russia because of its high economic value. Especially in Karnataka state in India, that too around Mysore district, many farmers are depending on tobacco crop because of suitable climate conditions and soil. It created a gainful employment to several lakhs of people in India. Roughly 80 percent of the flue cured variety (*FCV*) of tobacco grown in Karnataka is being exported abroad to meet the demand of multinational industries for various purposes.

Harvesting is an important stage in tobacco crop. Tobacco crop is grown for production of quality leaves. The quality of a leaf depends upon the ripeness of the leaf while it is harvested. Therefore, while harvesting, farmers should look into factors such as unripe or ripe or over-ripe properties of leaves based on degree of ripeness of leaves. Ripeness of leaf begins after 50 days of plantation of tobacco seedlings. Harvesting usually begins after 60 days of plantation of tobacco seedlings. Leaves are removed at intervals as they ripened. Manual classification of unripe, ripe and over-ripe leaves is laborious, time consuming, inefficient and costly process. Automation of this process helps the tobacco farmers to gain more profit. Computer vision and image processing techniques can be exploited for classification of tobacco leaves supporting automatic harvesting, which increases the speed and accuracy of harvesting in addition to, reducing the number of human labors and cost.

With this backdrop, this work is to propose a model to automatically classifying tobacco leaves using computer vision technologies. Following are the overall contributions of this work.

- Development of a model which fuses the different texture features and selects the best discriminating features for classification of tobacco leaves on a plant for the purpose of harvesting ripen leaves.
- Segmentation of tobacco leaves from the background using *CIELAB* color model.

- Creation of a relatively large dataset of harvesting tobacco leaves due to non-availability of a benchmarking dataset.
- Conduction of experimentations on the created large tobacco dataset for demonstrating the effectiveness of the proposed model.

2. Related Work

Few attempts could be traced on ripeness evaluation of different crops for automatic harvesting. Medjool date fruits were taken as a case study to demonstrate the performance of a novel color quantization and color analysis technique for fruit maturity evaluation and surface defect detection (Lee *et al*., 2008). Direct color mapping method (Lee *et al*., 2011) was proposed for maturity evaluation of tomato and date fruits. This color mapping method maps the RGB values of colors of interest into 1-D color space using polynomial equations. It uses a single index value to represent each color in the specified range for the purpose of maturity evaluation of tomato and date fruits. A robotic system for harvesting ripe tomatoes in greenhouse (Yin *et al*., 2009) was designed. In this work, L*a*b color space was used to segment tomatoes from complex background and K-means clustering method was applied on segmented tomatoes to recognize ripe tomatoes. Recently L*a*b color features and their combination along with texture features have been applied for the purpose of grading of mangoes using hierarchical classification approach (Anitha *et al*., 2020). A novel and robust color space conversion and color index distribution analysis technique for automated date maturity evaluation (Lee *et al*., 2008) was proposed. Computer vision technology for detecting fruit size, color, bruise, surface defects and evaluation of fruit overall quality (Gao *et al*., 2010) were discussed. A genetic algorithm based neural network detecting system (Xu, 2009) was developed for evaluating maturity of strawberry fruits. In this paper, H frequency of HIS color model was used to distinguish maturity levels of strawberry fruits in a variable illumination conditions. An intelligent and robust algorithm (Furfaro *et al*., 2007) was proposed to estimate absolute percentages of under-ripe (green), ripe (yellow), and over-ripe (brown) coffee cherries displayed on the canopy surface. The proposed algorithm was tested on the multispectral images. It was based on a coupled leaf/canopy radiative transfer model (*LCM2*). Feasibility of monitoring coffee field ripeness with airborne multispectral imagery (Johnson *et al*., 2004) was proposed. In this work, reflectance spectrum was recorded from four major components of coffee field *viz.*, green leaf, under-ripe fruit, ripe fruit and over-ripe fruit. Based on reflectance spectrum, ripeness evaluation of coffee field was performed. A Bayesian classifier considering a multivariate, three-class problem (Baltazar *et al*., 2008) was incorporated for data fusion to classify fresh intact tomatoes based on their ripening stages. In this work, data extracted from multiple sensors were fused. Further, fused data was used for the purpose of classification.

In our recent publication (Guru *et al*., 2012), a model for classification of tobacco leaves for automatic harvesting of tobacco leaves using texture models was proposed. Apart from this, no attempts have been made on classification of tobacco leaves for automatic harvesting. In the proposed work, the classification accuracy has been improved by applying feature level fusion and feature selection methods.

3. Proposed Model

The proposed model consists of five stages – segmentation, feature extraction, feature level fusion, feature selection and classification. The color space model *CIELAB* is used to

segment tobacco leaf area from the background. Features are extracted from segmented tobacco leaf using various texture models such as *LBP* (Local Binary Pattern), *LBPV* (Local Binary Pattern Variance), *GLTP* (Gray Level Local Texture Pattern), *GFR* (Gabor Filter Response) and *WD* (Wavelets Decomposition). These features are fused on different combination of texture models. The obtained fused feature vector is normalized. Features are selected from fused feature vector using wrapper feature selection methods such as *SFS* (Sequential forward selection), *SFFS* (Sequential floating forward selection), *SBS* (Sequential backward selection) and *SFBS* (Sequential floating backward selection). Then, *K-NN* classifier is used for classification of tobacco leaves in to three classes – unripe, ripe and over-ripe.

3.1. Segmentation

We have selected *CIELAB* (Viscarra *et al*., 2006) color model to segment a leaf area from their background (soil, stones and noise). Since the color of a leaf varies from green to yellow, the chromacity coordinate is used to segment the leaf from its background. For an illustration, we have shown three different samples (Figures 2, 3 and 4) of tobacco leaves and also the results of the segmentation.

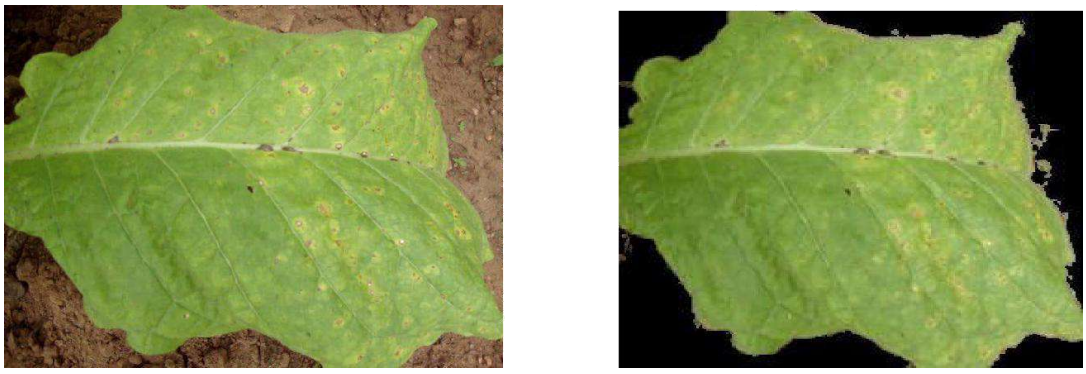
3.2. Feature extraction

Top surface of a leaf with rare maturity spots (see Figure 2) is smoother and its roughness increases as number of maturity spots increases (see Figures 3 and 4). This roughness is reflected by transitions in intensity levels on the surface of a leaf in the form of uniform and non-uniform patterns. To exploit this, we recommend to extract texture features from gray scale images of segmented tobacco leaves using the various texture based models *viz*., *LBP* (Ojala *et al*., 2002), *LBPV* (Guo *et al*., 2010), *GLTP* (He and Wang, 1990) (Surliandi and Kumar, 2008), *GFR* and *WD*.

3.3. Feature level fusion

Feature level fusion refers to combining different feature vectors that are obtained by employing multiple feature extraction algorithms. When the feature vectors are homogeneous, a single resultant feature vector can be obtained as a weighted average of the individual feature vectors. When the feature vectors are non-homogeneous, we can concatenate them to form a single feature vector (Jain *et al*., 2005).

The extracted feature vectors of *LBP*, *LBPV*, *GLTP*, *GFR* and *WD* are fused in all possible combinations by concatenating the feature vectors. The fused feature vectors are normalized using min-max method.



(a) (b)

Figure 2: (a) A sample tobacco leaf with rare maturity spots (b) Segmented image

(a) (b)

Figure 3: (a) A sample tobacco leaf with moderate maturity spots (b) Segmented image

(a) (b)

Figure 4: (a) A sample tobacco leaf with rich maturity spots (b) Segmented image

3.4. Feature selection

Feature selection is the process of selecting a subset of relevant features for building robust learning models. Feature selection is broadly classified into two categories such as

filter model and wrapper model. The filter model relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine best features for selection.

A well-known filter method Relief (Kira and Rendel, 1992) relies on relevance evaluation. Time Complexity of Relief for a dataset with M instances and N features is O(MN). However, the Relief method does not help to eliminate redundant features. Empirical evidence from feature selection literature shows that, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms and thus should be eliminated as well (Hall, 2000). Therefore, we have exploited feature selection methods based on wrapper model such as sequential forward selection (*SFS*), sequential floating forward selection (*SFFS*), sequential backward selection (*SBS*) and sequential floating backward selection (*SFBS*) (Ververidis and Kotropoulos, 2005, 2008). The criterion employed in these methods is the correct classification rate of the Bayes classifier assuming that the features obey the multivariate Gaussian distribution. These methods eliminate irrelevant features as well as redundant features but they are computationally slightly expensive than any filter method.

3.5. Classification

In the proposed model, the *K-NN* classifier based on G-statistic, Chi-square and Euclidean distance measure has been used to classify tobacco leaves into unripe, ripe and over-ripe for the purpose of harvesting.

3.5.1. Performance measures

To evaluate the correctness of classification algorithms, one should look into confusion matrix. A confusion matrix is a matrix plot of predicted versus actual classes of the samples.

Let k be the number of classes. Let r_i be the total number of samples of i^{th} class. Let c_i be the number of samples classified (labeled) as i^{th} class. Let T_i be the number of samples correctly labeled as i^{th} class. Then precision, recall, *F*-measure and classification accuracy (Espindola and Ebecken, 2005) are defined as follows.

Precision (P): Precision of the classifier model with respect to i^{th} class is the ratio of the number of samples correctly labeled as i^{th} class to the total number of samples labeled as i^{th} class. The precision of the classifier model with respect to i^{th} class is given by

$$\boldsymbol{P_i} = \frac{\boldsymbol{T_i}}{\mathbf{c_i}} \tag{1}$$

Recall (R): Recall of the classifier model with respect to i^{th} class is the ratio of the number of samples correctly labeled as i^{th} class to the total number of samples of i^{th} class. The recall of the classifier model with respect to i^{th} class is given by

$$R_i = \frac{T_i}{\mathrm{r}_i} \tag{2}$$

F-measure (F): F-measure is the harmonic mean of precision and recall and it is given by

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

Classification Accuracy (CA): It is the ratio of correctly classified samples to the total number of samples classified.

$$CA = \frac{\sum_{i=1}^{k} T_i}{\sum_{i=1}^{k} r_i} \tag{4}$$

4. Experimental Result

4.1. Dataset

Color images of tobacco leaves in real tobacco field are acquired using a Sony digital color camera. The leaves used for imaging are randomly selected from the tobacco field at Central Tobacco Research Institute (*CTRI*), Hunsur, Karnataka, India. Images are acquired at variable illumination conditions (sunny and cloudy). A total of 1300 sample images of size 250×250 are used for evaluating the proposed texture-based model.

Table 1: Number of samples of individual classes of tobacco leaves

Tobacco leaf Class	Number of samples	Total samples
Unripe leaf	323	1300
Ripe leaf	667	
Over-ripe leaf	310	

4.2. Experimentation

In the first set of experimentation, we conducted experiments for the proposed model based on individual texture models. During experimentation, we conducted four different sets of experiments. In the first set of experiment, we used 30% of the samples of each class of a harvesting dataset to create class representative vectors (training) and the remaining 70% of the samples for testing purpose. In the second set, third set and fourth set of experiments, the numbers of training and testing samples are in the ratio of 40:60, 50:50 and 60:40 respectively. In each set of experiment, experiments are repeated 20 times by choosing the training samples randomly. As measures of goodness of the proposed model based on individual texture model, we computed minimum, maximum, average and standard deviation of classification accuracy of all the 20 trails using the *K-NN* classifier. Classification accuracy of the proposed model based on individual texture models (*LBP*, *LBPV*, *GLTP*, *GFR* and *WD*) for 30% training, 40% training, 50% training and 60% training are depicted in Figures 5-8 respectively. It is observed from Figures 5-8 that the *GLTP* texture model has achieved a better average classification accuracy in experiment 4 (60% training samples) when compared to the other texture models. It is also observed that the proposed model has achieved a good classification accuracy for all the individual texture models in experiment 4 (60% training samples) when compared to that of experiment 1 (30% training samples), experiment 2 (40% training samples) and experiment 3 (50% training samples).

Therefore, we present experimental results obtained for 60% training only for fusion of texture features and application of feature selection method. In the second set of experimentation, we conducted experiments for the proposed model based on fusion of texture features. That is, the extracted feature vectors of *LBP*, *LBPV*, *GLTP*, *GFR* and *WD* are fused in all possible combinations by concatenating the feature vectors. During experimentation, experiments are repeated 20 times by choosing the training samples randomly. As measures of goodness, we computed minimum, maximum, average and standard deviation of classification accuracy of all the 20 trails using *K-NN* classifier. Classification accuracy of the proposed model based on fusion of 2 texture models at a time, 3 texture models at a time, 4 texture models at a time and all 5 texture models are depicted respectively in Figures 9-12. It is observed from Figures 9-12 that overall fusion of *GLTP* and *WD* features has achieved the best average classification accuracy when compared to the other fusion of texture models in any combination. Fusion of *GLTP* and *WD* features has achieved an improvement in classification accuracy when compared to that of an individual texture model. It is also observed that for all combinations of fusion of texture features, good classification accuracy is achieved for G-statistic distance measure when compared to the Chi-square and the Euclidean distance measures. Also, Confusion matrix for fusion of *GLTP* and *WD* features is tabulated in Table 2 and performance measures such as Precision, Recall and F-measure of individual classes are tabulated in Table 3.

In third set of experimentation, we conducted experiments based on fusion of texture features and the application of feature selection method. That is, we applied the wrapper feature selection method (*SFS*, *SBS*, *SFFS* and *SFBS*) on fused texture feature vector to reduce the dimension of feature matrix and to obtain discriminative texture features. During experimentation, experiments are repeated 20 times by choosing the training samples randomly. Classification accuracy of the proposed model based on fusion of texture features and feature selection method for 2 texture models at a time, 3 texture models at a time, 4 texture models at a time and all 5 features are depicted respectively in Figures 13- 16. Here we presented results of G-statistic for 60% training as it was observed to have good results for the G-statistic based *K-NN* classifier. It is observed from Figures 13-16 that the fusion of *GLTP* and *WD* features with *SBS* feature selection method has achieved best average classification accuracy when compared to other combinations. Also, Confusion matrix for fusion of *GLTP* and *WD* features with *SBS* feature selection method is tabulated in Table 4 and performance measures such as Precision, Recall and F-measure of individual classes are tabulated in Table 5.

5. Discussion

From the experimental results, it is observed that the *GLTP* texture model has dominant features when compared to *LBP*, *LBPV*, *GFR* and *WD*. Since the *GLTP* is built by the advantages of TS and *LBP*, it reveals more local texture information when compared to texture models such as Gabor response and Wavelet decomposition. The *GLTP* assigns a label (uniform label or non-uniform label) for each pixel in an image based on the uniformity or non-uniformity of neighborhood, where as the *GFR* is based on frequency and orientation of edge information. Though, the *GFR* is rotation invariant local texture information, fusion of Gabor response with the other texture models such as *LBP*, *LBPV*, *GLTP* and Wavelet decomposition has deteriorated the performance because the Gabor response will not represent edge information in the form of uniform patterns and non-uniform patterns. Fusion of *LBPV* with other texture models such as *LBP*, *GLTP*, Gabor response and Wavelet decomposition has also deteriorated the performance because global information such as

variance is embedded with local information in the *LBPV*. Therefore, the *LBPV* will deteriorate the classification accuracy.

Feature selection after fusion has improved the classification results for all combination of fusion of texture models. In all feature selection methods, the dominant features are selected. When analyzed we observed that, the *SBS* method on fusion of *GLTP* and Wavelets has selected only 15 features out of 55 features (*GLTP* – 46 features and Wavelets – 9 features). Out of 15 features, 10 features are from the *GLTP* and 5 features are from the *WD*. Therefore, the *GLTP* has more number of discriminating features with *WD* features and improve the classification accuracy. Similarly, the *SFS* on fusion of *GLTP* and *LBP* has selected only 6 features out of 56 features (*GLTP* – 46 features and *LBP* – 10 features). All 6 features are from the *GLTP* alone. This indicates that the *LBP* has no discriminating dominant features when it is with the *GLTP* features. Further, the *SFBS* on fusion of *GLTP*, *LBP* and *WD* has selected only 8 features. Out of 8 features, 4 features are from the *GLTP* and 4 features are from the *WD*. No discriminating features of *LBP* are selected when they are with *GLTP* and *WD* features.

The above observations appraise that *GLTP* and *WD* features have more discriminating and dominating features when compared to the other texture models such as *LBP*, *LBPV* and *GFR*.

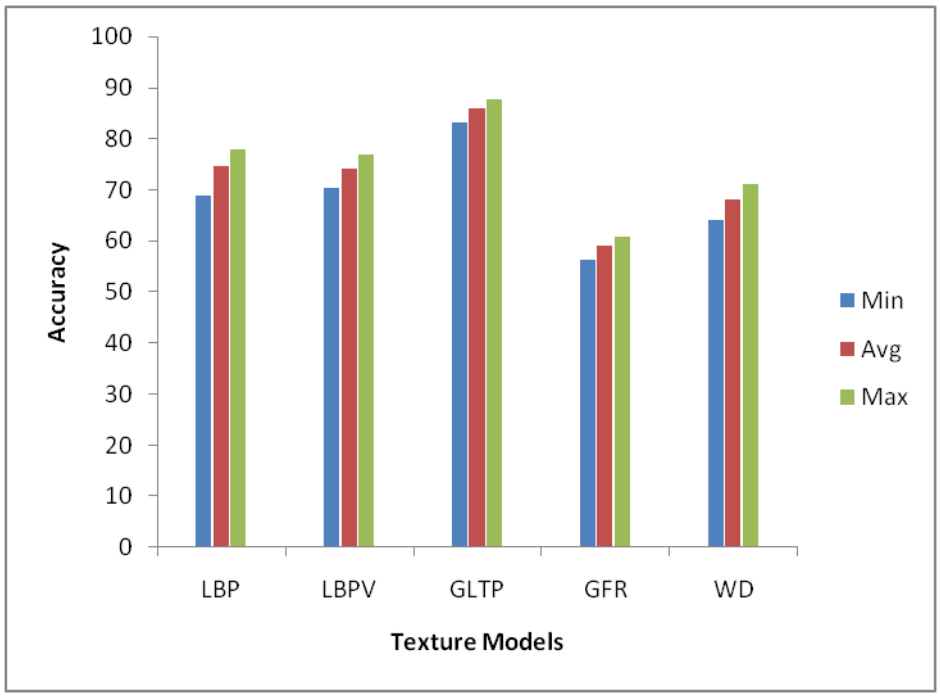


Figure 5: Classification accuracy of the proposed model based on individual texture models for 30% training

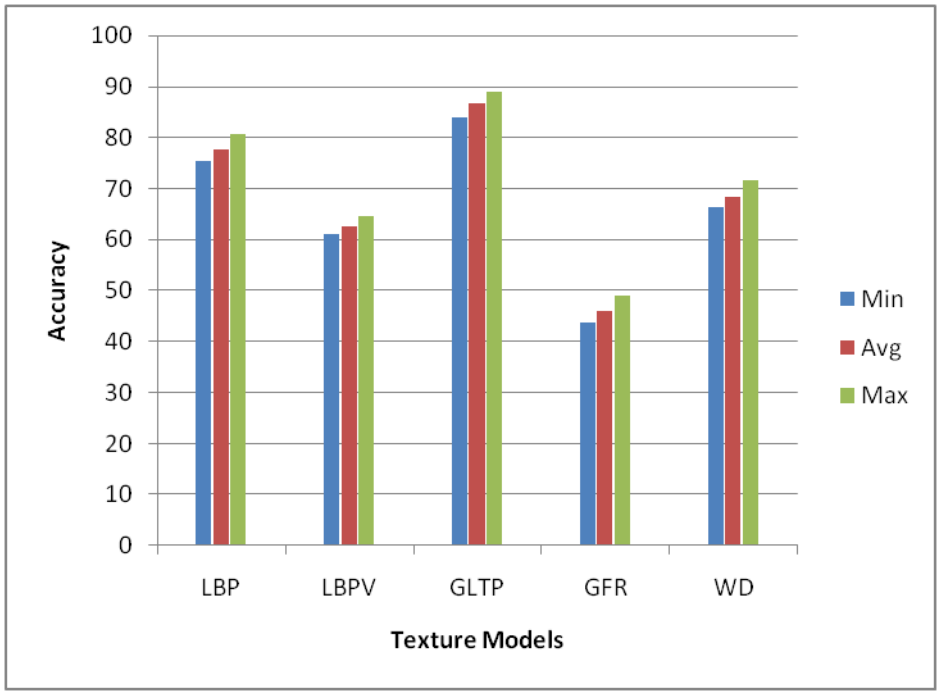


Figure 6: Classification accuracy of the proposed model based on individual texture models for 40% training

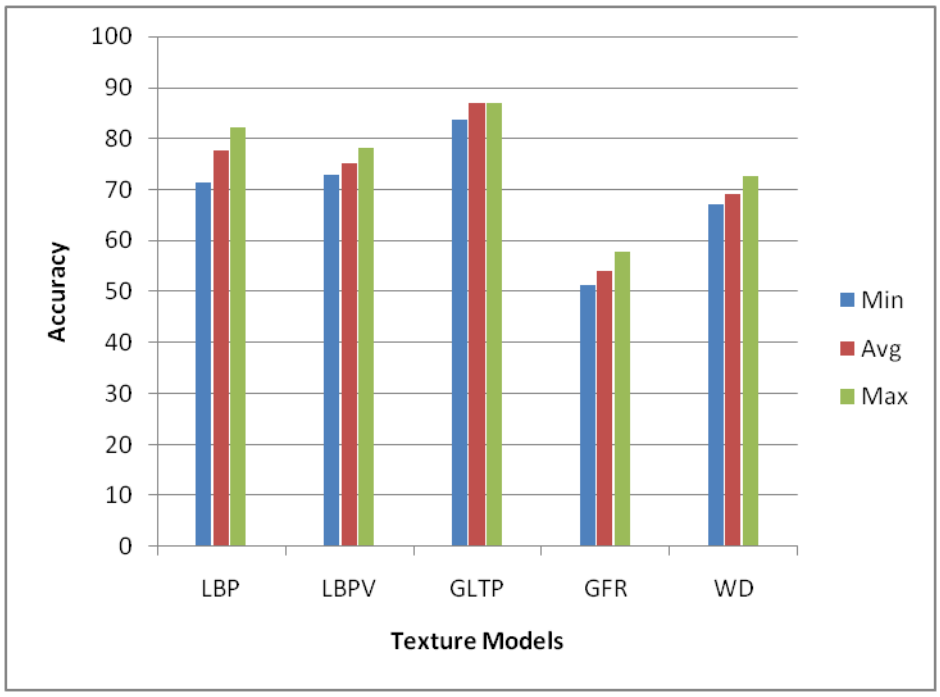


Figure 7: Classification accuracy of the proposed model based on individual texture models for 50% training

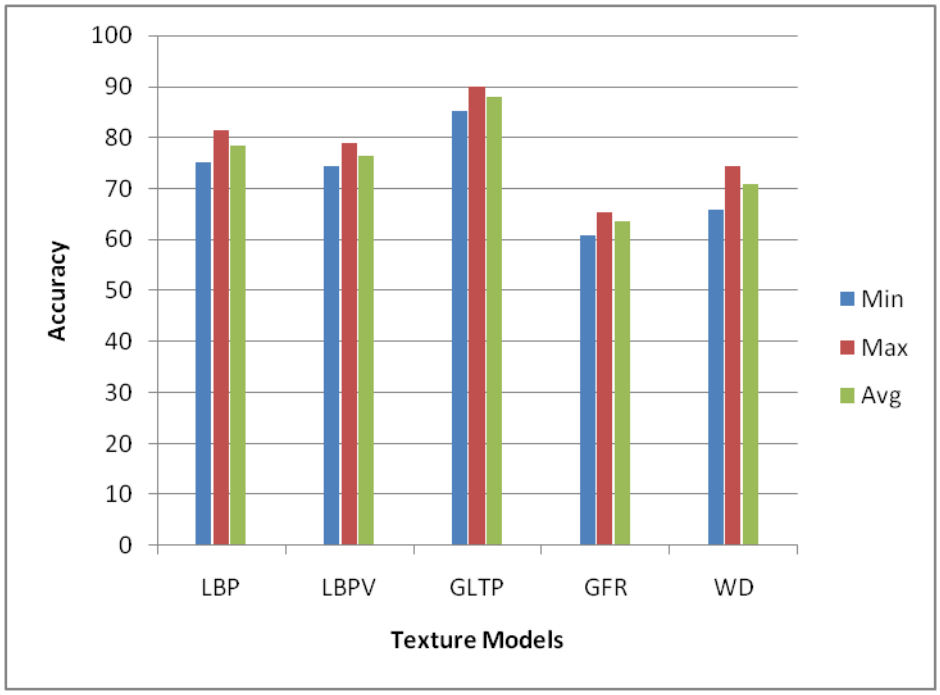


Figure 8: Classification accuracy of the proposed model based on individual texture models for 60% training

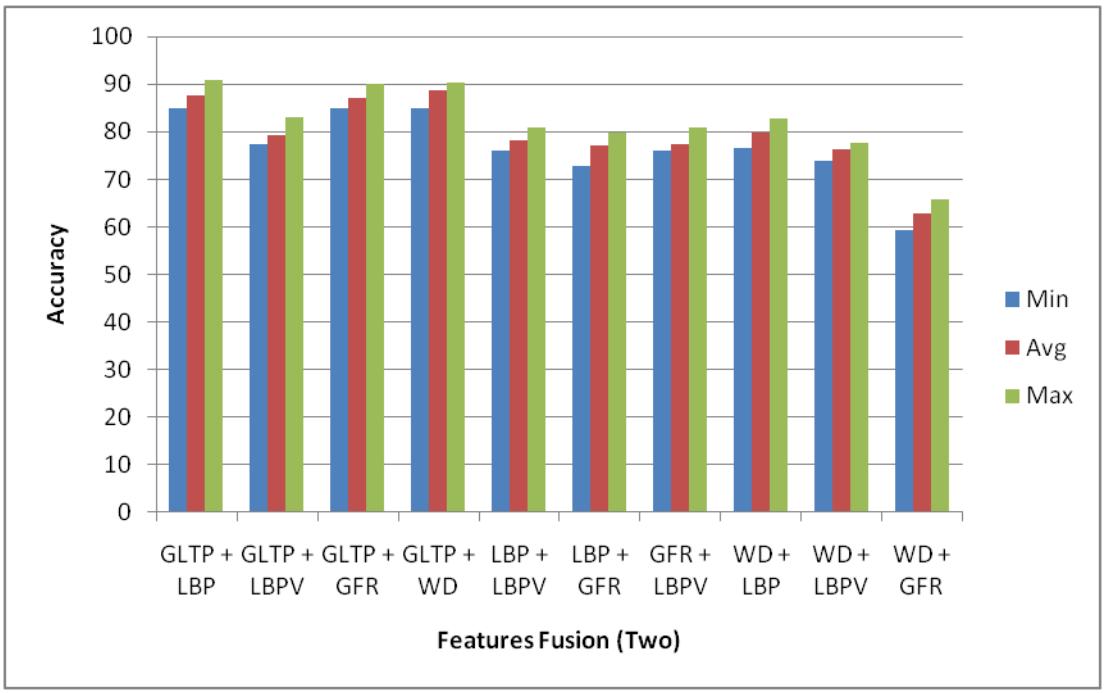


Figure 9: Classification accuracy of the proposed model based on fusion of 2 texture models at a time

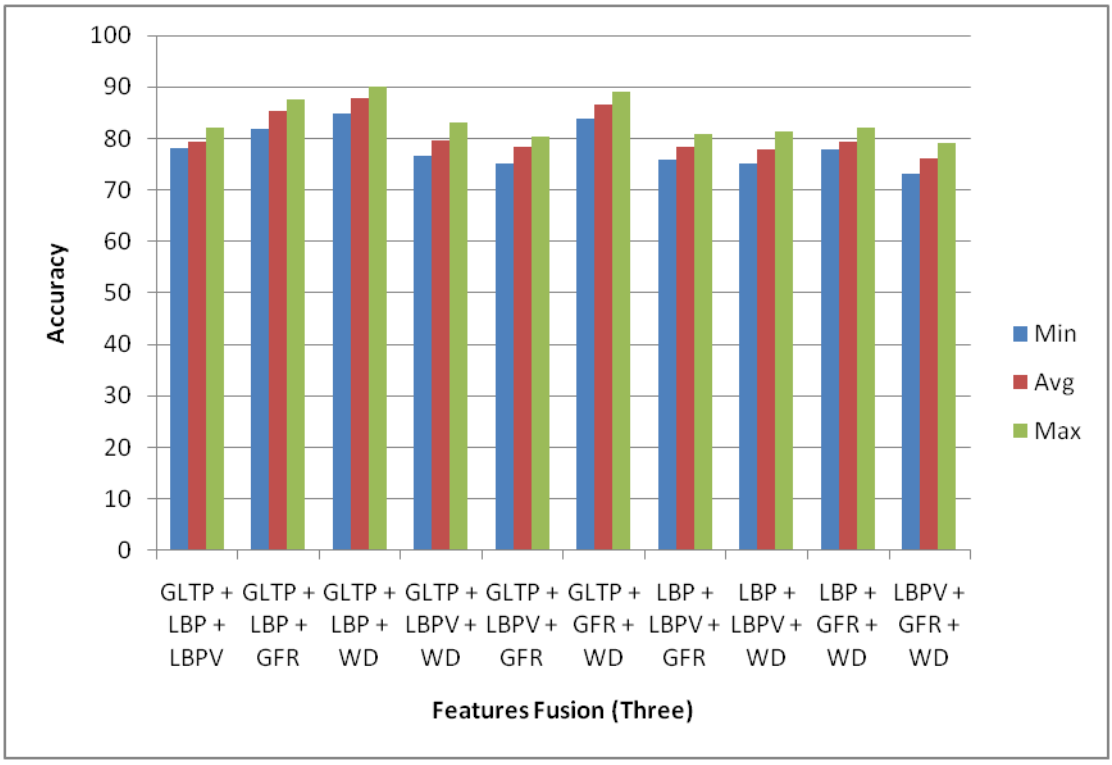


Figure 10: Classification accuracy of the proposed model based on fusion of 3 texture models at a time

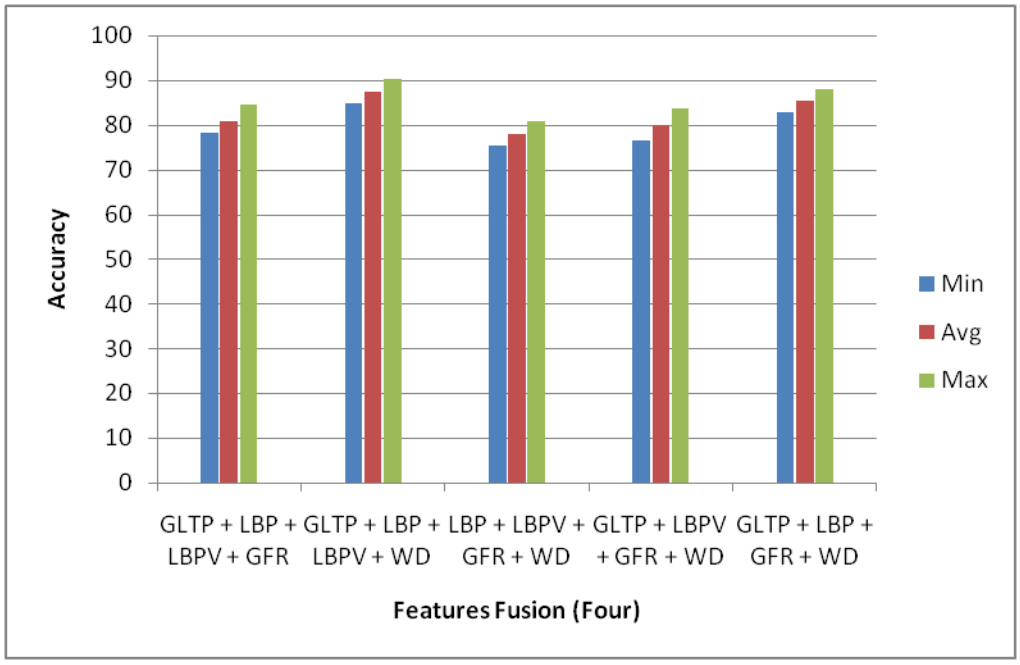


Figure 11: Classification accuracy of the proposed model based on fusion of 4 texture models at a time

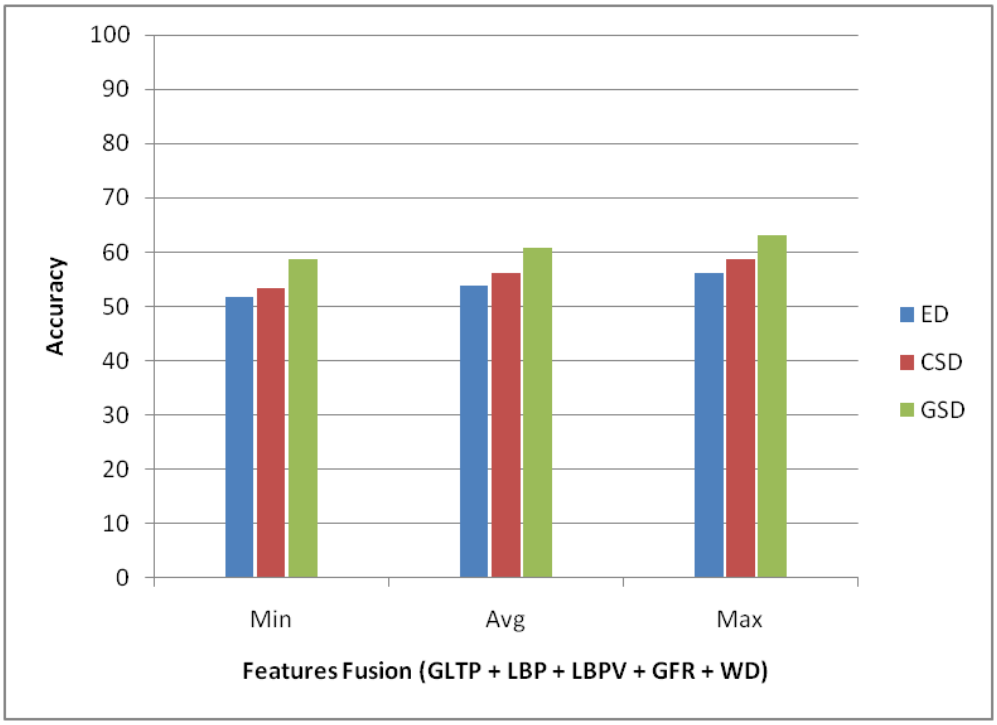


Figure 12: Classification accuracy of the proposed model based on fusion of 5 texture models at a time

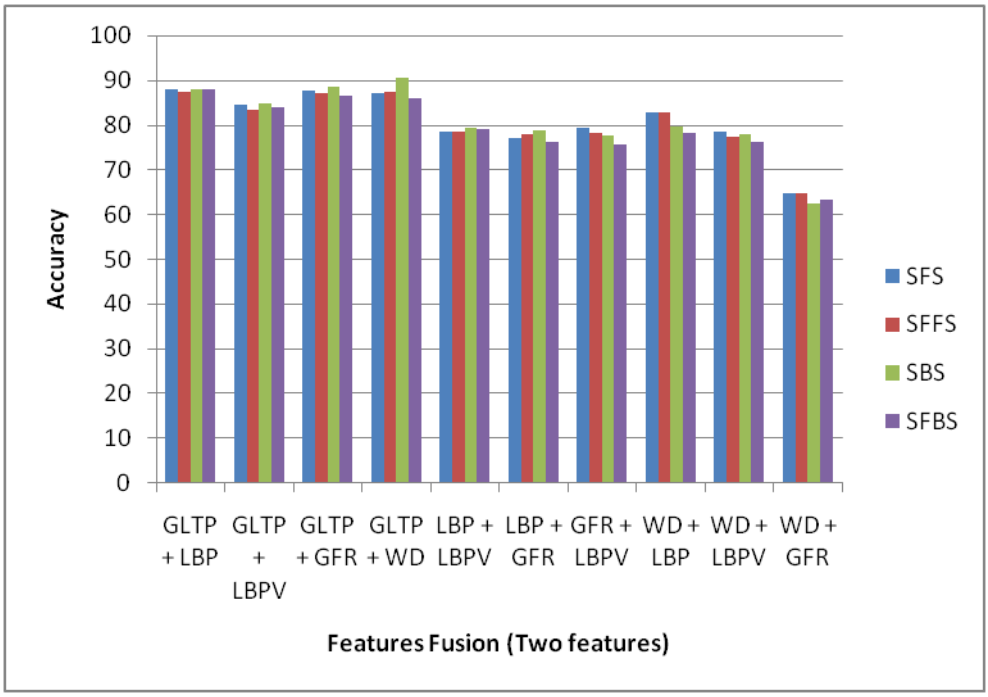


Figure 13: Classification accuracy of the proposed model based on fusion of 2 texture models at a time and feature selection methods

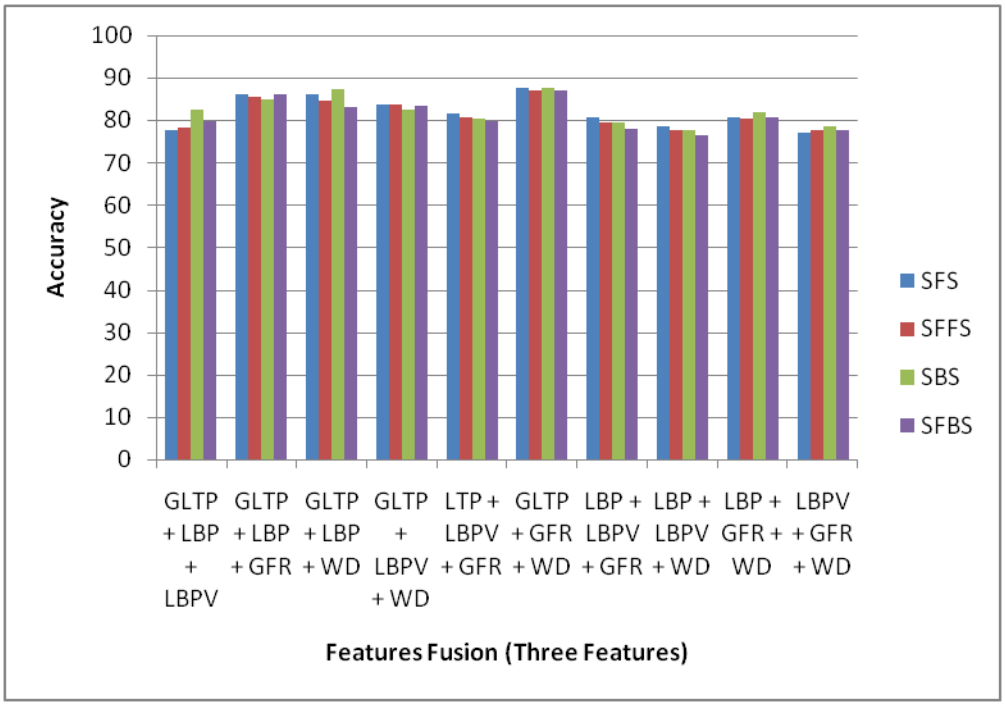


Figure 14: Classification accuracy of the proposed model based on fusion of 3 texture models at a time and feature selection methods

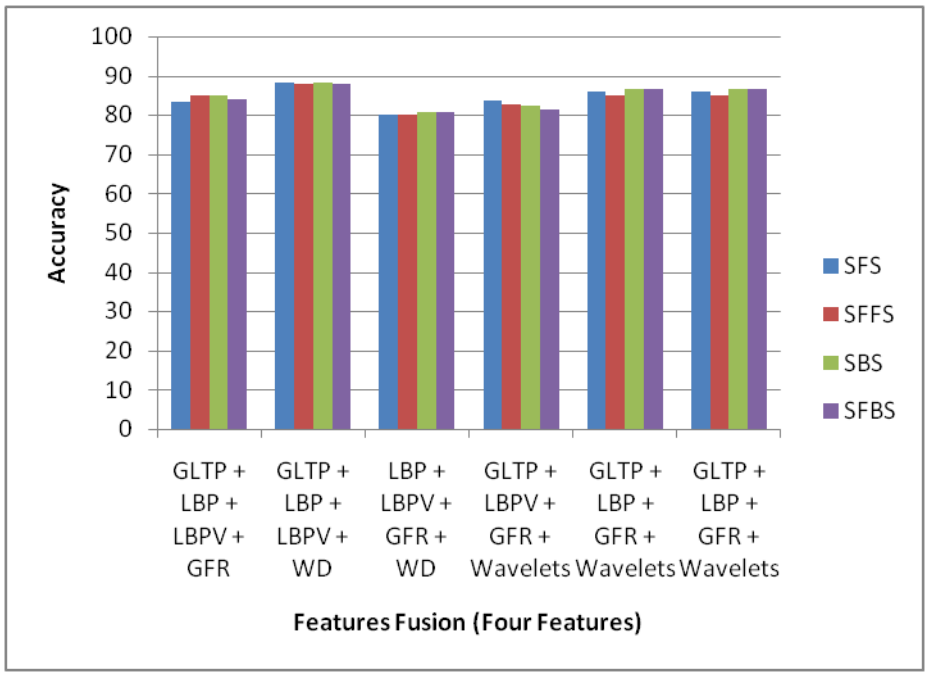


Figure 15: Classification accuracy of the proposed model based on fusion of 4 texture models at a time and feature selection methods

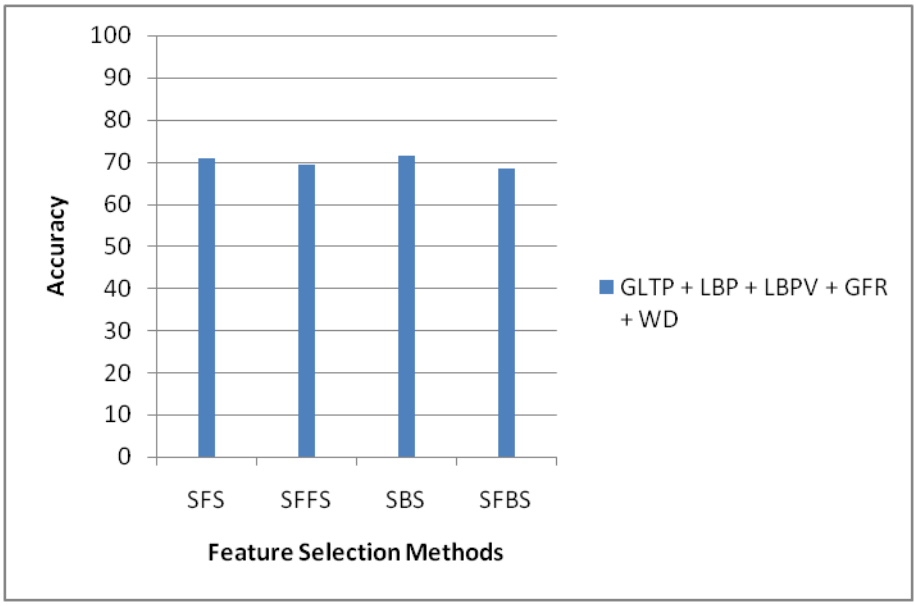


Figure 16: Classification accuracy of the proposed model based on fusion of 5 texture models at a time and feature selection methods

Table 2: Confusion matrix across leaf types using the proposed model based on fusion of 2 texture models (*GLTP* and *WD*) at a time

		Predicted Class		
		Unripe	Ripe	Over-ripe
Actual Class	Unripe	109	20	00
	Ripe	10	240	16
	Over-ripe	00	18	106

Table 3: Performance of the proposed model based on fusion of 2 texture models (*GLTP* and *WD*) at a time

Leaf Class	Precision	Recall	F-measure
Unripe	0.91	0.84	0.87
Ripe	0.86	0.90	0.87
Over-ripe	0.86	0.85	0.85

Table 4: Confusion matrix across leaf types using the proposed model based on fusion of 2 texture models (*GLTP* and *WD*) and *SBS* feature selection method

		Predicted Class		
		Unripe	Ripe	Over-ripe
Actual Class	Unripe	112	17	00
	Ripe	07	250	09
	Over-ripe	00	16	108

Table 5: Performance of the proposed model based on fusion of 2 texture models (*GLTP* and *WD*) and *SBS* feature selection method

Leaf Class	Precision	Recall	F-measure
Unripe	0.94	0.86	0.89
Ripe	0.88	0.93	0.90
Over-ripe	0.92	0.87	0.89

6. Conclusion

In this paper, a model based on texture features for classification of tobacco leaves for the purpose of harvesting is presented. A successful attempt is made to explore the applicability of texture features and wrapper feature selection methods for effective classification of tobacco leaves for the purpose of selective harvesting. The future work is expanding this for video data and developing in a real time environment.

References

Anitha, R., Guru, D. S., Mahesh, K. R. and Sumithra, R. (2020). Hierarchical approach for ripeness grading of mangoes. *Artificial Intelligence in Agriculture*, **4**, 243-252.

Baltazar, A., JorgeIsidro, A. and González-Aguilar Gustavo. (2008). Bayesian classification of ripening stages of tomato fruit using acoustic impact and colorimeter sensor data. *Computers and Electronics in Agriculture*, **60**, 113-121.

Espindola, R. P. and Ebecken, N. F. F. (2005). On extending F-measure and G-mean metrics to multi-class problems. *WIT Transactions on Information and Communication Technologies,* **35**, 25-34.

Furfaro, R., Ganapol, B. D., Johnson, L. F. and Herwitz, S. R. (2007). Neural network algorithm for coffee ripeness evaluation using airborne images. *Applied Engineering in Agriculture*, **23**(**3**), 379-387.

Gao, H., Cai, J. and Liu, X. (2010). Automatic grading of the post-harvest fruit: a review. *Computer and Computing Technologies in Agriculture* III, **371**, 141-146.

Goovaerts, P. (2000). Estimation or simulation of soil properties? An optimization problem with conflicting criteria. *Geoderma*, **3**, 165-186.

Guru, D. S., Mallikarjuna, P. B., Manjunath, S. and Shenoi, M. M. (2012). Machine Vision Based Classification of Tobacco Leaves for Automatic Harvesting. *Intelligent Automation and Soft Computing*, **18(5),** 577 – 586.

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford, USA, 359-366.

He, D. and Wang, L. (1990). Texture unit texture spectrum and texture analysis. *IEEE Transactions on Geosciences and Remote Sensing*, **28(4)**, 509-512.

Jain, A., Nandakumaran, K. and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, **38**, 2270-2285.

Ojala, T., Pietikainen, M. and Maenapaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24(7)**, 971-987.

Johnson, L. F., Herwitz, S. R., Lobitz, B. M. and Dunagan, S. E. (2004). Feasibility of monitoring coffee field ripeness with airborne multispectral imagery. *Applied Engineering in Agriculture*, **20(6)**, 845-849.

Kira, K. and Rendel, L. (1992). The feature selection problem: traditional methods and a new algorithm. *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, USA, 129-134.

Lee, D. J., Chang, Y., Archibald, J. K. and Greco, C. G. (2008). Color quantization and image analysis for automated fruit quality evaluation. *IEEE International Conference on Automation Science and Engineering*, Trieste, Italy, 194-199.

Srinivasan, A. 1999. Precision farming in Asia: Progress and prospects, *Proceedings of Fourth International Conference on Precision Agriculture*. Geospatial Analysis Center, Regional Science Institute, Hokkaido, Japan.

Lee, D., Archibald, J. K. and Xiong, G. (2011). Rapid color grading for fruit quality evaluation using direct color mapping. *IEEE Transactions on Automation Science and Engineering*, **8**, 292-302.

Lee, D., Archibald, J. K., Chang, Y. and Greco, C. R. (2008). Robust color space conversion and color distribution analysis techniques for date maturity evaluation. *Journal of Food Engineering*, **88**, 364-372.

Ojala, T., Pietikainen, M. and Maenapaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24(7)**, 971-987.

Surliandi, A and Ramar, K. (2008). Local Texture Patterns - A univariate texture model for classification of images. *Proceedings of the 16th International Conference on Advanced Computing and Communications (ADCOM08)*, Tamilnadu, India, 32-39.

Ververidis, D. and Kotropoulos, C. (2008). Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. *Signal Processing*, **88(12)**, 2956-2970.

Viscarra, R. A., Minasny, B., Roudier, P. and McBratney, A. B. (2006). Colour space models for soil science. *Geoderma*, **133,** 320-337.

Xu, L., (2009). Strawberry maturity neural network detecting system based on genetic algorithm. *Computer and Computing Technologies in Agriculture II*, **2**, 1201-1208.

Yin, H., Chai, Y., Yang, S. X. and Mitta, G. S. (2009). Ripe tomato extraction for a harvesting robotic system. *IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, USA, 2984-2989.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 51–73

The Transmuted Inverted Nadarajah-Haghighi Distribution: Different Estimation Methods and Applications

R. M. Mandouh
Department of Mathematical Statistics
Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Received: 05 July 2020; Revised: 02 February 2021; Accepted: 05 February 2021

Abstract

In this paper, a new inverted model called the transmuted inverted Nadarajah-Haghighi distribution is introduced. Different estimation methods of the unknown parameters of the new distribution are utilized. These methods are maximum likelihood (MLE), least squares and weighted least squares, maximum product spacing estimation, AD and RAD estimation, CVM estimation and Bayesian estimation. Also, the potentiality of the new model is discussed via a real data set.

Key words: Inverted distributions; Nadarajah-Haghighi distribution; Maximum likelihood estimation (MLE); Least squares and weighted least squares estimation; Maximum product spacing estimation; AD and RAD estimation; CVM estimation; Bayesian estimation.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Two-Parameter Nadarajah-Haghighi (NH) distribution was introduced by Nadarajah and Haghighi (2011) as an extension of exponential distribution and also as an alternative to the gamma, Weibull and exponentiated exponential distributions. They provided three motivations for introducing their distribution, for more details see their paper. Let Z to have Nadarajah-Haghighi distribution, $Z \sim NH(\alpha, \beta)$, then the cdf of Z takes the form

$$F(z) = 1 - e^{1-(1+\beta z)^{\alpha}}, \quad z > 0, \alpha, \beta > 0, \tag{1}$$

where β is the scale parameter and α is the shape parameter. When $\alpha = 1$, the exponential distribution is obtained. Nadarajah and Haghighi (2011) showed that its density can take decreasing and unimodal shapes and the hazard rate can take increasing, constant and decreasing shapes. In order to provide some flexibility, alternative generalizations of the Nadarajah and Haghighi distribution have been proposed. For example, Lemonte *et al.* (2015) introduced the Marshall-Olkin Nadarajah Haghighi distribution via the Marshall-Olkin generator (Marshall and Olkin, 1997). Its cdf is given by

$$F(z) = \frac{1 - e^{1-(1+\beta z)^{\alpha}}}{(1 - (1-\theta)e^{1-(1+\beta z)^{\alpha}})}, \quad z > 0,$$

Corresponding Author: R. M. Mandouh
Email: rshmndoh@cu.edu.eg

where the parameter $\alpha > 0$ and $\theta > 0$ control the shapes of the distribution, and the parameter $\beta > 0$ is the scale parameter. If $\beta = 1$, the NH distribution is obtained. They noted that this distribution is quite flexible and can be used effectively in modeling survival data, reliability problems, fatigue life studies and hydrological data. Also, it can have constant, decreasing, increasing, upside-down bathtub (unimodal), bathtub-shaped and decreasing-increasing-decreasing hazard rate functions.

Yousof and Karkmaz (2017) introduced the Topp-Leone Nadarajah-Haghighi model using the Topp-Leone generated family of distributions (Sangsanit and Bodhisuwan, 2016). If a random variable Z follows the Topp-Leone Nadarajah-Haghighi distribution, then its cdf of Z takes the form

$$F(z) = (1 - e^{2(1-(1+\beta z)^{\alpha})})^{\theta}, \quad z > 0, \alpha, \beta, \theta > 0,$$

They provided some plots of the pdf and hazard rate function for the distribution and showed that its hazard function allows different shapes.

Ogunde *et al.* (2017) introduced transmuted Nadarajah-Haghighi distribution as another generalization of Nadarajah-Haghighi distribution. Its cdf takes the form

$$F(z) = (1 - e^{1-(1+\beta z)^{\alpha}})(1 + \lambda e^{1-(1+\beta z)^{\alpha}}), \quad z > 0, \alpha, \beta > 0 \text{ and } |\lambda| \leq 1.$$

They showed that its hazard function allows different shapes such as decreasing and bathtub shapes.

On the other hand inverted distributions of random variables with positive support provide a valuable alternative for the regular distributions when the assumptions for the use of these distributions are not valid. Also, they may be used in Bayesian analysis of prior and posterior distribution of some parameters such as the scale parameter. Sheikh and Ahmed (1987) discussed characteristic features of the hazard functions based on this inverted class of distributions and explored their possible uses. Hazard functions and mean residual life of inverted normal inverted Gamma and inverted Weibull are compared with the normal, Gamma and Weibull hazards. For a general discussion of inverted distributions, see Folks (1983), Lehmann and Shaffer (1988) and Habibullah and Ahmed (2006).

Some authors discussed the inverse transformation method of baseline variables to obtain inverted distributions due to its usefulness to explore additional properties of the phonomenons which non inverted distributions cannot. Some of these distributions are: inverse exponential distribution (Keller and Kamath in 1982), inverse Rayleigh distribution (Voda in 1972), inverse Lindley distribution (Sharma *et al.*, 2015), inverted Nadarajah-Haghighi (Tahir *et al.*, 2018), inverse xgamma (Yadav *et al.*, 2019) etc.

Furthermore, some authors used the quadratic rank transmutation map (QRTM) approach to generate a generalization of an inverted distribution such as: Mahmoud and Mandouh (2013); Elbatal (2013); Khan (2019) ect. According to this approach, a random variable Z is said to have a transmuted distribution if its cumulative distribution function (cdf) satisfies the following relationship:

$$G(x) = (1 + \lambda)F(x) - \lambda F(x)^2, |\lambda| \leq 1, \tag{2}$$

where F(x) is the cdf of the baseline model and the corresponding probability density function takes the form:

$$g(x) = f(x)[(1+\lambda) - 2\lambda F(x)], |\lambda| \leq 1$$

(see Shaw and Buckley (2009)). The same approach has been used to introduce the transmuted form of inverted Nadarajah-Haghighi distribution.

2. The Transmuted Inverted N-H Distribution

Let the random variable X=1/Z, where Z follows the NH distribution whose cdf is given in (1), then cdf of the inverted N-H distribution takes the form

$$F(x) = e^{1-(1+\beta x^{-1})^{\alpha}}, \quad x > 0, \alpha, \beta > 0.$$

Using (2) and taking the inverted N-H distribution as the base distribution, one can generate the cdf of the transmuted inverted N-H (TINH) distribution as follows

$$G(x) = e^{1-(1+\beta x^{-1})^{\alpha}}(1+\lambda-\lambda e^{1-(1+\beta x^{-1})^{\alpha}}), \quad x > 0, \alpha, \beta > 0 \text{ and } |\lambda| \leq 1. \tag{3}$$

The corresponding pdf and hazard function (failure rate function) are given respectively

$$g(x) = \alpha\beta x^{-2}(1+\beta x^{-1})^{\alpha-1} e^{1-(1+\beta x^{-1})^{\alpha}}(1+\lambda-\lambda e^{1-(1+\beta x^{-1})^{\alpha}}), \tag{4}$$

and

$$h(x) = \frac{\alpha\beta x^{-2}(1+\beta x^{-1})^{\alpha-1} e^{1-(1+\beta x^{-1})^{\alpha}}(1+\lambda-\lambda e^{1-(1+\beta x^{-1})^{\alpha}})}{1-e^{1-(1+\beta x^{-1})^{\alpha}}(1+\lambda-\lambda e^{1-(1+\beta x^{-1})^{\alpha}})}. \tag{5}$$

The new distribution is flexible to model positive real data sets which display decreasing and upside-down bathtub (UBT) hazard rate shapes. Some plots of density and hazard functions are displayed in Figures (1) and (2) for different values of the parameters. In Figure (1), the plots indicate that the TINH density can be decreasing and unimodal. The plots in Figure (2) show that the TINH hazard function can be decreasing and UBT. The new distribution has no finite moments.

The inverse of the cumulative function (3) yields the following quantile function

$$Q(u) = \beta((1 - ln(\frac{1+\lambda-\sqrt{(1+\lambda)^2-4\lambda u}}{2\lambda}))^{1/\alpha} - 1)^{-1}, \quad u \in (0,1) \tag{6}$$

The specification of a distribution through its quantile function takes away the need to describe a distribution through its moments. The following alternative measures in terms of quantiles that reduce the shortcomings of the moment-based ones:
The median as a measure of location is defined by

$$M = Q(0.5) = \beta((1 - ln(\frac{1+\lambda-\sqrt{(1+\lambda)^2-2\lambda}}{2\lambda}))^{1/\alpha} - 1)^{-1}.$$

The interquartile range as a measure of dispersion is defined by

$$IQR = Q_3 - Q_1 = Q(0.75) - Q(0.25).$$

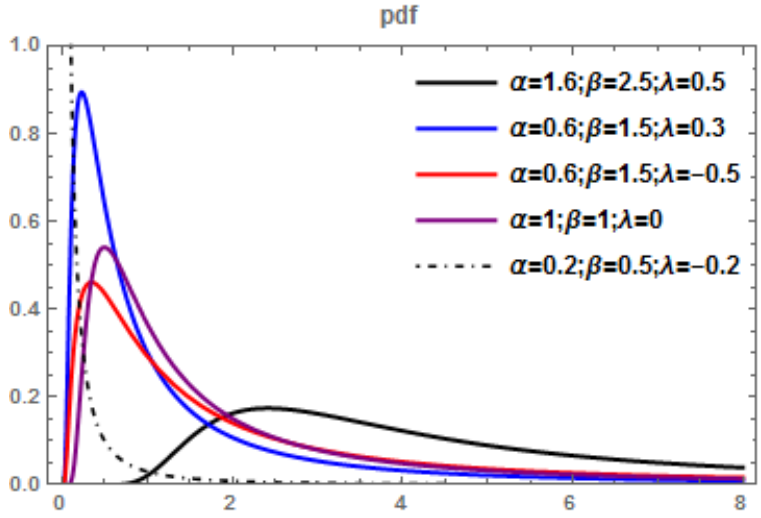


Figure 1: Plots of the TINH density for different parameter values

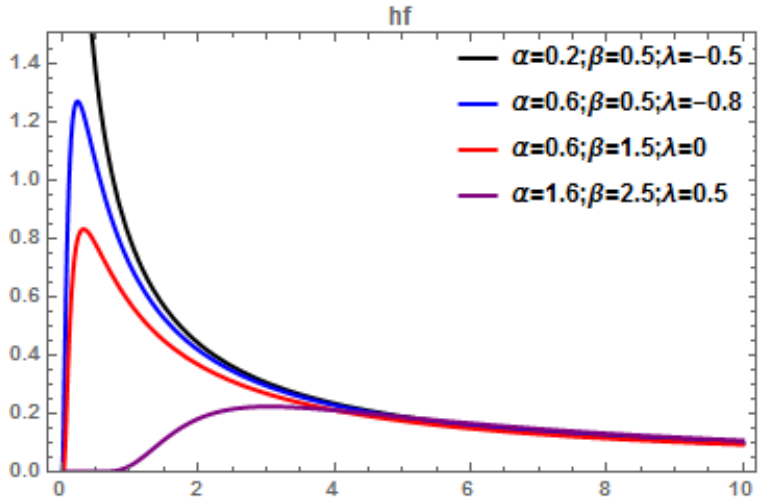


Figure 2: Plots of the TINH hazard rate for different parameter values

Skewness is measured by Galton's coefficient

$$Sk = \frac{Q_3 + Q_1 - 2M}{Q3 - Q1}. \tag{7}$$

Moors, 1988 proposed the measure

$$Ku = (Q(0.875) - Q(0.625) + Q(0.375) - Q(0.125))/IQR \tag{8}$$

as a measure of kurtosis. Given the form of $Q(u)$, the calculations of all the coefficients are very simple, as one needs to only substitute the appropriate fractions for u. For example, one can use formulas (6), (7) and (8) to calculate the skewness and kurtosis for the TINH distribution. Table 1 shows the skewness and kurtosis of the TINH distribution for different values of parameters. One can note that for fixed values of α, the skewness and kurtosis decrease as λ approaching to 1 and for fixed values of λ, the skewness and kurtosis decrease as α increases. Also, for generating random numbers from the TINH distribution, one can

use formula (6). Quantile functions have several interesting properties that are not shared by distributions, which makes it more convenient for analysis. For more details see Nair *et al.*,2013.

Table 1: Skewness and Kurtosis of the TINH distribution for some values of parameters

Sk	$\alpha\|\lambda$	-1	-0.8	-0.4	0.4	0.8	1
	0.5	0.4873	0.4889	0.4988	0.4827	0.4171	0.3770
	0.8	0.4788	0.4778	0.4810	0.4518	0.3807	0.3388
	1.0	0.4763	0.4751	0.4770	0.4443	0.3719	0.3295
	1.5	0.4742	0.4725	0.4721	0.4368	0.3629	0.3201
	3.0	0.4729	0.4710	0.4706	0.4322	0.3575	0.3144
	5.0	0.4727	0.4706	0.4700	0.4312	0.3563	0.3132
Ku	$\alpha\|\lambda$	-1	-0.8	-0.4	0.4	0.8	1
	0.5	2.1593	2.1621	2.1766	2.1347	1.8452	1.6651
	0.8	2.1450	2.1445	2.1481	2.0693	1.7814	1.6109
	1.5	2.1417	2.1406	2.1417	2.0547	1.7675	1.5993
	1.5	2.1386	2.1367	2.1354	2.0405	1.7541	1.5882
	3.0	2.1367	2.1344	2.1317	2.0321	1.7463	1.5817
	5.0	2.1363	2.1339	2.1309	2.0304	1.7446	1.5804

Note: neither skewness nor kurtosis of the TINH distribution depends on the value of β.

3. Non-Bayesian Estimation Methods

Here, we use different methods for estimating the parameters of the TINH distribution (α, β and λ). These methods are maximum likelihood estimation (MLE), least squares and weighted least squares estimation, maximum product spacing estimation, AD and RAD estimation and CVM estimation.

3.1. Maximum likelihood estimation (mle)

The mle is the most popular technique for obtaining estimators and it has desirable properties such as constructing confidence intervals. Now, we consider $X_1, X_2, \ldots, X_n$ as a random sample from TINH distribution, defined in (3), with observed values $x_1, x_2, \ldots, x_n$. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (\alpha, \beta, \lambda)^t$ can be expressed by

$$\begin{aligned} l &= nln(\alpha) + nln(\beta) - 2\sum_{i=1}^{n} lnx_i + (\alpha-1)\sum_{i=1}^{n} ln(1+\beta x_i^{-1}) + \sum_{i=1}^{n}(1-(1+\beta x_i^{-1})^{\alpha}) \\ &+ \sum_{i=1}^{n} ln(1+\lambda - 2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})}). \end{aligned} \tag{9}$$

The components of the score vector $U(\boldsymbol{\theta})$ take the forms

$$\begin{aligned} U_\alpha &= n/\alpha + \sum_{i=1}^{n} ln(1+\beta x_i^{-1}) + \sum_{i=1}^{n}(1+\beta x_i^{-1})^{\alpha} ln(1+\beta x_i^{-1}) \\ &+ \sum_{i=1}^{n} \frac{2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})}(1+\beta x_i^{-1})^{\alpha}) ln(1+\beta x_i^{-1})}{(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})}, \\ U_\beta &= n/\beta + (\alpha-1)\sum_{i=1}^{n} x_i^{-1} ln(1+\beta x_i^{-1}) - \alpha \sum_{i=1}^{n} x_i^{-1}(1+\beta x_i^{-1})^{\alpha-1} \\ &+ \sum_{i=1}^{n} \frac{2\lambda\alpha x_i^{-1} e^{(1-(1+\beta x_i^{-1})^{\alpha})}(1+\beta x_i^{-1})^{\alpha-1}}{(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})}, \\ U_\lambda &= \frac{1-2e^{(1-(1+\beta x_i^{-1})^{\alpha})}}{(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})}. \end{aligned} \tag{10}$$

Equating formulas in (10) to zero and solving them simultaneously yield the mle estimates of the unknown parameters. To construct confidence interval of the model parameter, this requires the 3×3 observed information matrix $J(\boldsymbol{\theta}) = -J_{sk}$, for $s, k = \alpha, \beta, \lambda$ and $\boldsymbol{\theta} = (\alpha, \beta, \lambda)^t$, whose elements are obtained by taking the second derivative of (9). In the observed information matrix, we replace the model parameters by its mles. Maximum likelihood estimation of the model parameters may be difficult to obtain in certain cases-particularly where the support of the model is unknown. Moreover the mle may not be robust to departures from the assumed model. These considerations motivated the following estimation methods described below.

3.2. Minimum distance estimation

In this section, we use some methods of estimation based on minimum distance between the cdf of TINH distribution and the empirical cdf. These methods are divided into two approaches; the first group is known as least-square approach; the second group is related to goodness-of-fit statistics.

3.2.1.Least-square approach (LSE)

Swain *et al.* (1988) used least-square approach to parameter estimation to summarize a set of data by a distribution function in Johnson's translation system. They investigated this approach via minimizing the distance between the vector of "uniformized" order statistics and the corresponding vector of expected values. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n with the cdf $F(.)$ in (3) and let $x_{(1:n)} < x_{(2:n)} < \cdots < x_{(n:n)}$ be the ordered observations. The LSEs of α, β and λ, say $\hat{\alpha}_{LSE}, \hat{\beta}_{LSE}$ and $\hat{\lambda}_{LSE}$, can be obtained by minimizing the following formula with respect to α, β, λ.

$$Dls(\alpha, \beta, \lambda) = \sum_{i=1}^{n} (F(x_{(i:n)}; \alpha, \beta, \lambda) - \frac{i}{n+1})^2$$

Also, one can determine these estimators by solving

$$\sum_{i=1}^{n} (F(x_{i:n}; \alpha, \beta, \lambda) - \frac{i}{n+1}) \rho_1(x_{i:n}; \alpha, \beta, \lambda) = 0,$$

$$\sum_{i=1}^{n}(F(x_{i:n};\alpha,\beta,\lambda)-\frac{i}{n+1})\rho_2(x_{i:n};\alpha,\beta,\lambda)=0,$$

and

$$\sum_{i=1}^{n}(F(x_{i:n};\alpha,\beta,\lambda)-\frac{i}{n+1})\rho_3(x_{i:n};\alpha,\beta,\lambda)=0,$$

where

$$\begin{aligned}\rho_1(x_{i:n};\alpha,\beta,\lambda) &= (1+\beta x^{-1})^{\alpha} ln(1+\beta x^{-1}) e^{1-(1+\beta x^{-1})^{\alpha}} \\ &* \quad (-(1+\lambda)+2\lambda e^{1-(1+\beta x^{-1})^{\alpha}}), \qquad (11)\\ \rho_2(x_{i:n};\alpha,\beta,\lambda) &= \alpha(1+\beta x^{-1})^{\alpha-1}x^{-1}e^{1-(1+\beta x^{-1})^{\alpha}}(-(1+\lambda)+2\lambda e^{1-(1+\beta x^{-1})^{\alpha}}), \qquad (12)\end{aligned}$$

and

$$\rho_3(x_{i:n};\alpha,\beta,\lambda) = e^{1-(1+\beta x^{-1})^{\alpha}}(1-e^{1-(1+\beta x^{-1})^{\alpha}}). \qquad (13)$$

Weighted least-square estimators, $\hat{\alpha}_{WLSE}, \hat{\beta}_{WLSE}$ and $\hat{\lambda}_{WLSE}$, can be determined by minimizing (see Tahir *et al.*, 2018)

$$W(\alpha,\beta,\lambda)=\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}(F(x_{(i:n)};\alpha,\beta,\lambda)-\frac{i}{n+1})^2$$

Also, one can obtain these estimators by solving

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}(F(x_{i:n};\alpha,\beta,\lambda)-\frac{i}{n+1})\rho_1(x_{i:n};\alpha,\beta,\lambda)=0,$$

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}(F(x_{i:n};\alpha,\beta,\lambda)-\frac{i}{n+1})\rho_2(x_{i:n};\alpha,\beta,\lambda)=0,$$

and

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}(F(x_{i:n};\alpha,\beta,\lambda)-\frac{i}{n+1})\rho_3(x_{i:n};\alpha,\beta,\lambda)=0,$$

where $\rho_i(.;\alpha,\beta,\lambda), i=1,2,3$ are given by (11)-(13).

3.2.2.The approach based on the goodness-of-fit statistics

Anderson and Darling (1952) proposed a general class of tests for testing the goodness of fit of a sample of n observations to a specified continuous distribution function $F(x)$. Their test was based on the difference between the specified distribution and the empirical distribution $F_n(x)$ of the sample. From the following measure

$$W_n^2 = n\int_{-\infty}^{\infty}(F_n(x)-GF(x))^2\psi(F(x))dF, \qquad (14)$$

where $\psi(t)(\geq 0)$ is some preassigned weight function., when $\psi(t)=1, W_n^2$ reduces to nw^2, where w^2 is the Cramér-von-Mises test statistic (see Anderson and Darling (1954). Put $\psi(t)=1/t(1-t)$ in (14) W_n^2 reduces to the statistic A_n^2 which was studied by (Anderson and Darling (1952, 1954)). Here, we estimate the parameters of TINH distribution based

on minimization of the goodness-of-fit statistics with respect to α, β and λ. These statistics are Cramér-von-Mises; Anderson-Darling. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n with the cdf G(.) in (3) and let $x_{(1:n)} < x_{(2:n)} < \cdots < x_{(n:n)}$ be the ordered observations. After computing the last integration, the formulae of the two statistics will be obtained in (15) and (16).

Cramér-von Mises (CVM) estimation

The CVM estimators of α, β and λ, say $\hat{\alpha}_{CVM}, \hat{\beta}_{CVM}$ and $\hat{\lambda}_{CVM}$, can be obtained by minimizing the following formula with respect to α, β, λ.(see MacDonald, 1971)

$$CM(\alpha, \beta, \lambda) = \frac{1}{12n} + \sum_{i=1}^{n}(F(x_{(i:n)}; \alpha, \beta, \lambda) - \frac{2i-1}{2n})^2. \tag{15}$$

Also, one can obtain these estimators by solving the following non-linear equations

$$\sum_{i=1}^{n}(F(x_{i:n}; \alpha, \beta, \lambda) - \frac{2i-1}{2n})\rho_1(x_{i:n}; \alpha, \beta, \lambda) = 0,$$

$$\sum_{i=1}^{n}(F(x_{i:n}; \alpha, \beta, \lambda) - \frac{2i-1}{2n})\rho_2(x_{i:n}; \alpha, \beta, \lambda) = 0,$$

and

$$\sum_{i=1}^{n}(F(x_{i:n}; \alpha, \beta, \lambda) - \frac{2i-1}{2n})\rho_3(x_{i:n}; \alpha, \beta, \lambda) = 0,$$

where $\rho_i(.; \alpha, \beta, \lambda), i = 1, 2, 3$ are given by (11)-(13).

Anderson-Darling estimation

The AD estimators of α, β and λ, say $\hat{\alpha}_{AD}, \hat{\beta}_{AD}$ and $\hat{\lambda}_{AD}$, can be obtained by minimizing the following formula with respect to α, β, λ.(see MacDonald, 1971)

$$AD(\alpha, \beta, \lambda) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(log(F(x_{i:n}; \alpha, \beta, \lambda)) + log(\bar{F}(x_{n+1-i:n}; \alpha, \beta, \lambda)), \tag{16}$$

where $\bar{F}(x) = 1 - F(x)$. Also, one can obtain these estimators by solving the following non-linear equations

$$\sum_{i=1}^{n}(2i-1)(\frac{\rho_1(x_{i:n}; \alpha, \beta, \lambda)}{F(x_{i:n}; \alpha, \beta, \lambda)} - \frac{\rho_1(x_{n+1-i:n}; \alpha, \beta, \lambda)}{\bar{F}(x_{n+1-i:n}; \alpha, \beta, \lambda)}) = 0,$$

$$\sum_{i=1}^{n}(2i-1)(\frac{\rho_2(x_{i:n}; \alpha, \beta, \lambda)}{F(x_{i:n}; \alpha, \beta, \lambda)} - \frac{\rho_2(x_{n+1-i:n}; \alpha, \beta, \lambda)}{\bar{F}(x_{n+1-i:n}; \alpha, \beta, \lambda)}) = 0,$$

and

$$\sum_{i=1}^{n}(2i-1)(\frac{\rho_3(x_{i:n}; \alpha, \beta, \lambda)}{F(x_{i:n}; \alpha, \beta, \lambda)} - \frac{\rho_3(x_{n+1-i:n}; \alpha, \beta, \lambda)}{\bar{F}(x_{n+1-i:n}; \alpha, \beta, \lambda)}) = 0.$$

where $\rho_i(.;\alpha,\beta,\lambda), i=1,2,3$ are given by (11)-(13).

The Right-tail Anderson-Darling (RAD) estimators of α,β and λ, say $\hat{\alpha}_{RAD}, \hat{\beta}_{RAD}$ and $\hat{\lambda}_{RAD}$, can be obtained by minimizing the formula (17) with respect to α,β,λ. (see Tahir *et al.* 2018)

$$RAD(\alpha,\beta,\lambda) = \frac{n}{2} - 2\sum_{i=1}^{n} F(x_{i:n};\alpha,\beta,\lambda) - \frac{1}{n}(2i-1)log(\bar{F}(x_{n+1-i:n};\alpha,\beta,\lambda)), \tag{17}$$

where $\bar{F}(x) = 1 - F(x)$. Also, these estimators can be obtained by solving the following non-linear equations

$$-2\sum_{i=1}^{n}\rho_1(x_{i:n};\alpha,\beta,\lambda) + \frac{1}{n}\sum_{i=1}^{n}(2i-1)\frac{\rho_1(x_{n+1-i:n};\alpha,\beta,\lambda)}{\bar{F}(x_{n+1-i:n};\alpha,\beta,\lambda)} = 0,$$

$$-2\sum_{i=1}^{n}\rho_2(x_{i:n};\alpha,\beta,\lambda) + \frac{1}{n}\sum_{i=1}^{n}(2i-1)\frac{\rho_2(x_{n+1-i:n};\alpha,\beta,\lambda)}{\bar{F}(x_{n+1-i:n};\alpha,\beta,\lambda)} = 0,$$

and

$$-2\sum_{i=1}^{n}\rho_3(x_{i:n};\alpha,\beta,\lambda) + \frac{1}{n}\sum_{i=1}^{n}(2i-1)\frac{\rho_3(x_{n+1-i:n};\alpha,\beta,\lambda)}{\bar{F}(x_{n+1-i:n};\alpha,\beta,\lambda)} = 0.$$

where $\rho_i(.;\alpha,\beta,\lambda), i=1,2,3$ are given by (11)-(13).

3.3. Maximum product of spacing (MPS) estimation

This approach was introduced using two methods. The first was by Cheng and Amin (1983) via the idea of spacings. They proposed it as a general method of estimating parameters in continuous univariate distributions. They studied some properties of their approach such as efficiency; consistency and others. Also, they compared it with the mle method via some examples. The second was introduced by Ranneby (1984) who used an approximation of Kullback-Leibler information like the mle method to derive this approach.
Let $x_1, x_2, \ldots, x_n$ be a random sample of size n with the cdf $G(.)$ in (2.1) and let $x_{(1:n)} < x_{(2:n)} < \cdots < x_{(n:n)}$ be the ordered observations. The uniform spacings of the sample is defined as

$$D_i(\alpha,\beta,\lambda) = G(x_{(i:n)};\alpha,\beta,\lambda) - G(x_{(i-1:n)};\alpha,\beta,\lambda), \quad i=1,2,\cdots,n, \tag{18}$$

where $G(x_{(0:n)};\alpha,\beta,\lambda) = 0, G(x_{(n+1:n)};\alpha,\beta,\lambda) = 1$ and $\sum_{i=1}^{n+1} D_i(\alpha,\beta,\lambda) = 1$.

The maximum product of spacings estimators $\hat{\alpha}_{MPS}, \hat{\beta}_{MPS}$, and $\hat{\lambda}_{MPS}$ of the parameters α,β and λ are obtained by maximizing the geometric mean of the spacings with respect to α,β and λ , i.e. maximizing $(\prod_{i=1}^{n+1} D_i(\alpha,\beta,\lambda))^{(1/(n+1))}$. Or, equivalently, maximizing the function

$$\mathcal{D}(\alpha,\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1} lnD_i(\alpha,\beta,\lambda).$$

The estimators $\hat{\alpha}_{MPS}, \hat{\beta}_{MPS},$ and $\hat{\lambda}_{MPS}$ of the parameters α, β and λ are obtained by solving the following non-linear equations

$$\frac{\partial}{\partial\alpha}\mathcal{D}(\alpha,\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{1}{D_i(\alpha,\beta,\lambda)}(\rho_1(x_{i:n};\alpha,\beta,\lambda) - \rho_1(x_{i-1:n};\alpha,\beta,\lambda)) = 0,$$

$$\frac{\partial}{\partial\beta}\mathcal{D}(\alpha,\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{1}{D_i(\alpha,\beta,\lambda)}(\rho_2(x_{i:n};\alpha,\beta,\lambda) - \rho_2(x_{i-1:n};\alpha,\beta,\lambda)) = 0,$$

and

$$\frac{\partial}{\partial\lambda}\mathcal{D}(\alpha,\beta,\lambda) = \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{1}{D_i(\alpha,\beta,\lambda)}(\rho_3(x_{i:n};\alpha,\beta,\lambda) - \rho_3(x_{i-1:n};\alpha,\beta,\lambda)) = 0,$$

where $\rho_i(.;\alpha,\beta,\lambda), i = 1,2,3$ are given by (11)-(13).

4. Numerical Study for Different Estimation Methods

Now, a numerical study is carried out to compare the performance of the frequentist estimators discussed above. To do this we compute absolute value of relative bias (ARbias), scaled root mean square error (SRMSE), average absolute (D_{abs}) and maximum absolute (D_{max}) differences between the theoretical and empirical distribution function at the estimate values (see Tahir *et al.*, 2018). The formulas of these statistics take the forms:

$$ARbias(\hat{\theta}) = |bias(\hat{\theta})|/\theta, \quad bais(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta),$$

$$SRMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}/\theta, \quad MSE(\hat{\theta}) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta)^2,$$

$$D_{abs} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}|F(x_{ij},\theta) - F(x_{ij},\hat{\theta})|, \quad D_{max} = \frac{1}{m}\sum_{i=1}^{m}\underset{j}{max}|F(x_{ij},\theta) - F(x_{ij},\hat{\theta})|,$$

and $\sum Ranks$ gives the partial sum of the ranks. A superscript indicates the rank of each of the estimators for that metric. For example, Table 2 shows the ARbias of the MLE ($\hat{\alpha}$) as 0.072^7 for n=30. This indicates that the ARbias of ($\hat{\alpha}$)obtained using the method of ML ranks 7th among all other estimators. For different sample sizes (n=30, 50, 100, 150), we generate (m=1000) random samples from TINH distribution with parameters $\alpha = 2, 0.8, \beta = 1.5, 0.5$ and $\lambda = 0.5, -0.5$. The results are reported in Tables 2-5 and one can note that in most cases the ARbias and SRMSE of all estimators decrease when the sample size increases. Also, D_{abs} is smaller than D_{max} for all estimation methods and these statistics are smaller when n increases. According to $\sum Ranks$, CVM and LS are the best compared to the other methods. Although the results are not reported here, we also performed simulation study by taking several different values of λ. The trend of the results are quite similar as reported in Table 2 through 5.

Table 2: Numerical results for $\alpha = 2, \beta = 0.5,$ and $\lambda = 0.5$

n	Estimate	MLE	MPS	CVM	AD	RAD	LS	WLS
30	ARbias(α)	0.0721^{7}	0.0317^{6}	0.0110^{1}	0.0223^{3}	0.0157^{2}	0.0228^{4}	0.0234^{5}
	SRMSE($\hat{\alpha}$)	0.2674^{7}	0.2165^{6}	0.1475^{1}	0.1875^{5}	0.1784^{4}	0.1541^{2}	0.1692^{3}
	ARbias($\hat{\beta}$)	0.0482^{2}	0.1013^{5}	0.0384^{1}	0.1140^{7}	0.1104^{6}	$0.0574^{3.5}$	$0.0574^{3.5}$
	SRMSE($\hat{\beta}$)	0.3528^{7}	0.3312^{6}	0.1951^{1}	0.2923^{5}	0.2700^{4}	0.2089^{2}	0.2142^{3}
	ARbias($\hat{\lambda}$)	0.1519^{6}	0.1671^{7}	0.0569^{3}	0.0846^{4}	0.0980^{5}	0.0401^{2}	0.0055^{1}
	SRMSE($\hat{\lambda}$)	0.5368^{6}	0.6072^{7}	0.4622^{3}	0.4258^{1}	0.4643^{4}	0.4475^{2}	0.4713^{5}
	D_{abs}	0.0364^{7}	0.0193^{4}	0.0141^{3}	0.0215^{5}	0.0232^{6}	0.0096^{2}	0.0096^{1}
	$Dmax$	0.0630^{6}	0.0689^{7}	0.0238^{3}	0.0353^{4}	0.0403^{5}	0.0177^{2}	0.0172^{1}
	$\sum Ranks$	$48^{6.5}$	$48^{6.5}$	16^{1}	34^{4}	36^{5}	19.5^{2}	22.5^{3}
50	ARbias($\hat{\alpha}$)	0.0412^{6}	0.2977^{7}	0.0119^{2}	0.0153^{5}	0.01369^{3}	0.0099^{1}	0.0150^{4}
	SRMSE($\hat{\alpha}$)	0.2311^{7}	0.1874^{6}	01296^{1}	0.1544^{4}	0.15624^{5}	0.1302^{2}	0.1429^{3}
	ARbias($\hat{\beta}$)	0.0605^{6}	0.0881^{7}	0.0175^{1}	0.0774^{2}	0.0793^{3}	$0.0270^{4.5}$	$0.0270^{4.5}$
	SRMSE($\hat{\beta}$)	0.3372^{7}	0.3070^{6}	0.1486^{1}	0.2290^{5}	0.2124^{4}	0.1507^{2}	0.1772^{3}
	ARbias($\hat{\lambda}$)	0.1682^{7}	0.1548^{6}	0.0732^{3}	0.1118^{4}	0.1219^{5}	0.0039^{1}	0.0443^{2}
	SRMSE($\hat{\lambda}$)	0.4760^{6}	0.5103^{7}	0.4140^{5}	0.3786^{1}	0.4139^{4}	0.3931^{2}	0.4130^{3}
	D_{abs}	0.0310^{7}	0.0164^{4}	0.0102^{3}	0.0168^{5}	0.0181^{6}	0.0034^{1}	0.0087^{2}
	$Dmax$	0.0729^{7}	0.0680^{6}	0.0322^{3}	0.0488^{4}	0.0532^{5}	0.0059^{1}	0.0193^{2}
	$\sum Ranks$	53^{7}	49^{6}	19^{2}	30^{4}	35^{5}	14.5^{1}	23.5^{3}
100	ARbias(α)	0.0397^{7}	0.0130^{3}	0.0040^{1}	0.0217^{4}	0.03246	0.0096^{2}	0.0223^{5}
	SRMSE(α)	0.2055^{7}	0.1472^{6}	0.1013^{1}	0.1261^{4}	0.1290^{5}	0.1061^{2}	0.1103^{3}
	ARbias(β)	0.0327^{4}	0.0450^{5}	0.0106^{3}	0.0696^{6}	0.0848^{7}	0.0192^{1}	0.0192^{2}
	SRMSE(β)	0.2938^{7}	0.2233^{6}	0.1132^{1}	0.1978^{5}	0.1792^{4}	0.1227^{2}	0.1408^{3}
	ARbias(λ)	0.1176^{7}	0.0967^{4}	0.0429^{3}	0.1063^{6}	0.1034^{5}	0.0044^{1}	0.0639^{3}
	SRMSE(λ)	0.3691^{6}	0.3979^{7}	0.3383^{5}	0.2991^{1}	0.3294^{4}	0.3243^{2}	0.3282^{3}
	D_{abs}	0.0230^{6}	0.1010^{7}	0.0053^{2}	0.0132^{5}	0.0131^{4}	0.0018^{1}	0.0078^{3}
	$Dmax$	0.0548^{7}	0.0454^{4}	0.0211^{2}	0.0497^{6}	0.0483^{5}	0.0027^{1}	0.0299^{3}
	$\sum Ranks$	51^{7}	42^{6}	17^{2}	37^{4}	40^{5}	12^{1}	25^{3}
150	ARbias(α)	0.0062^{2}	0.0143^{4}	0.0025^{1}	0.0289^{5}	0.0419^{7}	0.0101^{3}	0.0343^{6}
	SRMSE(α)	0.1871^{7}	0.1256^{6}	0.0856^{2}	0.1105^{4}	0.1130^{5}	0.0031^{1}	0.0989^{3}
	ARbias(β)	0.0670^{5}	0.0377^{4}	0.0129^{1}	0.0729^{6}	0.0932^{7}	$0.0161^{2.5}$	$0.0161^{2.5}$
	SRMSE(β)	0.2859^{7}	0.1943^{5}	0.1030^{1}	0.1948^{6}	0.1664^{4}	0.1062^{2}	0.1285^{3}
	ARbias(λ)	0.0987^{7}	0.0566^{4}	0.0097^{1}	0.0862^{5}	0.0901^{6}	0.0160^{2}	0.0544^{3}
	SRMSE(λ)	0.3151^{6}	0.3302^{7}	0.2798^{5}	0.2327^{1}	0.2716^{4}	0.2737^{3}	0.2615^{2}
	D_{abs}	0.0204^{7}	0.0066^{3}	0.0026^{2}	0.0110^{5}	0.0115^{6}	0.0022^{1}	0.0070^{4}
	$Dmax$	0.0468^{7}	0.0270^{4}	0.0046^{1}	0.0411^{5}	0.0429^{6}	0.0077^{2}	0.0259^{3}
	$\sum Ranks$	48^{7}	$37^{4.5}$	14^{1}	$37^{4.5}$	45^{6}	16.5^{2}	26.5^{3}

Table 3: Numerical results for $\alpha = 2, \beta = 1.5,$ and $\lambda = 0.5$

n	Estimate	MLE	MPS	CVM	AD	RAD	LS	WLS
30	ARbias(α)	0.0715^{7}	0.0370^{2}	0.0208^{1}	0.0487^{4}	0.0473^{3}	0.0610^{6}	0.0526^{5}
	SRMSE(α)	0.2985^{7}	0.2263^{6}	0.1587^{1}	0.2058^{5}	0.1885^{3}	0.1723^{2}	0.1898^{4}
	ARbias(β)	0.0691^{1}	0.1050^{3}	0.0818^{2}	0.1471^{7}	0.1466^{6}	0.1200^{5}	0.1073^{4}
	SRMSE(β)	0.4113^{7}	0.3474^{6}	0.2173^{1}	0.2966^{5}	0.2919^{4}	0.2309^{2}	0.2584^{3}
	ARbias(λ)	0.1613^{6}	0.1702^{7}	0.0995^{4}	0.0887^{3}	0.1195^{5}	0.0135^{2}	0.0131^{1}
	SRMSE(λ)	0.5427^{6}	0.5944^{7}	0.4704^{4}	0.4399^{1}	0.4948^{5}	0.4582^{2}	0.4685^{3}
	D_{abs}	0.0409^{7}	0.0184^{4}	0.0158^{3}	0.0203^{5}	0.0217^{6}	0.0064^{1}	0.0096^{2}
	$Dmax$	0.0676^{6}	0.0700^{7}	0.0367^{4}	0.0360^{3}	0.0483^{5}	0.0108^{1}	0.0162^{2}
	$\sum Ranks$	47^{7}	42^{6}	20^{1}	33^{4}	37^{5}	21^{2}	24^{3}
50	ARbias(α)	0.0602^{7}	0.0367^{2}	0.0277^{1}	0.0468^{4}	0.0499^{5}	0.0576^{6}	0.0414^{3}
	SRMSE(α)	0.2629^{7}	0.2018^{6}	0.1483^{1}	0.1808^{5}	0.1739^{4}	0.1575^{2}	0.1658^{3}
	ARbias(β)	0.0426^{1}	0.0972^{5}	0.0751^{2}	0.1237^{7}	0.1226^{6}	0.09713.5	$0.0971^{3.5}$
	SRMSE(β)	0.3647^{7}	0.3227^{6}	0.1934^{1}	0.2652^{5}	0.2584^{4}	0.2074^{2}	0.2296^{3}
	ARbias(λ)	0.1070^{6}	0.1248^{7}	0.0835^{4}	0.0842^{5}	0.0791^{3}	0.0096^{1}	0.0313^{2}
	SRMSE(λ)	0.4719^{6}	0.5223^{7}	0.4079^{3}	0.3880^{1}	0.4234^{5}	0.3962^{2}	0.4124^{4}
	D_{abs}	0.0304^{7}	0.0153^{5}	0.0118^{3}	0.0161^{6}	0.0147^{4}	0.0046^{1}	0.0087^{2}
	$Dmax$	0.0482^{6}	0.0549^{7}	0.0367^{5}	0.0365^{4}	0.0344^{3}	0.0074^{1}	0.0137^{2}
	$\sum Ranks$	47^{7}	45^{6}	20^{2}	37^{5}	34^{4}	18.5^{1}	22.5^{3}
100	ARbias(α)	0.0387^{2}	0.0428^{3}	0.0342^{1}	0.0487^{6}	0.0516^{7}	0.0482^{5}	0.0448^{4}
	SRMSE(α)	0.2100^{7}	0.1646^{6}	0.1224^{1}	0.1467^{5}	0.1463^{4}	0.1268^{2}	0.1335^{3}
	ARbias(β)	0.0460^{1}	0.0982^{5}	0.0794^{2}	0.1147^{7}	0.1142^{6}	$0.0874^{3.5}$	$0.0874^{3.5}$
	SRMSE(β)	0.3992^{7}	0.2718^{6}	0.1698^{1}	0.2150^{4}	0.2226^{5}	0.1749^{2}	0.1955^{3}
	ARbias(λ)	0.0802^{4}	0.0918^{6}	0.0841^{5}	0.0922^{7}	0.0776^{3}	0.0413^{1}	0.0588^{2}
	SRMSE(λ)	0.3705^{6}	0.4097^{7}	0.3270^{4}	0.3073^{1}	0.3270^{5}	0.3137^{2}	0.3232^{3}
	D_{abs}	0.0221^{7}	0.0124^{4}	0.0108^{3}	0.0140^{6}	0.0125^{5}	0.0066^{1}	0.0095^{2}
	$Dmax$	0.0373^{4}	0.0429^{6}	0.0394^{5}	0.0430^{7}	0.0361^{3}	0.0192^{1}	0.0274^{2}
	$\sum Ranks$	$38^{4.5}$	$43^{6.5}$	22^{2}	$43^{6.5}$	$38^{4.5}$	17.5^{1}	22.5^{3}
150	ARbias(α)	0.0106^{1}	0.0375^{2}	0.0401^{3}	0.0571^{6}	0.0584^{7}	0.0496^{5}	0.0475^{4}
	SRMSE(α)	0.1955^{7}	0.1594^{6}	0.1144^{1}	0.1391^{5}	0.1350^{4}	0.1185^{2}	0.1225^{3}
	ARbias(β)	0.0639^{1}	0.0895^{5}	0.0807^{2}	0.1215^{7}	0.1193^{6}	$0.0860^{3.5}$	$0.0860^{3.5}$
	SRMSE(β)	0.3025^{7}	0.2604^{7}	0.1612^{1}	0.2035^{4}	0.2117^{5}	0.1655^{2}	0.1861^{3}
	ARbias(λ)	0.0601^{5}	0.0554^{4}	0.0491^{3}	0.0740^{7}	0.0622^{6}	0.0187^{1}	0.0425^{2}
	SRMSE(λ)	0.3234^{6}	0.3413^{7}	0.2784^{4}	0.2640^{1}	0.2886^{5}	0.2697^{2}	0.2778^{3}
	D_{abs}	0.0195^{7}	0.0108^{4}	0.0079^{2}	0.0121^{6}	0.0108^{5}	0.0050^{1}	0.0081^{3}
	$Dmax$	0.0297^{6}	0.0260^{4}	0.0233^{3}	0.0351^{7}	0.0295^{5}	0.0088^{1}	0.0201^{2}
	$\sum Ranks$	40^{5}	38^{4}	19^{2}	$43^{6.5}$	$43^{6.5}$	17.5^{1}	23.5^{3}

Table 4: Numerical results for $\alpha = 0.8, \beta = 1.5,$ and$\lambda = 0.5$

n	Estimate	MLE	MPS	CVM	AD	RAD	LS	WLS
30	ARbias(α)	0.0841^{7}	0.0240^{2}	0.0113^{1}	0.0365^{4}	0.0306^{3}	0.0547^{6}	0.0424^{5}
	SRMSE(α)	0.2615^{7}	0.2171^{6}	0.1645^{1}	0.1778^{4}	0.1914^{5}	0.1692^{3}	0.1671^{2}
	ARbias(β)	0.0506^{1}	0.1462^{5}	0.0948^{2}	0.1741^{7}	0.1649^{6}	$0.1292^{3.5}$	$0.1292^{3.5}$
	SRMSE(β)	0.4542^{7}	0.4360^{6}	0.2716^{1}	0.3205^{4}	0.3506^{5}	0.2864^{2}	0.2917^{3}
	ARbias(λ)	0.1479^{6}	0.2263^{7}	0.0943^{3}	0.1258^{5}	0.1156^{4}	0.0251^{2}	0.0228^{1}
	SRMSE(λ)	0.6135^{6}	0.6522^{7}	0.5121^{2}	0.4906^{1}	0.5163^{3}	0.5212^{4}	0.5313^{5}
	D_{abs}	0.0369^{7}	0.0239^{6}	0.0157^{3}	0.0210^{4}	0.0211^{5}	0.0067^{1}	0.0087^{2}
	$Dmax$	0.0612^{6}	0.08894^{7}	0.0376^{3}	0.0495^{5}	0.0454^{4}	0.0118^{1}	0.0135^{2}
	$\sum Ranks$	47^{7}	46^{6}	16^{1}	34^{4}	35^{5}	22.5^{2}	23.5^{3}
50	ARbias(α)	0.0498^{7}	0.0257^{2}	0.0238^{1}	0.0410^{5}	0.0386^{4}	0.0491^{6}	0.0365^{3}
	SRMSE(α)	0.2123^{7}	0.1810^{6}	0.1371^{2}	0.1508^{4}	0.1598^{5}	0.1423^{3}	0.1357^{1}
	ARbias(β)	0.0678^{1}	0.1309^{5}	0.1056^{2}	0.1731^{7}	0.1494^{6}	$0.1216^{3.5}$	$0.1216^{3.5}$
	SRMSE(β)	0.4271^{7}	0.3980^{6}	0.2470^{1}	0.3016^{4}	0.3055^{5}	0.2530^{2}	0.25784^{3}
	ARbias(λ)	0.0980^{5}	0.1297^{7}	0.0837^{4}	0.1207^{6}	0.0750^{3}	0.0001^{1}	0.0269^{2}
	SRMSE(λ)	0.5355^{6}	0.5842^{7}	0.4386^{3}	0.4344^{2}	0.4395^{4}	0.4304^{1}	0.4648^{5}
	D_{abs}	0.0289^{7}	0.0183^{5}	0.0134^{3}	0.0195^{6}	0.0152^{4}	0.0054^{1}	0.0083^{2}
	$Dmax$	0.0445^{5}	0.0563^{7}	0.0363^{4}	0.0519^{6}	0.0320^{3}	0.0103^{1}	0.0124^{2}
	$\sum Ranks$	$45^{6.5}$	$45^{6.5}$	20^{2}	40^{5}	34^{4}	18.5^{1}	21.5^{3}
100	ARbias(α)	0.0247^{1}	0.0309^{2}	0.0339^{3}	0.0448^{6}	0.0427^{5}	0.0477^{7}	0.0361^{4}
	SRMSE(α)	0.1632^{7}	0.1518^{6}	0.1078^{2}	0.1167^{4}	0.1264^{5}	0.1115^{3}	0.1030^{1}
	ARbias(β)	0.0659^{1}	0.1458^{6}	0.1129^{2}	0.1576^{7}	0.1375^{5}	$0.1225^{3.5}$	$0.1225^{3.5}$
	SRMSE(β)	0.3705^{6}	0.3946^{7}	0.2186^{1}	0.2538^{4}	0.2689^{5}	0.2236^{2}	0.2400^{3}
	ARbias(λ)	0.0980^{3}	0.1401^{6}	0.1192^{5}	0.1447^{7}	0.1014^{4}	0.0762^{1}	0.0871^{2}
	SRMSE(λ)	0.4657^{6}	0.5196^{7}	0.3696^{2}	0.3869^{4}	0.3858^{3}	0.3686^{1}	0.4070^{5}
	D_{abs}	0.0216^{7}	0.0192^{6}	0.0127^{3}	0.0169^{5}	0.0131^{4}	0.0085^{1}	0.0104^{2}
	$Dmax$	0.0452^{3}	0.0648^{6}	0.0554^{5}	0.0670^{7}	0.0469^{4}	0.0353^{1}	0.0404^{2}
	$\sum Ranks$	34^{4}	46^{7}	23^{3}	44^{6}	35^{5}	19.5^{1}	22.5^{2}
150	ARbias(α)	0.0021^{1}	0.0426^{3}	0.0451^{4}	0.0574^{7}	0.0454^{5}	0.0536^{6}	0.0412^{2}
	SRMSE(α)	0.1380^{6}	0.1402^{7}	0.1017^{2}	0.1109^{5}	0.1089^{4}	0.1043^{3}	0.0917^{1}
	ARbias(β)	0.1010^{1}	0.1664^{6}	0.1394^{3}	0.1854^{7}	0.1344^{2}	$0.1439^{4.5}$	$0.1439^{4.5}$
	SRMSE(β)	0.3635^{6}	0.3843^{7}	0.2292^{1}	0.2412^{4}	0.2395^{3}	0.2322^{2}	0.2435^{5}
	ARbias(λ)	0.0972^{3}	0.1316^{5}	0.1324^{6}	0.1597^{7}	0.0971^{2}	0.1010^{4}	0.0917^{1}
	SRMSE(λ)	0.4207^{6}	0.4562^{7}	0.3144^{2}	0.3365^{4}	0.3298^{3}	0.3084^{1}	0.3445^{5}
	D_{abs}	0.0194^{7}	0.0183^{6}	0.0138^{4}	0.0174^{5}	0.0117^{3}	0.0108^{2}	0.0106^{1}
	$Dmax$	0.0460^{2}	0.0623^{5}	0.0630^{6}	0.0758^{7}	0.0461^{3}	0.0480^{4}	0.0436^{1}
	$\sum Ranks$	32^{5}	$46^{6.5}$	28^{4}	$46^{6.5}$	25^{2}	26.5^{3}	20.5^{1}

Table 5: Numerical results for $\alpha = 0.8, \beta = 0.5,$ and $\lambda = 0.5$

n	Estimate	MLE	MPS	CVM	AD	RAD	LS	WLS
30	ARbias(α)	0.0603^{7}	0.0356^{5}	0.0027^{1}	0.0271^{3}	0.0223^{2}	0.0413^{6}	0.0314^{4}
	SRMSE(α)	0.2334^{7}	0.2203^{6}	0.1472^{1}	0.1656^{4}	0.1684^{5}	0.1561^{3}	0.1479^{2}
	ARbias(β)	0.0849^{2}	0.2042^{7}	0.0725^{1}	0.1555^{6}	0.1477^{5}	$0.0989^{3.5}$	$0.0989^{3.5}$
	SRMSE(β)	0.4387^{6}	0.5016^{7}	0.2511^{1}	0.3126^{4}	0.3248^{5}	0.2635^{3}	0.2542^{2}
	ARbias(λ)	0.1630^{6}	0.2335^{7}	0.0405^{2}	0.0890^{5}	0.0867^{4}	0.0829^{3}	0.0356^{1}
	SRMSE(λ)	0.5989^{6}	0.6852^{7}	0.4927^{2}	0.4623^{1}	0.5096^{4}	0.4946^{3}	0.5343^{5}
	D_{abs}	0.0365^{7}	0.0291^{6}	0.0128^{3}	0.0200^{4}	0.0202^{5}	0.0111^{2}	0.0087^{1}
	$Dmax$	0.0658^{6}	0.0932^{7}	0.0187^{2}	0.0353^{5}	0.0347^{3}	0.0353^{4}	0.0160^{1}
	$\sum Ranks$	47^{6}	51^{7}	13^{1}	32^{4}	33^{5}	27.5^{3}	19.5^{2}
50	ARbias(α)	0.0428^{7}	0.0385^{6}	0.0128^{1}	0.03106^{4}	0.0274^{3}	0.0361^{5}	0.0244^{2}
	SRMSE(α)	0.1874^{7}	0.1817^{6}	0.1245^{2}	0.1369^{4}	0.1474^{5}	0.1301^{3}	0.1130^{1}
	ARbias(β)	0.0533^{1}	0.1757^{7}	0.0665^{2}	0.1343^{6}	0.1115^{5}	$0.0781^{3.5}$	$0.0781^{3.5}$
	SRMSE(β)	0.3883^{6}	0.4497^{7}	0.2145^{3}	0.2809^{5}	0.2639^{4}	0.2139^{2}	0.2128^{1}
	ARbias(λ)	0.1125^{5}	0.1795^{7}	0.0626^{3}	0.1131^{6}	0.0720^{4}	0.0183^{2}	0.0089^{1}
	SRMSE(λ)	0.5420^{6}	0.6263^{7}	0.43226^{2}	0.4320^{1}	0.4612^{4}	0.4346^{3}	0.4817^{5}
	D_{abs}	0.02515^{7}	0.02252^{6}	0.009755^{3}	0.01655^{5}	0.01277^{4}	0.0039^{2}	0.0038^{1}
	$Dmax$	0.0490^{4}	0.0781^{7}	0.0274^{2}	0.0492^{5}	0.0311^{3}	0.0087^{1}	0.0613^{6}
	$\sum Ranks$	43^{6}	53^{7}	18^{1}	36^{5}	32^{4}	21.5^{3}	20.5^{2}
100	ARbias(α)	0.0076^{1}	0.0486^{6}	0.0244^{2}	0.0494^{7}	0.0393^{5}	0.0354^{4}	0.0325^{3}
	SRMSE(α)	0.1515^{7}	0.1456^{6}	0.0977^{2}	0.1154^{5}	0.1052^{4}	0.1005^{3}	0.0887^{1}
	ARbias(β)	0.1026^{4}	0.1863^{7}	0.0911^{1}	0.1809^{6}	0.1332^{5}	$0.0948^{2.5}$	$0.0948^{2.5}$
	SRMSE(β)	0.3691^{6}	0.3912^{7}	0.1885^{1}	0.2554^{5}	0.2314^{4}	0.1886^{2}	0.1943^{3}
	ARbias(λ)	0.1147^{5}	0.1591^{7}	0.07464^{3}	0.15804^{6}	0.1005^{4}	0.0288^{1}	0.0571^{2}
	SRMSE(λ)	0.4450^{6}	0.5113^{7}	0.3426^{2}	0.3454^{3}	0.3645^{4}	0.3372^{1}	0.3783^{5}
	D_{abs}	0.0231^{7}	0.0203^{6}	0.0106^{3}	0.0193^{5}	0.0135^{4}	0.0062^{1}	0.0086^{2}
	$Dmax$	0.0528^{4}	0.0734^{7}	0.0345^{3}	0.0730^{6}	0.0464^{5}	0.0131^{1}	0.0263^{2}
	$\sum Ranks$	40^{5}	53^{7}	20^{2}	43^{6}	35^{4}	15.5^{1}	20.5^{3}
150	ARbias(α)	0.0028^{1}	0.0367^{4}	0.0303^{3}	0.0544^{7}	0.0416^{6}	0.0384^{5}	0.0294^{2}
	SRMSE(α)	0.1347^{7}	0.1326^{6}	0.0902^{2}	0.1079^{5}	0.0923^{3}	0.0949^{4}	0.0775^{1}
	ARbias(β)	0.1020^{2}	0.1439^{5}	0.0997^{1}	0.1844^{6}	0.1243^{4}	$0.1050^{3.5}$	$0.1050^{3.5}$
	SRMSE(β)	0.3407^{6}	0.3449^{7}	0.1846^{1}	0.2442^{5}	0.2099^{4}	0.1917^{2}	0.1918^{3}
	ARbias(λ)	0.1186^{5}	0.1258^{6}	0.1009^{3}	0.1782^{7}	0.1060^{4}	0.0717^{2}	0.0690^{1}
	SRMSE(λ)	0.4072^{6}	0.4384^{7}	0.2982^{2}	0.3103^{3}	0.3188^{4}	0.2905^{1}	0.3281^{5}
	D_{abs}	0.0201^{7}	0.0167^{5}	0.0112^{3}	0.0189^{6}	0.0117^{4}	0.0085^{2}	0.0081^{1}
	$Dmax$	0.0563^{5}	0.0600^{6}	0.0481^{3}	0.0847^{7}	0.0505^{4}	0.0341^{2}	0.0329^{1}
	$\sum Ranks$	39^{5}	$46^{6.5}$	18^{2}	$46^{6.5}$	33^{4}	21.5^{3}	17.5^{1}

5. Bayesian Estimation

In this section, Bayesian estimation of the three unknown parameters of the TINH distribution will be discussed. Approximate Bayes estimates are computed using the Gibbs sampling procedure with generating samples from the posterior distributions. This requires prior density functions of the unknown parameters (α, β and λ). Here, we assume that α, β and λ

are independent random variables. The parameters α and β have gamma distributions while λ follows uniform distribution. Their pdfs, respectively are $g_1(\alpha) \propto \alpha^{(a-1)}e^{(-b\alpha)}, g_2(\beta) \propto \beta^{(c-1)}e^{(-d\beta)}$ and $g_3(\lambda) =$ constant. The hyper-parameters a, b, c and d are assumed to be known.

The joint prior distribution for (α, β and λ) takes the form $g(\alpha, \beta, \lambda) \propto \alpha^{(a-1)}\beta^{(c-1)}e^{(-b\alpha-d\beta)}$ and the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{x}; \alpha, \beta, \lambda) &\propto \alpha^n \beta^n e^{-\alpha \sum_{i=1}^n ln(1+\beta x_i^{-1})^{-1}} e^{-\sum_{i=1}^n (1+\beta x_i^{-1})} e^{-\alpha \sum_{i=1}^n (1+\beta x_i^{-1})} \\ &* \quad e^{-\sum_{i=1}^n ln(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})^{-1}}. \end{aligned} \tag{19}$$

Then the joint posterior is given by

$$\begin{aligned} g(\alpha, \beta, \lambda|\boldsymbol{x}) &\propto \alpha^{n+a-1}\beta^{n+c-1} e^{-\alpha(b+\sum_{i=1}^n ln(1+\beta x_i^{-1})^{-1})} e^{-(d\beta+\sum_{i=1}^n (1+\beta x_i^{-1}))} e^{-\alpha \sum_{i=1}^n (1+\beta x_i^{-1})} \\ &* \quad e^{-\sum_{i=1}^n ln(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})^{-1}}. \end{aligned} \tag{20}$$

The conditional posterior distributions used in the Gibbs sampling algorithm are given by

$$\begin{aligned} g(\alpha|\beta, \lambda, \boldsymbol{x}) &\propto \alpha^{n+a-1} e^{-\alpha(b+\sum_{i=1}^n ln(1+\beta x_i^{-1})^{-1})} e^{-\alpha \sum_{i=1}^n (1+\beta x_i^{-1})} \\ &* \quad e^{-\sum_{i=1}^n ln(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})^{-1}}, \end{aligned} \tag{21}$$

$$\begin{aligned} g(\beta|\alpha, \lambda, \boldsymbol{x}) &\propto \beta^{n+c-1} e^{-\alpha(b+\sum_{i=1}^n ln(1+\beta x_i^{-1})^{-1})} e^{-(d\beta+\sum_{i=1}^n (1+\beta x_i^{-1}))} e^{-\alpha \sum_{i=1}^n (1+\beta x_i^{-1})} \\ &* \quad e^{-\sum_{i=1}^n ln(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})^{-1}}, \end{aligned} \tag{22}$$

and

$$g(\lambda|\alpha, \beta, \boldsymbol{x}) \propto e^{-\sum_{i=1}^n ln(1+\lambda-2\lambda e^{(1-(1+\beta x_i^{-1})^{\alpha})})^{-1}}. \tag{23}$$

The computation can be achieved using the WinBUGS software which requires only the specification of the joint distribution for the data and the prior distributions for the model parameters. Gibbs sampling algorithm works as follows

1. Specify the size of the samples we wish to generate, say m.
2. Choose an initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$.
3. For iteration i from 1 to m, generate $\theta_j^{(i)}$ from $g(\theta_j|\theta_1^{(i)}, ..., \theta_{(j-1)}^{(i)}, \theta_{(j+1)}^{(i-1)}, ..., \theta_p^{(i-1)})$, for j from 1 to p.
4. Return the values $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(m)}$.

Discarding the early m_0 number of burn-in draws and using the remaining $m - m_0$, $\boldsymbol{\theta}^{(m_0+1)}, \boldsymbol{\theta}^{(m_0+2)}, \ldots, \boldsymbol{\theta}^{(m)}$, as the chosen draws from the joint posterior distribution, the Bayes estimate of θ_j is

$$\hat{\theta}_j = \frac{\sum_{i=m_0+1}^{m-m_0} \theta_j^{(i)}}{m - m_0}, j = 1, 2, 3.$$

Furthermore, the lower and upper bounds of the $100(1-\nu)\%, 0 < \nu < 1$, Bayesian probability interval of θ_j can be obtained using $\nu/2$ 100 th and $(1-\nu/2)$ 100 th percentiles of the sequence of the $m - m_0$ draws; $\theta^{(m_0+1)}, \theta^{(m_0+2)}, \ldots, \theta^{(m)}$.
Here, we generate 10,000 samples of α, β and λ, after a "burn-in-sample" of size 1000 and the approximate Bayes estimates with some posterior summaries, such as MC error, 95% credible interval, median, are given in Table 6. Table 7 has the results of Bayesian estimation for real data set (mentioned in the section 5) and the graphical representation of the marginal posteriors of α, β and λ are displayed in Figure 3. One can note that the posteriors of α and β are approximately normal while skewed for λ. Another MCMC method called Metropolis-Hastings algorithm is used to generate random draws from the joint posterior distribution without deriving its explicit form. Metropolis-Hastings algorithm unlike Gibbs-sampling, it requires a proposal distribution and a common choice of it is the multivariate normal distribution. Metropolis-Hastings algorithm steps are

1. Set the size of the random draws we wish to generate, say m.
2. Choose an initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$.
3. For $i = 1, 2, \ldots, m$, repeat the following steps:
   (a) Set $\theta^{(i)} = \theta^{(i-1)}$.
   (b) Generate a candidate value θ^* from a proposal distribution $p(\theta^{(*)}|\theta^{(i)})$.
   (c) Calculate the ratio $\kappa = min(1, \frac{g(\boldsymbol{\theta}^{(*)}|data)/p(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(i)}}{g(\boldsymbol{\theta}^{(i)}|data)/p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(*)})})$.
   (d) Generate a random value u from uniform distribution on (0, 1).
   (e) Put $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$, if $\kappa \geq u$, otherwise put $\theta^{(i)} = \theta^{(i-1)}$.
4. Return the values $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$.

The lower and upper bounds of the $100(1-\nu)\%$ Bayesian probability interval of θ_j as given above. The computations are carried out using R software. We use the two previously MCMC methods to analyze the same real dataset. We generate 10,000 samples of α, β and λ, after a "burn-in-sample" of size 1000 with assuming gamma priors for α and β and uniform prior for λ. The results of Bayesian estimation for real dataset are given in Tables 7-8. Table 7 has the results of Bayesian estimation for real data set (mentioned in the section 5) and the graphical representation of the marginal posteriors of α, β and λ are displayed in Figure 5. One can note that the posteriors of α and β are approximately normal while skewed for λ. Table 8 displays the posterior mean, median, standard deviation and the limits of a 95% credible interval of each parameter. Figure 6 shows The approximated marginal Posterior density functions of α, β and λ.

Table 6: Summary results for the posterior parameters in the case of the TINH model

Parameter	n	Estimate	SD	MC error	95% Credible Interval	Median
$\alpha = 2$	30	2.0140	0.4468	0.013930	(1.2990, 3.0340)	1.9660
	50	2.3050	0.4715	0.017700	(1.5210,3.3660)	2.2540
	100	1.9060	0.4109	0.021700	(1.2910,2.8780)	1.8400
	150	1.7690	0.3769	0.022720	(1.2200,2.6930)	1.7090
$\beta = 0.5$	30	0.5695	0.1794	0.005541	(0.2963,0.0986)	0.5428
	50	0.5178	0.1397	0.005364	(0.3031,0.8414)	0.5000
	100	0.5614	0.1701	0.008805	(0.2924,0.9551)	0.5417
	150	0.5597	0.1729	0.010350	(0.2776,0.9335)	0.5384
$\lambda = 0.5$	30	0.4009	0.2585	0.004148	(0.0172,0.9304)	0.3726
	50	0.8234	0.1485	0.002417	(0.4459,0.9933)	0.8612
	100	0.3069	0.1925	0.005165	(0.0181,0.7155)	0.2857
	150	0.4439	0.1966	0.007221	(0.0585,0.8048)	0.4533
$\alpha = 2$	30	1.9590	0.3896	0.011880	(1.3330, 2.8670)	1.9090
	50	2.1540	0.3982	0.013320	(1.5060,3.0490)	2.1120
	100	2.2250	0.4046	0.017630	(1.5400,3.1500)	2.1760
	150	2.1170	0.3882	0.019640	(1.4970,2.9960)	2.0550
$\beta = 1.5$	30	1.5260	0.3962	0.012380	(1.4820,2.4280)	2.4280
	50	1.7240	0.4165	0.014080	(1.0370,2.6490)	1.6800
	100	1.4540	0.3508	0.014830	(0.8763,2.2910)	1.4230
	150	1.4660	0.3605	0.018850	(0.8729,2.2600)	1.4400
$\lambda = 0.5$	30	0.7376	0.2166	0.002997	(0.1965,0.9917)	0.9917
	50	0.7676	0.1897	0.003344	(0.2849,0.9909)	0.8152
	100	0.6903	0.1674	0.002574	(0.3211,0.9637)	0.7069
	150	0.3957	0.1627	0.003313	(0.0792,0.7070)	0.3990
$\alpha = 0.8$	30	0.6644	0.1216	0.003085	(0.4721, 0.9483)	0.6484
	50	0.7987	0.1368	0.003655	(0.5742,1.0200)	0.7822
	100	0.7989	0.1256	0.005275	(0.5978,1.0800)	0.7839
	150	0.8240	0.1186	0.005274	(0.6321,1.0950)	0.8093
$\beta = 0.5$	30	0.5976	0.2113	0.005306	(0.2687,1.0880)	0.5706
	50	0.5635	0.1921	0.005415	(0.2774,1.0300)	0.5332
	100	0.5212	0.1750	0.007392	(0.2570,0.9230)	0.4954
	150	0.4990	0.1490	0.006292	(0.2595,0.8474)	0.4822
$\lambda = 0.5$	30	0.2354	0.1876	0.003029	(0.00653,0.6865)	0.1903
	50	0.3402	0.2279	0.004328	(0.01394,0.8417)	0.3101
	100	0.4408	0.2238	0.006032	(0.03971,0.8596)	0.4437
	150	0.4422	0.1905	0.005595	(0.05382,0.7761)	0.4587
$\alpha = 0.8$	30	0.9075	0.16500	0.004467	(0.6420, 1.2850)	0.8889
	50	0.6691	0.09102	0.002224	(0.5147,0.8685)	0.6598
	100	0.9733	0.14330	0.005941	(0.7385,1.2970)	0.9598
	150	0.8045	0.10710	0.004466	(0.6268,1.0430)	0.7936

Continued on next page

Table 6 – *Continued from previous page*

Parameter	n	Estimate	SD	MC error	95% Credible Interval	Median
$\beta = 1.5$	30	1.2040	0.3524	0.009401	(0.6320,2.0060)	1.1590
	50	1.8200	0.4637	0.011370	(1.0370,2.8300)	1.7790
	100	1.3500	0.3489	0.014150	(0.7682,2.1310)	1.3140
	150	1.4280	0.4045	0.016750	(0.7979,2.3730)	1.3710
$\lambda = 0.5$	30	0.6908	0.2128	0.002818	(0.1850,0.9612)	0.7307
	50	0.1735	0.1394	0.002083	(0.00468,0.5163)	0.1409
	100	0.7858	0.1563	0.003428	(0.4087,0.9894)	0.8163
	150	0.3297	0.1892	0.005103	(0.02239,0.7159)	0.3160

Table 7: Summary results for the posterior parameters in the case of the TINH model based on 128 bladder cancer patients (Gibbs sampling)

Parameter	n	Estimate	SD	MC error	95% Credible Interval	Median
α	128	0.6774	0.04840	9.059E-4	(0.5871, 0.77580)	0.67620
β		5.1520	0.75810	0.01403	(3.8190,6.78600)	5.10800
λ		0.05183	0.05317	8.52E-4	(2.815E-4,0.1935)	0.03513

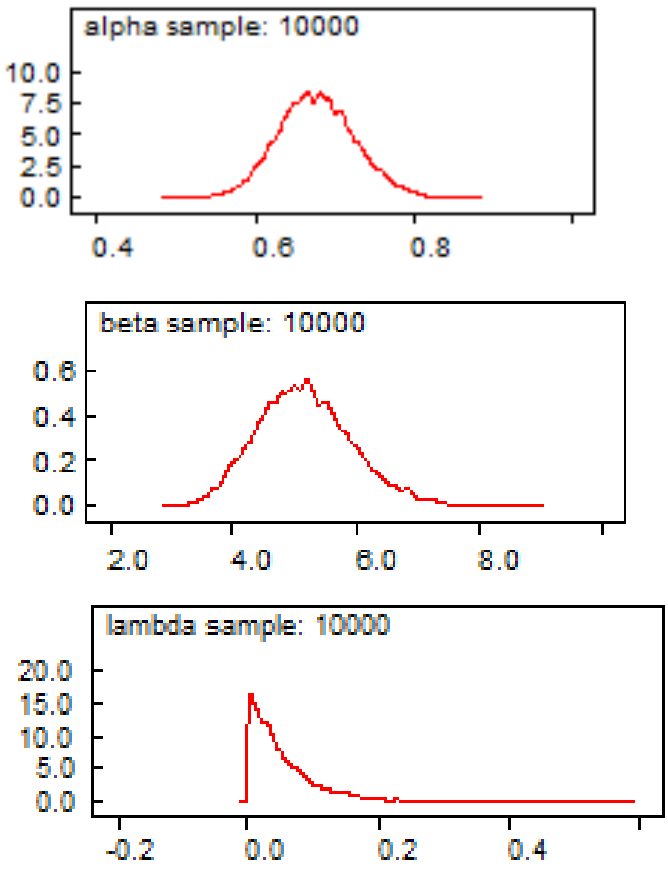


Figure 3: Posteriors of α, β and λ using Gibbs sampling for real data set

Table 8: Summary results for the posterior parameters in the case of the TINH model based on 128 bladder cancer patients (Metropolis-Hasting algorithm)

Parameter	n	Estimate	SD	95% Credible Interval	Median
α	128	0.6481	0.03593	(0.5834, 0.7227)	0.64583
β		0.0640	0.38463	(4.6554, 5.9891)	5.6639
λ		0.05183	0.04823	(0.01154, 0.1956)	0.05007

From Tables 7-8, one can note that all Bayesian point estimates are close however the Metropolis-Hastings provides narrower credible intervals.

6. Applications

Now, to illustrate the potentiality of the TINH distribution, we use a real data set and show that the new distribution is fit to this data set. The data set represents the remission times (in months) of a random sample of 128 bladder cancer patients. Bladder cancer is a disease in which abnormal cells multiply without control in the bladder. The most common type of bladder cancer recapitulates the normal histology of the urothelium and is known as transitional cell carcinoma. The data are as follows: 0.08, 0.20, 0.40, 0.50, 0.51, 0.81, 0.90, 1.05, 1.19, 1.26, 1.35, 1.40, 1.46, 1.76, 2.02, 2.02, 2.07, 2.09, 2.23, 2.26, 2.46, 2.54, 2.62, 2.64, 2.69, 2.69, 2.75, 2.83, 2.87, 3.02, 3.25, 3.31, 3.36, 3.36, 3.48, 3.52, 3.57, 3.64, 3.70, 3.82, 3.88, 4.18, 4.23, 4.26, 4.33, 4.34, 4.40, 4.50, 4.51, 4.87, 4.98, 5.06, 5.09, 5.17, 5.32, 5.32, 5.34, 5.41, 5.41, 5.49, 5.62, 5.71, 5.85, 6.25, 6.54, 6.76, 6.93, 6.94, 6.97, 7.09, 7.26, 7.28, 7.32, 7.39, 7.59, 7.62, 7.63, 7.66, 7.87, 7.93, 8.26, 8.37, 8.53, 8.65, 8.66, 9.02, 9.22, 9.47, 9.74, 10.06, 10.34, 10.66, 10.75, 11.25, 11.64, 11.79, 11.98, 12.02, 12.03, 12.07, 12.63, 13.11, 13.29, 13.80, 14.24, 14.76, 14.77, 14.83, 15.96, 16.62, 17.12, 17.14, 17.36, 18.10, 19.13, 20.28, 21.73, 22.69, 23.63, 25.74, 25.82, 26.31, 32.15, 34.26, 36.66, 43.01, 46.12, 79.05. These data were studied by Zea *et al.* (2012), among others. According to real data set the maximum likelihood estimates are obtained for the TINH distribution as follow:
$\hat{\alpha} = 3.15, \hat{\beta} = 1.8$ and $\hat{\lambda} = 0.85$.
Given the cumulative distribution function $F_0(x)$ of the hypothesized distribution (here TINH distribution) and the empirical distribution function Fdata (x) of the observed data, the popular Kolmogorov-Smirnov goodness of fit test (K-S) was carried out at 5% level of significance. The test statistic is given by:$D = \sup_x |F_0(x) - F_{data}(x)|$.For above data set, K-S statistic $D = 0.117$ with p-value $0.1 > 0.05$.

In many applications, there is qualitative information about the hazard rate shape, which can help with selecting a particular model. The empirical scaled TTT transform (Aarset, 1987) can be used to identify the shape of the hazard function. The scaled TTT transform is convex (concave) if the hazard rate is decreasing (increasing), and for bathtub (unimodal) hazard rates, the scaled TTT transform is first convex (concave) and then concave (convex). The TTT plot for complete data is the plot of $(i/n, G(i/n))$, where $G(\frac{i}{n}) = \frac{\sum_{j=1}^{i} T_{j:n} + (n-i)T_{i:n}}{\sum_{j=1}^{n} T_{j:n}}$ for $i = 1, 2, \ldots, n$, $\sum_{j=1}^{i} T_{j:n} + (n-i)T_{i:n}$ is the total time on test at the ith failure for $i = 1, 2, \ldots, n$ and $T_{(j:n)}, j = 1, 2, \ldots, n$, are the order statistics of the sample. Figure 4 presents TTT of complete data. As displayed in Figure 4: the TTT plot has first

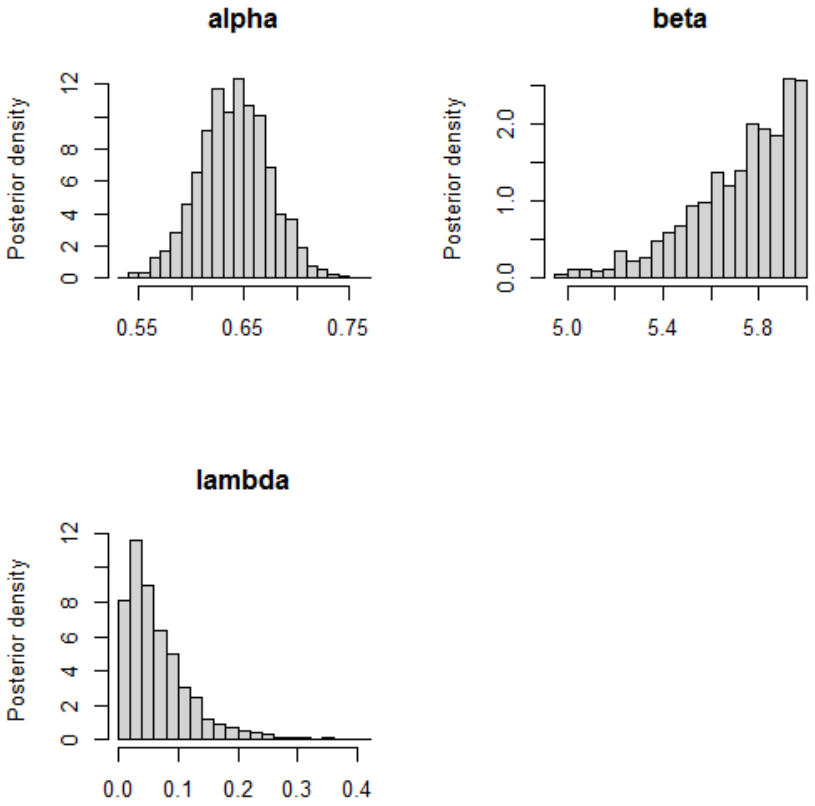


Figure 4: The approximated marginal Posterior density functions of α, β and λ using Metropolis-Hastings for real data set

a concave shape and then a convex shape. It depicts a unimodal shaped failure rate which agrees with the estimated parameters.

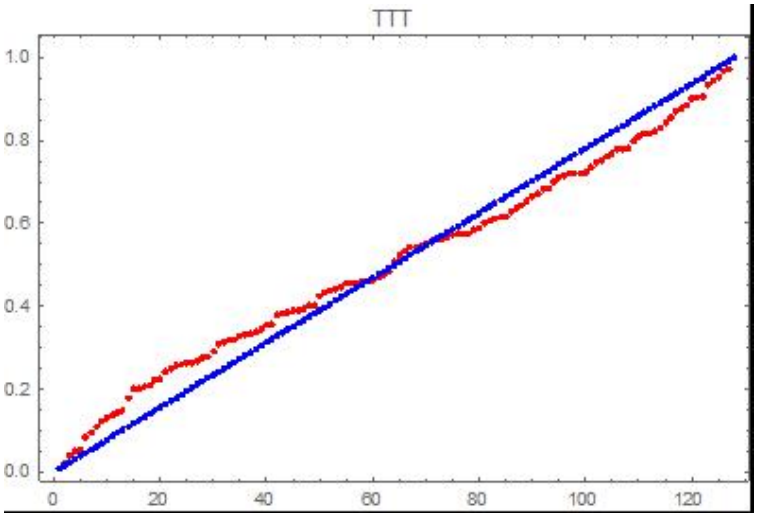


Figure 5: the TTT for real data set

Now we will compare the fits of the TINH, TMIW, TMIR, TMIE and MIW models by mean of another real data set to illustrate the potentiality of the TINH model. The cdfs associated with the competitive models are given by:

$$\begin{aligned}
F_{TMIW}(x) &= e^{-\gamma x^{-1}-\beta x^{-\alpha}}(1+\lambda-\lambda e^{-\gamma x^{-1}-\beta x^{-\alpha}}), \quad z>0, \alpha, \beta, \gamma>0 \text{ and } |\lambda| \leq 1, \\
F_{TMIR}(x) &= e^{-\gamma x^{-1}-\beta x^{-2}}(1+\lambda-\lambda e^{-\gamma x^{-1}-\beta x^{-2}}), \quad z>0, \beta, \gamma>0 \text{ and } |\lambda| \leq 1, \\
F_{TMIE}(x) &= e^{-(\gamma+\beta) x^{-1}}(1+\lambda-\lambda e^{-(\gamma+\beta) x^{-1}}), \quad z>0, \beta, \gamma>0 \text{ and } |\lambda| \leq 1, \\
F_{MIW}(x) &= e^{-\gamma x^{-1}-\beta x^{-\alpha}}, \quad z>0, \alpha, \beta>0, \text{ and } \gamma>0.
\end{aligned}$$

The following data represents a complete data with the exact times of failure. This data is considered a data set of the life of fatigue fracture of Kevlar 373/*epoxy* that are subject to constant pressure at the 90% stress level until all had failed. The data are: 0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 1.2766, 1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330, 2.2100, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960. This data is considered by Ogunde *et al.* (2017). For model comparison, we consider some well-known measures such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the consistent Akaike information criterion (CAIC) and the Hannan-Quinn information criterion (HQIC). These criterions are defined by:

$$\begin{aligned}
AIC &= -2l(\hat{\boldsymbol{\theta}}) + 2p; \\
BIC &= -2l(\hat{\boldsymbol{\theta}}) + plog(n); \\
CAIC &= -2l(\hat{\boldsymbol{\theta}}) + \frac{2pn}{n-p-1}; \\
HQIC &= -2l(\hat{\boldsymbol{\theta}}) + 2log(log(n)).
\end{aligned}$$

where $l(\hat{\boldsymbol{\theta}})$ denotes the log-likelihood function evaluated at the maximum likelihood estimates for parameters $\boldsymbol{\theta}$, p is the number of parameters and n is the sample size. The model with minimum AIC (or BIC, CAIC and HQIC) value is chosen as the best model to fit the data. Also we consider the statistics AD (A*) and CVM (W*) to compare the models, where lower values of these statistics indicate a good fit. Table 9 lists the mles of the model parameters, the values of the measures AIC, BIC, CAIC and HQIC and from this table one can conclude that the TINH model provides a better fit to the current data than the other models. Furthermore, the values of the statistics in Table 10 indicate the TINH model provides the best fit compared to the other models.

Acknowledgements

I am very grateful to the Chair Editor and the reviewer for valuable comments and suggestions which have improved considerably the first version of the manuscript.

Table 9: The MLEs and some measures for the fitted models

Model	Estimates	$l(\hat{\boldsymbol{\theta}})$	AIC	BIC	CAIC	HQIC
TINH	$\hat{\alpha} = 0.60, \hat{\beta} = 1.99, \hat{\lambda} = .08$	-145.6	297.13	304.12	297.46	299.92
TMIW	$\hat{\alpha} = 0.69, \hat{\beta} = 1.09, \hat{\lambda} = 0.97, \hat{\gamma} = 0.17$	-150.4	308.73	318.05	309.29	312.45
TMIR	$\hat{\beta} = 0.01, \hat{\lambda} = 0.57, \hat{\gamma} = 0.68$	-189.3	384.69	391.68	385.02	387.48
TMIE	$\hat{\beta} = 0.50, \hat{\lambda} = 0.02, \hat{\gamma} = 0.12$	-163.5	332.92	339.91	333.25	335.71
MIW	$\hat{\alpha} = 0.71, \hat{\beta} = 0.74, \hat{\gamma} = 0.14$	-155.4	316.7	323.7	317.1	319.5

Table 10: Statistics A* and W*

Distribution	A*	W*
TINH	4.120	0.693
TMIW	4.983	0.851
TMIR	8.529	1.542
TMIE	6.851	1.206
MIW	5.720	0.988

References

Aarset, M.V. (1987). How to identify a bathtub hazard rate? *IEEE Transactions on Reliability*, **R-36(1)**, 106-108.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *The Annals of Mathematical Statistics*, **23(2)**, 193-212.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, **49(268)**, 765-769.

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society*, **B45(3)**, 394-403.

Elbatal, I. (2013). Transmuted modified inverse Weibull distribution: A generalization of the modified inverse Weibull probability distribution. *International Journal of Mathematical Archive*, **4(8)**, 117-129.

Habibullah, S. N. and Ahmed, M. (2006). On a new class of univariate continuous distributions that are closed under inversion. *Pakistan Journal of Statistics and Operation Research*, **II(2)**, 151-159.

Khan, M. S. (2019). Transmuted Modified Inverse Weibull distribution: Properties and application. *Pakistan Journal of Statistics and Operation Research*, **XV(III)**, 667-677.

Folks, J. L. (1983). Inverse Distributions. *in Encyclopedia of Statistical Sciences (Vol. 4)*, eds. Kotz S. and Johnson N. L., New York: John Wiley, pp. 244-249.

Lehmann, E. L. and Shaffer, J. P. (1988). Inverted distributions. *The American Statistician*, **42(3)**, 191-194.

Lemonte, A. J., Cordeiro, G. M. and Moreno-Arenas, G. (2016). A new useful three-parameter extension of the exponential distribution. *Statistics*, **50(2)**, 312-337.

MacDonald, P. D. M. (1971). Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren. *Journal of the Royal Statistical Society*, **B33(2)**, 326-329.

Mahmoud, M. R. and Mandouh, R. M. (2013). On the Transmuted Fréchet Distribution. *Journal of Applied Sciences Research*, **9(10)**, 5553-5561.

Marshal, 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.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *The Statistician*, **37**, 25–32.

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

Nair, N. U., Sankaran, P. G. and Balakrishnan, N. (2013). *Quantile-Based Reliability Analysis*. Springer, NewYork.

Ogunde, A. A., Ibraheem, A. G. and Audu, A. T. (2017). Performance rating of transmuted Nadarajah and Haghighi exponential distribution: An analytical approach. *Journal of Statistics: Advances in Theory and Applications*, **17(2)**, 137-151.

Sangsanit, Y. and Bodhisuwan, W. (2016). The Topp-Leone generator of distributions: properties and inferences. *Songklanakarin Journal of Science and Technology*, **38(5)**, 537-548.

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

Shaw, W. T. and Buckley, I. R. C. (2009). The Alchemy of Probability Distributions: beyond Gram-Charlier Expansions, and a Skew-Kurtotic-Normal Distribution from a Rank Transmutation Map. https://arxiv.org/abs/0901.0434.

Sheikh, A. K. and Ahmad, M. and Ali, Z. (1987). Some remarks on the hazard functions of the inverted distributions. *Reliability Engineering*, **19(4)**, 255-261.

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

Tahir, M. H., Cordeiro, G. M., Ali, S., Dey, S. and Manzoor, A. (2018). The Inverted Nadarajah-Haghighi: Properties, estimation methods and applications. *Journal of Statistical Computation and Simulation*, **June 2018**, 1-24.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, **11(2)**, 93-112.

Yadav, A. S., Maiti, S. S., Saha, M. and Pandey, A. (2019). The inverse xgamma distribution: Statistical properties and different methods of estimation. *Annals of Data Science*, 1-19. https://doi.org/10.1007/s40745-019-00211-w

Yousof, H. M. and Korkmaz, M. Ç. (2017). Topp-Leone Nadarajah-Haghighi distribution. *Journal of Statisticians: Statistics and Actuarial Sciences*, **2**, 119-128.

Zea, L. M., Silva, R. B., Bourguignon, M., Santos, A. M. and Cordeiro, G. M. (2012). The beta exponentiated Pareto distribution with application to bladder cancer susceptibility. *International Journal of Statistics and Probability*, **1(2)**, 8-19.

Statistics and Application {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 75-87

Lexicon Based Sentiment Analysis (LBSA) to Improve the Accuracy of Acronyms, Emoticons, and Contextual Words

T. Nikil Prakash and A. Aloysius
Department of Computer Science, St. Joseph's College, Thiruchirappalli.

Received: 02 August 2020; Revised: 30 January 2021; Accepted: 08 February 2021

Abstract

Sentiment analysis (SA) is a process of text analysis and is also called an area of Natural Language Processing (NLP). NLP is concerned with finding the emotions or opinions within the text. NLP is used to classify the opinions with the user's expression in text reviews and to analyze whether the user text is positive or negative. The data are drawn from various resources such as web blogs, social media sites, e-commerce, online banking, etc. The research work focuses on the lexicon method and the data are collected from Twitter comments data using the Kaggle website. A new Senti_Con_Acro algorithm is proposed in this paper. The proposed work is focused on sentiment acronyms, emoticons, and contextual sentiment. The result of the proposed work is to improve the accuracy and efficiency of existing work.

Key words: Sentiment analysis; Lexicon based approaches; Contextual words; Emoticons; Acronyms.

1. Introduction

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether the data is positive, negative, or neutral. Sentiment analysis on textual information is also done to allow companies to track consumer feedback on brand and product sentiment and to understand customer needs. The advancement of social media and an increasingly wide range of communicating networks are sharing ideas and opinions among people everywhere in the world. Such comments and reviews are believed to essential assets for the users (Sankar, 2019). In recent years, a large number of people are accessing the Internet and social media. Smart devices have empowered users to share their expressions or opinions explicitly over social media and this information can reach a large audience in a fraction of seconds.

Sentiment analysis relate to the identification and other information on the feelings and attitudes expressed in natural language texts, opinions, and beneficial evaluations. It is well established and subject to considerable research to identify information relating to products, companies, and other commercial entities automatically. Sentiment analysis can be useful for handling customer feedback automatically, for concentrating advertising and for analyzing consumer trends and trends (Mullen, 2006).

Sentiment Analysis provides tools to analyse this information such as user feelings, emotions, product reviews, social media chats, comments and posts. SA has become popular in online communities to find business giants in mining consumer minds and enhanced business

Corresponding Author: T. Nikil Prakash
Email: nikilprakash1992@gmail.com

performance in recent years. It is considered to be an effective method for classifying user reviews into either positive or negative polarities. Data analytics researches and organizations to find the opinion of customers thinking and emotions. Sentiment analysis tools have identified their feeling and emotions to text. Mostly sentiment tools have been done at sentence and document level reviews. SA has three main approaches namely lexicon-based method, Machine Learning (ML) and hybrid approaches Liu (2015).

The lexicon-based method is tagged with polarity detection and the word or phrases whether determine positive or negative. ML tools are used to classify the training dataset. It automatically identifies the product reviews, online banking, etc (Devika, 2016). The hybrid approach of sentiment analysis exploits both Machine Learning methods and lexicon-based methods for polarity detection (Katrekar 2010).

Many challenges are similar to those of traditional feeling analysis, but they are not always identical. It is well known that people have ambiguous expressions of their sentiments and opinions. The Proposed work will identify the contextual words, acronyms, and emoticons in the text. The Literature Review describes existing work and a comparison of existing works and the proposed work describes the Senti_Con_Acro algorithm model and the experiments results. Finally, the conclusion and references are the final sections in this research work.

2. Literature Review

Yousif (2017) proposed and analyzed scientific citation sentiment analysis challenges and issues. They classified citation function and citation recommendation to have huge consideration of Sentiments. They identified data preprocessing methods used in scientific SA and presents citation context extraction, future extraction, and user's data sources.

Xie (2017) proposed maximum entropy probabilistic latent semantic analysis (ME-PLSA). This method extracts seed emotion words from Wikipedia and training corpus data. The maximum entropy model tests the process of emotion classification. *K*-fold model divides the training set and the test set. The emotional classification method classifies words such as relevant of words, part of speech in context, some similarity emotional words, and the relevance of degree adverbs and so on.

Felipe (2018) proposed a recommendation process in SA to textual data. They classified Facebook and Twitter datasets and classified all negative review posts. They evaluated the issues of data scarcity in e-commerce. The recommended system increased the assertiveness of the recommendation process. Support Vector Machine (SVM) algorithm classified the dataset and increased the performance of real data.

Chen (2018), proposed pre-trained character embedding with a Dual-Channel Convolutional Neural Network (char-DCCNN) in Sentiment analysis. They represent vector-matrix using the input text in a two-channel convolutional neural network. One channel is static and another one is fine-tuned. They collected the microblog sentiment dataset and identified the reviews such as film, sports, social and other filed of datasets. The Char-DCCNN method classified this dataset.

Dasa (2018) proposed real-time SA of Twitter streaming data for stock prediction. Streaming data find the source of data analysis collected in real-time. Streaming data normally deals with a continuous flow of data. The data carries information such as websites, social

media, mobile phone applications, server logs, *etc*. The active learning algorithm analyzed the data and predicted user behavior in a ceaseless manner. Long short-term memory helps to stream online data prediction and provided better results.

Thelwall (2017) proposed a lexicon sentiment analysis for identifying gender biases. The lexical sentiment algorithm detects strong positive from female categories and negative from male categories. A social media monitoring algorithm is used to find the male and female attitudes in social media. They found out that the people's opinions and the SentiStrength tool gave a better performance. They compared both male and female different levels of communication in sentiment analysis and accurate data.

Hassan (2016) proposed contextual semantic sentiment analysis for Twitter data. They detected the sentiment analysis of two levels such as entity level and tweet-level. They described three data sets such as StentiStrength during lexical words and the term of strength that is fixed data or unchanged data and another one is SentiCircles it dynamically changes the contextual words. They used SentiCircles data sets for lexicon-based sentiment identification at both levels of sentiment detection. These approaches enhanced the performance of other datasets.

Alexander (2010) proposed sentiment classification using an automatic corpus collection method in the training set. They used TreeTagger for POS tagging and compared the difference in polarity sets. They used synthetic structures, it is described to emotions and or opinion state facts. They collected corpus training data sets and classified the data as positive, negative, and neutral. The classification method such as naïve Bayes uses the N-gram and Part of Speech (POS) tagging method.

Raghavendra (2019) proposed the Rule-Based Modeling (RBM) method for sentiment lexicon analysis. They collected from the dataset of Cornell's review data and identified the data whether it is positive or negative reviews in text files. The RBM classified the data and gave better performance results in existing lexicon methods. The RBM is highly regarded as both a sentiment analysis and feature extraction tools.

Table 1: Comparative details of existing work

Authors	Approaches/ Classification	Dataset	Advantages	Disadvantages
Jiangfeng, (2019)	• Aspect-Level Sentiment Classification, • NLP, • Gaussian kernel, • Information retrieval	Social Media	• The sentence is hidden on the LSTM layer. • Improved the performance of aspect-level sentiment classification	• Influence among different aspects when one opinionated sentence owns more than one aspect terms
Reinald, (2018)	• Sentence classification. • NLP	Social Media	• Classification performance and achieves state-of-the-art performance	• Complex NLP tasks. • Inaccurate translations thus data producing

				noisy sentence vector
Al-Kabi (2018)	• Arabic text classification • Opinion mining • Naïve Bayes • SVM • K-NN	Arabic	• Effectiveness of two free online tools such as social mention and SentiStrength	• Unable to extract the stem of all slang words • Dataset is not stemmed • The problem of spelling mistakes and repetition of letters and characters
Parinda (2019)	• Machine learning techniques • Information Extraction • NLP • Naive Bayes • POS tagging	Social Media	• Flexible and customizable way of generating connections between data sources	Low quality data size
Douglas, (2013).	• Dictionary-based approaches • Conjunction-based approach	Blog posts	• Applied to older corpuses • Develop a class of minimally-supervised	• The size of the dictionary • The sensitivity of various dictionaries
Jonathan, (2017),	• Machine Learning • Semantic parsing	Social media	• Parse utterances in unseen domains by decoupling structure mapping • Improved generalization such as dependency trees, syntactic CCG parses	• Structure distribution in the target domain is very different from the source • Datasets where only denotations are provided • Average accuracy
Hassan, (2016),	• Supportvector machine(SVM) • Maximumentropy(maxent)	Twitter	• Higher performance in detecting neutral entities	• Influence on performance. • Different sentiment orientations
Mike, (2017).	Phrase-level sentiment analysis	Twitter	• Automatically Identify the contextual polarity for a large subset	Low accuracy

3. Proposed Work

The Proposed work section highlights on analysis of Twitter data, based on lexicon approaches. The Senti_Con_Acro algorithm framework model is proposed in this research work. Figure 1. shows the proposed Senti_Con_Acro algorithm framework model. The proposed Senti_Con_Acro algorithm model process contains several phases that follow as:

3.1. Senti_Con_Acro model framework

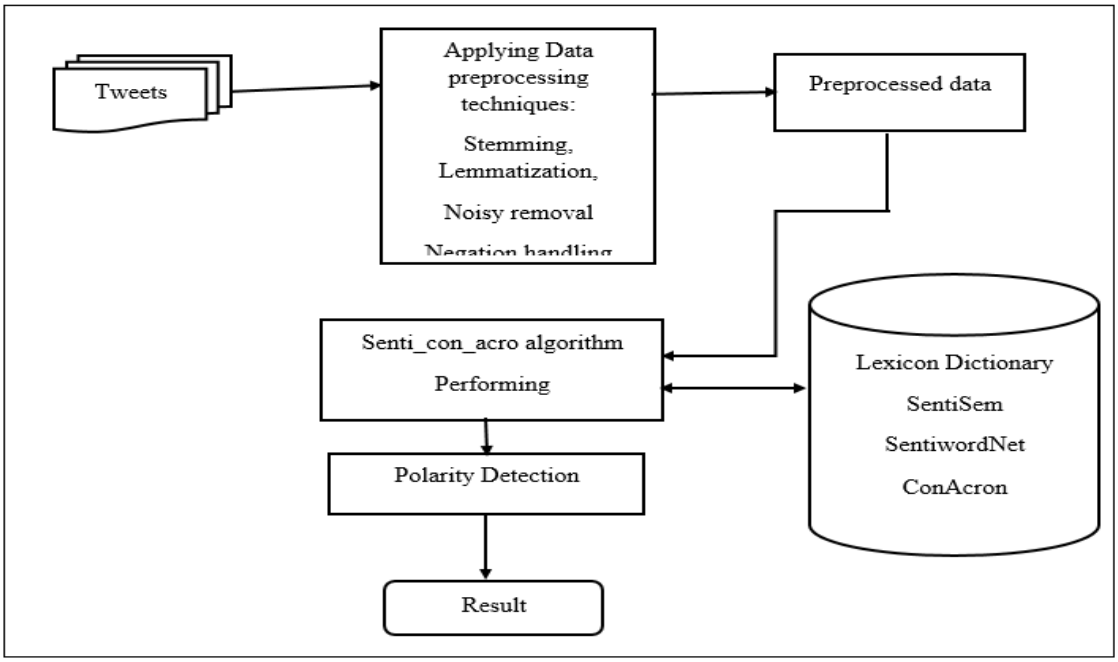


Figure 1: Framework for Senti_Con_Acro algorithm

Phase 1: Data Preprocessing

Data Preprocessing is a process of making unstructured data into structured data. It is often inconsistent, incomplete, and contains a lot of certain behaviors or product reviews and data many errors. Data preprocessing contains data preparation and data cleaning.

Data Preparation

Data preparation is a collection of a dataset from various resources like social media, e-commerce, web blogs, *etc*. In the proposed work, the data is collected from the Twitter dataset which is collected from the Kaggle website.

Data Cleaning

Data cleaning is the next step of data preprocessing. It refers to the process of cleaning to missing values, noisy data and inconsistent data. It also removes stop words and non-English words, *etc*. the Twitter dataset is cleaned and processed from the proposed work. The data cleaning contains three important steps that are as follows:

Tokenization

Tokenization is the process of splitting longer text into small pieces of text. It is also called text segmentation or lexical analysis.

Stemming

Stemming is a process of reducing words to the word stem from the dataset. Stemming algorithms have two types namely Porter stemmer and Lancaster Stemmer. Porter stemming removes morphological words and Lancaster stemmer removes aggressive words. It is used to determine domain vocabularies in domain analysis.

Lemmatization

Lemmatization is to access lexical knowledge bases and to get the correct sentence from the words. It is used for the WorldNet corpus and stops words in the corpus.

Phase 2: Feature Extraction

Feature extraction is the second stage of the proposed work. The text features are extracted from different ways that are as follows:

i. Unigram feature

In the proposed work Unigram features assume that the occurrence of each word is independent of its previous word. The proposed work counts all the input words using the unigram method. Hence each word becomes a gram (feature) here. For example:
"I", "have", "a", "lovely", "dog."

ii. Contextual words

The contextual words are called the different set of words or phrases. The proposed work identifies the user behavior or product performance in sentiment analysis. The contextual dictionary increasing content constantly which provides unmatched opportunities to support decision-making processes and advocacy efforts. Table 2 shows some examples of contextual words to convert acronyms of dictionary words.

Table 2: Contextual words to a dictionary word

Context Words	Dictionary Words
Abrupt	Sudden
Up-To-Date	Informed
Percepts	Perceived
Present	Existing
Common	Public
Constant	Continual

iii. Emoticons

The emoticons are emojis that identify the user behaviors and expressions. There is n number of emojis/ emoticons available in the emoticon dictionary. These emojis identify the positive and negative expressions in sentiment analysis algorithms. The proposed work identifies the emoticons and determines them as a positive or negative value. Figure 2 displays sentiment polarity using emoticons.

Figure 2: Emoticon Dictionary

iv. Acronyms

Abbreviations or acronyms are widely used in text materials to reduce space. The text in such areas consists of one to two sentences or a few sentences such as text messages, social media comments and blog posts. Customers may use or add new abbreviations or short word types, i.e. fast communication acronyms which rarely appear in regular or modern text, for these messages. Text as "TIA" for "Thank You in advance" is, for instance, common in these fields and for the machine. The textual significance of the texts could hardly be accurately understood. The high-rate text adds new abbreviations that can impact the reliability of the emotional analysis. To solve this problem, abbreviations must be extracted and identified before the sentiment method is performed.

Phase 3: Polarity Detection

Polarity detection is the third stage of the proposed work. Sentiment polarity determines three types of sentiment analysis that is positive, negative, and neutral. Sentiment polarity returns the overall opinion of a text or document in one single issue. The opinions classify into two opposing sentiment polarities are called positive or negative or introduce as neutral while the position of opinion locates between these two polarities. An opinionated text and categorizing it based on overall positive, neutral, and negative classes is called sentiment polarity classification.

Phase 4: Frequency Occurrence

Once the feature is extracted, they are used as input for supervised lexicon-based approaches for further classification. Generally, the frequency of occurrence of a keyword is a more suitable feature in overall sentiment analysis and not necessarily indicated by repeated use of keywords.

Brevity's law (also called Zipf's law) states that if words of a language are sorted in the order of decreasing frequency of usage, a word's frequency is inversely proportional to its rank, or sequence number in the list. The Brevity's Mandelbrot law equation is federated as the frequency of the sentiments which is measured as low rank and high-rank ratio and categorized through the deviancy of the power law. Brevity's Mandelbrot law check the ranking value if k >k0 is greater than the k0 value which gives the ranking is same in order the k value less than k0 the value is added as k0 + k

$(DT) \leftarrow f_k \propto (k0 + k)\text{-}b$

where, f ← frequency of a word,
k ← ranking of a word
DT ← Input data

Algorithm: Senti_Con_Acro

Phase 1: Pre-Processing

```
Input:Tweets (DT)
Output: Processed Tweets <dt'1, dt'2, .., dt'n> equivalent to <dt1,dt2,dt3…dtn>
Begin
        for each tweet do
                apply unigram and skip-gram // feature selection
                remove URLs, username.
                replace a repeated character by two or more indication of the same
        character
        if w is a stop word next
                remove w from DT
        else if w is a neighboring word next
                Detect the contextual word from the DT to assign W
        end if
        end for
return processed Tweets DT'
End
```

Phase 2: Feature Extraction

```
If DT word found in the dictionary then
        If DT word is a Neighboring word then
                Replace the equivalent word in a contextual word
        Else if DT word is an acronym then
                Replace the equivalent word for acronyms
        Else if DT word is emoticons then
                Replace the equivalent word for emoticons
        Else
         Identify the word is acronyms, and emoticons insert into the dictionary with
        equivalent meaning.
Else
        Identify the word is acronyms, and emoticons insert into the dictionary with
equivalent meaning.
```

Phase 3: Polarity Detection

For each DT word
DT word polarity ← DTp +DTne+ DTnu
DTp ← $\sum_{i=1}^{n} dt\ p(i)$
DTne ←$\sum_{i=1}^{n} dt\ ne(i)$
DTnu ← $\sum_{i=1}^{n} dt\ nu(i)$
$\sum_{i=1}^{n} dt$ Polarity ← sum ($\sum_{i=0}^{n} dt\ p(i)$, $\sum_{i=0}^{n} dt\ ne(i)$, $\sum_{i=0}^{n} dt\ nu(i)$
For each dt
If dt p(i) > dt nu(k) > DT Class (i) ← positive
Else if dt ne(i) < dt nu(i) > DT Class (i) ← Negative
Else dt nu(i) ← Neutral
End for
If DTp > DTne && DT nu then DT is positive impact
Else if DTne< DTnu and DTp then DT is negative impact
Else
DT is Neutral
End for

Phase 4: Frequency Occurrence

For each DT type
Frequency occurrences (DT) ←$f_k \infty (k + k)$-b
Max(f) then
Rank ← min
End for
End

3.2. Results

3.2.1. Confusion matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Table 3. Shows the classification of a test dataset produces four outcomes – true positive, false positive, true negative, and false negative.

Table 3: Confusion matrix classification of a test dataset

	Predicted Positives	Predicted Negatives
Actual Positive Instances	Number of True Positive	Number of False Negatives
Actual Negative Instances	Number of False Positive	Number of True Negatives

- **True Positives (TP)** - These are the correctly predicted positive values which mean that the value of the actual class is yes and the value of the predicted class is also yes.

- **True Negatives (TN)** - These are the correctly predicted negative values which mean that the value of the actual class is no and value of the predicted class is also no.

- **False Positives (FP)** – When the actual class is no and the predicted class is yes.

- **False Negatives (FN)** – When actual class is yes but predicted class is no.

3.2.2. Basic measures derived from the confusion matrix

Various measures can be derived from a confusion matrix. These are follow as:

➢ **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false-positive rate. In this proposed work done 0.7875 pretty good precision.

$$\text{Precision} = TP/TP+FP$$

➢ **Recall (Sensitivity)** - Recall is the ratio of correctly predicted positive observations to all observations in actual class - yes. In this proposed work is done recall of 0.7552 is good for this model as it's above 0.5.

$$\text{Recall} = TP/TP+FN$$

➢ **F-Measure** – F-measure is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In the proposed work done the F-measure is 0.7717.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

➢ **Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. For the proposed model, done 0.8093 which means the proposed model is approx. 81% accurate.

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN$$

3.2.3. Comparison results of the existing work

Table 4: Comparison results of the proposed work and existing work

S. No	Author	Accuracy (%)
1	Seydeh Akram Saadat Neshan (2020)	76.3
2	Ahmad A. Al-Oqaily (2020)	68
3	M. Edison (2017)	68.75
4	Vallikannu Ramanathan (2019)	76
5	Proposed Work	80.93

3.2.4. Result for the proposed work

Table 5: Result for the Proposed Work

	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
Proposed Result	78.75	75.52	77.17	80.53

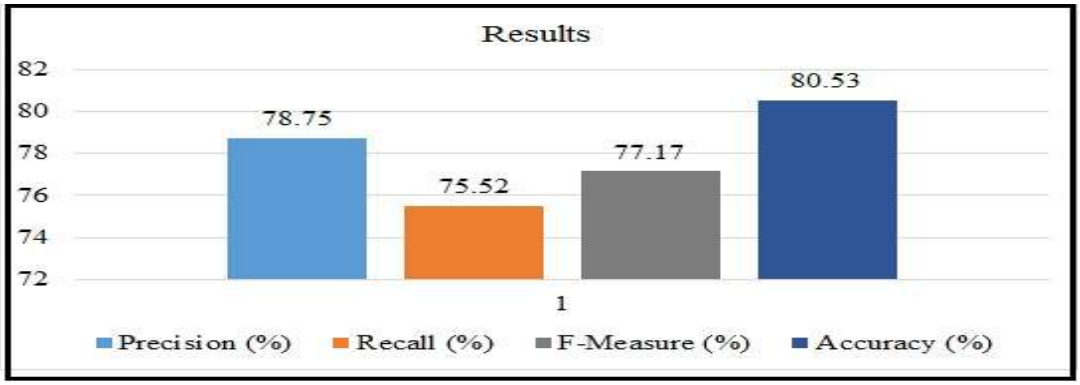


Figure 3: Evaluation measures from the confusion matrix

In this proposed work the statistical analysis of confusion matrix is applied to predict the result of emoticons, acronyms, contextual words with acronyms, and sentiments. Mainly sentiment analysis results are also predicted. The results are compared with the existing work and it brings better results than existing work.

4. Conclusion

Sentiment analysis is a platform of text analysis. It's identifying people's opinions, emotions, and sentiments, etc. This paper focused on lexicon-based sentiment analysis. A new Senti_Con_Acro Algorithm has been proposed to identify sentiment acronyms, emoticons, and contextual words. The emoticons, acronyms, contextual acronyms, and sentiments have been evaluated. This proposed work has given better results than the existing work. In the future, an image-based emotion detection method using different sentiment analysis approaches can be carried out to find the sentiments and to improve the accuracy to handle different evaluation metrics.

References

Akriti, S., Madhurima, H., Saru, D. and Madhulika, B. (2018). *An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach*. Indian Journal of Science and Technology, DOI: 10.17485/ijst/2018/v11i4/119594, **11(4)**, 1-6.

Alexander, P. and Patrick, P. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* Proceedings of the International Conference on Language Resources and Evaluation, 17-23.

Al-Kabi, M., Izzat, A., Rawan, K. and Heider, W. (2018). *Evaluating Social Context in Arabic Opinion.* The International Arab Journal of Information Technology, **15**, 1-7.

Al-Oqaily, A., Al-Hassan, M., Salah, K. and Elshqeirat, B. (2020). *Sentiment Analysis for Arabic Tweets Datasets: Lexicon-based and Machine Learning Approaches*. Journal of Theoretical and Applied Information Technology, **98**(**4**), 612-624.

Ashish, K. (2010). *An Introduction to Sentiment Analysis*. Global Logic Inc. 1-6.

Bing, L. (2015). *Sentiment Analysis.* Cambridge University Press, DOI: https://doi.org/10.1017/CBO9781139084789, 1-384.

Devika, M. D., Sunitha, C. and Amal, Ganesh (2016). *Sentiment Analysis: A Comparative Study on Different Approaches.* Fourth International Conference on Recent Trends in Computer Science and Engineering, DOI:10.1016/j.procs.2016.05.124. **87(2016)**. 44-49.

Douglas, R. and Christopher, Z. (2013). *Corpus-based Dictionaries for Sentiment Analysis of Specialized Vocabularies*. Prepared for presentation at the New Directions in Analyzing Text as Data Workshop, 1-17.

Edison, M. and Aloysius, A. (2017). Lexicon based Acronyms and Emoticons Classification of Sentiment Analysis (SA) on Big Data. *International Journal of Database Theory and Application*, **10**(**7**), 41-54. DOI: http://dx.doi.org/10.14257/ijdta.2017.10.7.04,

Felipe, G., Contratres, Solange, N., Alves S., Lucia, Vilela, Leite, F. Luiz, S. and Souza, D. (2018). *Sentiment Analysis of Social Network Data for Cold-Start Relief in Recommender Systems*. Springer International Publishing AG, DOI: https://doi.org/10.1007/978-3-319-77712-2_12, 122–132.

Hassan, S., Yulan H., Miriam F. and Harith A. (2016). *Contextual Semantics for Sentiment Analysis of Twitter.* Elsevier. DOI: http://dx.doi.org/10.1016/j.ipm.2015.01.0050306-4573, **52**, Issue 1, 5-19.

Jiangfeng, Z., Xiao, M. and Zhou, K. (2019). *Enhancing Attention-Based LSTM with Position Context for Aspect-Level Sentiment Classification.* IEEE, **7**, 20462 – 20472.

Jonathan, H., Michal, Shmueli, S. and Konopnicki, D. (2017*). Emotion Detection from Text via Ensemble Classification Using Word Embedding's.* Conference: the ACM SIGIR International Conference, DOI: 10.1145/3121050.3121093, 269-272.

Mike, T. (2017). *Gender Bias in Sentiment Analysis. Emerald Publishing, Information Review,* DOI: https://doi.org/10.1108/OIR-05-2017-0139, **42**, Issue. 1, 45-57.

Parinda, P. and Premkumar, Sivakumar. (2019). *Context Dependency Relation Extraction Using Modified Evolutionary Algorithm Based on Web Mining*. Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-13-1498-8_23, 259-267.

Raghavendra, K. D., Mukul, A., Surendra, K., Keshari and Ashwini, K. (2019). *Sentiment Analysis and Feature Extraction Using Rule-Based Model (RBM).* Springer Nature Singapore, DOI: https://doi.org/10.1007/978-981-13-2354-6_7, 57-64.

Reinald, K., Amplayoy, Kyungjae, Leey, Jinyeong, Yeoz and Seung-won, Hwangy (2018). *Translations as Additional Contexts for Sentence Classification.* Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 1-7.

Sankar, H., Subramaniyaswamy, V., Vijayakumar, V., Sangaiah Arun Kumar, R., Logesh and Umamakeswari, A. (2019). *Intelligent Sentiment Analysis Approach using Edge Computing-based Deep Learning Technique.* Wiley Online Library, DOI: https://doi.org/10.1002/spe.2687, 1–13.

Seydeh, A., Saadat, N. and Reza, A. (2020). *A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis.* 6th International Conference on Web Research (ICWR), IEEE, DOI: 10.1109/ICWR49608.2020.9122298, 8-14.

Shuai, W., Mianwei, Z., Sahisnu, M., Bing, L. and Chang, Y. (2018). *Disentangling Aspect and OpinionWords in Target-based Sentiment Analysis using Lifelong Learning.* Research Gate, DOI: arXiv: 1802.05818, 1-7.

Sisi, C., Youdong, D., Zhifeng, X., Shenkai, L. and Huiming, D. (2018). *Chinese Weibo Sentiment Analysis Based on Character Embedding with Dual-Channel Convolutional Neural Network*. IEEE International Conference on Cloud Computing and Big Data Analysis, DOI: 978-1-5386-4301-3/18/$31.00, 107-111.

Suman, S. and Rajitha, N. (2018) *Unsupervised Deep Representations for Learning Audience Facial Behaviors*. IEEE, DOI: arXiv:1805.04136, 1-6.

Sushree, D., Ranjan, K., Behera, Mukesh, K., Santanu, K. and Ratha. (2018). *Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction*. Science Direct, DOI: 10.1016/j.procs.2018.05.111, 956-964.

Tony. M. and Robert, M. (2006). *A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse.* Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, 1-4.

Vallikannu, R., Meyyappan, T. (2019). *Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism*. 4th MEC International Conference on Big Data and Smart City (ICBDSC), IEEE, 1-5.

Xie, X., Songlin, G., Hu, F., Xie, M. and Jiang, N. (2017). *An Improved Algorithm for Sentiment Analysis based on Maximum Entropy.* Soft Computing, Springer. DOI: 10.1007/s00500-017-2904-0, 1-13.

Yousif, A., Niu, Z., Taurus, J. K. and Ahmad, A. (2017). *A Survey on Sentiment Analysis of Scientific Citations.* Artificial Intelligence Review. Springer, DOI: https://doi.org/10.1007/s10462-017-9597-8, 1-34.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 89–111

Correlated Inverse Gaussian Frailty Models Based on Reversed Hazard Rate

David D. Hanagal
Department of Statistics, Savitribai Phule Pune University, Pune, India
Email: david.hanagal@gmail.com

Received: 13 December 2020; Revised: 25 January 2021; Accepted: 09 February 2021

Abstract

Frailty models are used in the survival analysis to account for the unobserved heterogeneity in individual risks to disease and death. To analyze the bivariate data on related survival times (*e.g.* matched pairs experiments, twin or family data), the shared frailty models were suggested. Shared frailty models are used despite their limitations. To overcome their disadvantages correlated frailty models may be used. In this paper, we introduce the correlated inverse Gaussian frailty models based on reversed hazard rate with three different baseline distributions namely, the generalized log-logistic type I, the generalized log-logistic type II and the modified inverse Weibull. We introduce the Bayesian estimation procedure using Markov Chain Monte Carlo (MCMC) technique to estimate the parameters involved in these models. We present a simulation study to compare the true values of the parameters with the estimated values. We also apply the proposed models to the Australian twin data set and a better model is suggested. ...

Key words: Australian twin data; Bayesian estimation; Correlated inverse Gaussian frailty; Generalized log-logistic distribution; MCMC; Modified inverse Weibull distribution; Reversed hazard rate.

AMS Subject Classifications: 62F15; 62N01; 62P10

1. Introduction

Frailty models are extensively used in the survival analysis to account for the unobserved heterogeneity in individual risks to disease and death. The frailty model is a random effect model for time to event data which is an extension of the Cox's proportional hazards model. To analyze the bivariate data on related survival times (*e.g.* matched pairs experiments, twin or family data), the shared frailty models were suggested. Bivariate survival data arises whenever each study subjects experience two events. Particular examples include failure times of paired human organs, (*e.g.* kidneys, eyes, lungs, breasts, *etc.*) and the first and the second occurrences of a given disease. In the medical literature, several authors considered paired organs of an individual as a two-component system, which work under interdependency circumstances. In industrial applications, these data may come from systems whose survival depend on the survival of two similar components.

Research on the bivariate survival models has grown rapidly several years in the past. Clayton's (1978) random effect model of the bivariate survival was a key innovation. He introduced the notion of the shared relative risk. This model was further developed by Oakes (1982) to analyze the association between two non-negative random variables. Clayton and Cuzick (1985) added observed covariates to the bivariate survival model with the shared relative risk. Crowder (1985) and Hougaard (1986) proposed the random effect models of the bivariate Weibull distributions. A shared frailty model with a positive stable distribution of frailty was suggested by Hougaard (1987). He also discussed several other bivariate distributions with biomedical and reliability applications. Oakes (1989) developed a shared frailty model related to the "archimedean distributions" studied by Genest and MacKay (1986). He also proposed a local time dependent association measure between bivariate life spans and discussed its use for a large class of bivariate survival functions. Vaupel (1991), Vaupel *et al.* (1991), Nielsen *et al.* (1992) studied genetic and environmental influences on longevity using bivariate survival models.

Hanagal (2006) discussed the gamma frailty regression model in the bivariate survival data and Hanagal (2007) also presented the gamma frailty regression models in the mixture distributions. Hanagal and Dabade (2013), Hanagal and Bhambure (2015, 2016) and Hanagal and Pandey (2014a, 2014b, 2015a, 2015b, 2016, 2017) and Hanagal *et al.* (2017a, 2017b) analyzed kidney infection data and Australian twin data using shared gamma and inverse Gaussian frailty models with different baseline distributions for the multiplicative model. Hanagal and Sharma (2013, 2015a, 2015b, 2015c) analyzed acute leukemia data, kidney infection data and diabetic retinopathy data using shared gamma and inverse Gaussian frailty models for the multiplicative model. Hanagal and Bhambure (2014) developed shared inverse Gaussian frailty model based on the reversed hazard rate for Australian twin data. Hanagal *et al.*(2017b) discussed correlated gamma frailty models for bivariate survival data to analyze kidney infection data and Hanagal and Pandey (2017) proposed correlated gamma frailty models for bivariate survival data based on reversed hazard rate for Australian twin data. Hanagal (2017) gave extensive literature review on different shared frailty models.

Shared frailty explains correlation between subjects within clusters. However, it does have some limitations. Firstly, it forces the unobserved factors to be the same within the cluster, which may not always reflect reality. For example, at times it may be inappropriate to assume that all partners in a cluster share all their unobserved risk factors. Secondly, the dependence between survival times within the cluster is based on marginal distributions of survival times. However, when covariates are present in a proportional hazards model with gamma distributed frailty the dependence parameter and the population heterogeneity are confounded (Clayton and Cuzick, 1985). This implies that the joint distribution can be identified from the marginal distributions (Hougaard, 1986). Thirdly, in most cases, a one-dimensional frailty can only induce positive association within the cluster. However, there are some situations in which the survival times for subjects within the same cluster are negatively associated. For example, in the Stanford Heart Transplantation Study, generally the longer an individual must wait for an available heart, the shorter he or she is likely to survive after the transplantation. Therefore, the waiting time and the survival time afterwards may be negatively associated.

To avoid these limitations, correlated frailty models are being developed for the analysis of multivariate failure time data, in which associated random variables are used to

characterize the frailty effect for each cluster. Correlated frailty models provide not only variance parameters of the frailties as in shared frailty models, but they also contain additional parameter for modeling the correlation between frailties in each group. Frequently one is interested in construction of a bivariate extension of some univariate family distributions (*e.g.*, gamma). For example, for the purpose of genetic analysis of frailty one might be interested in estimation of correlation of frailty. It turns out that it is possible to carry out such extension for the class of infinitely-divisible distributions (Iachine 1995a, 1995b). In this case an additional parameter representing the correlation coefficient of the bivariate frailty distribution is introduced.

2. Reversed Hazard Rate and Correlated Frailty

In many practical situations reversed hazard rate (RHR) is more appropriate to analyze the survival data. Reversed hazard rate was proposed as a dual to the hazard rate by Barlow *et al.* (1963). Shaked and Shantikumar (1994) and Block *et al.* (1998) provided a general definition of reversed hazard rate (RHR) as,

$$m(t) \;=\; \lim_{\triangle t \to 0} P(t - \triangle t < T \leq t | T \leq t) / \triangle t, \quad t > 0. \tag{1}$$

The reversed hazard rate specifies the instantaneous rate of death or failure at time t, given that it failed before time t. Thus in a small interval, $m(t) \triangle t$ is the approximate probability of failure in the interval, given failure until the end of the interval $(t - \triangle t, t]$. In lifetime data analysis, the concepts of reversed hazard rate has potential application when the time elapsed since failure is a quantity of interest in order to predict the time of failure. The reversed hazard rate is more useful in estimating reliability function when the data are left censored or right truncated. Reversed hazard rate plays a vital role in the analysis of parallel systems, in reliability and survival analysis. For example, in certain systems or situations, sometimes the failure is prevented through numerous safety measures.

The correlated frailty model is the important concept in the area of multivariate frailty models. It is a natural extension of the shared frailty approach on the one hand, and of the univariate frailty model on the other. In the correlated frailty model, the frailties of individuals in a cluster are correlated but not necessarily shared. The conditional distribution function in the bivariate case (without observed covariates) is

$$F(t_1, t_2 | Z_1, Z_2) \;=\; S_1(t_1|Z_1) S_2(t_2|Z_2) = e^{-Z_1 M_{01}(t_1)} e^{-Z_2 M_{02}(t2)}, \tag{2}$$

where Z_1 and Z_2 are two correlated frailties and $M_{0i}(t_i) = \int_{t_i}^{\infty} m_{0i}(u) du, (i = 1, 2)$ is cumulative reversed hazard rate. The distribution of the random vector (Z_1, Z_2) needs to be specified and determines the association structure of the event times in the model.

The reversed hazard of the i-th $(i = 1, 2)$ individual of the j-th $(i = j, ..., n)$ pair has the form

$$m(t \mid X_{ij}, Z_{ij}) = Z_{ij} m_{0i}(t) e^{\beta' X_{ij}}, \tag{3}$$

where t denotes age or time, X_{ij} is a vector of observed covariates, β is a vector of regression parameters describing the effect of the covariates X_{ij}, $m_{0i}(.)$ are baseline reversed hazard functions, and Z_{ij} are frailties. Bivariate correlated frailty models are characterized by the joint distribution of a two-dimensional vector of frailties (Z_{1j}, Z_{2j}). If the two frailties are independent, the resulting lifetimes are independent, and no clustering is present in the model. If the two frailties are equal, the shared frailty model is obtained as a special case of the correlated frailty model with correlation one between the frailties.

In order to derive a marginal likelihood function, the assumption of conditional independence of lifespans, given the frailty, is used. Let δ_{ij} be a censoring indicator for individual $i(i = 1, 2)$ in pair $j(j = 1, ..., n)$. Indicator δ_{ij} is 1 if the individual has experienced the event of interest, and 0 otherwise. According to (2), the conditional distribution function of the ith individual in the jth pair is

$$F(t|X_{ij}, Z_{ij}) = e^{-Z_{ij}M_{0i}(t)e^{\beta' X_{ij}}}, \tag{4}$$

with $M_{0i}(t) = \int_t^\infty m_{0i}(u)du$ denoting the cumulative baseline hazard function. Here and in the following, F is used as a generic symbol for a distribution function. The contribution of individual $i(i = 1, 2)$ in pair $j(j = 1, ..., n)$ to the conditional likelihood is given by

$$\left[Z_{ij}m_{0i}(t)e^{\beta' X_{ij}}\right]^{\delta_{ij}} e^{-Z_{ij}M_{0i}(t_{ij})e^{\beta' X_{ij}}}, \tag{5}$$

where t_{ij} stands for observation time of individual i from pair j. Assuming the conditional independence of life spans, given the frailty, and integrating out the frailty, we obtain the marginal likelihood function

$$\prod_{j=1}^{n} \int_{R\times} \int_{R} \left[u_{1j}m_{01}(t_{1j})e^{\beta' X_{1j}}\right]^{\delta_{1j}} e^{-z_{1j}M_{01}(t_{1j})e^{\beta' X_{1j}}}$$
$$\left[u_{2j}m_{02}(t_{2j})e^{\beta' X_{2j}}\right]^{\delta_{2j}} e^{-z_{2j}M_{02}(t_{2j})e^{\beta' X_{2j}}} f(z_{1j}, z_{2j})dz_{1j}dz_{2j} \tag{6}$$

where $f(.,.)$ is the probability density function of the corresponding frailty distribution. All these formulas can be easily extended to the multivariate case, but need a specification of the correlation structure between individuals in a cluster in terms of the multivariate density function, which complicates analysis. For more details see Hanagal(2011) and Hanagal (2019).

3. Correlated Inverse Gaussian Frailty Model

Alternative to the gamma distribution, Hougaard (1984) introduced the inverse Gaussian as a frailty distribution. It provides much flexibility in modeling, when early occurrences of failures are dominant in a life time distribution and its failure rate is expected to be non-monotonic. In such situations, the inverse Gaussian distribution might provide a suitable choice for the lifetime model. Also inverse Gaussian is almost an increasing failure rate distribution when it is slightly skewed and hence is also applicable to describe lifetime distribution which is not dominated by early failures. Secondly, for the inverse Gaussian distribution, the surviving population becomes more homogeneous with respect to time, where as for gamma distribution the relative heterogeneity is constant.

Consider a continuous random variable Z follows inverse Gaussian distribution with parameters μ and σ^2 then density function of Z is,

$$f_Z(z) = \begin{cases} \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}} z^{-\frac{3}{2}} e^{-\frac{(z-\mu)^2}{2z\sigma^2\mu^2}} & ;\ z>0, \mu>0, \sigma^2>0 \\ 0 & ;\ otherwise, \end{cases} \tag{7}$$

and the Laplace transform is,

$$L_Z(s) = exp\left[\frac{1}{\mu\sigma^2} - \left(\frac{1}{\sigma^4\mu^2} + \frac{2s}{\sigma^2}\right)^{\frac{1}{2}}\right]. \tag{8}$$

The mean and variance of frailty variable are $E(Z) = \mu$ and $V(Z) = \mu^3\sigma^2$. For identifiability, we assume Z has expected value equal to one i.e. $\mu = 1$. Under this restriction, the density function and the Laplace transformation of the inverse Gaussian distribution reduces to,

$$f_Z(z) = \begin{cases} \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}} z^{-\frac{3}{2}} e^{-\frac{(z-1)^2}{2z\sigma^2}} & ;\ z>0, \sigma^2>0 \\ 0 & ;\ otherwise, \end{cases} \tag{9}$$

and the Laplace transform is,

$$L_Z(s) = exp\left[\frac{1-(1+2\sigma^2 s)^{\frac{1}{2}}}{\sigma^2}\right], \tag{10}$$

with variance of Z as σ^2. The frailty variable Z is degenerate at $Z = 1$ when σ^2 tends to zero.

Let Z be an infinitely divisible frailty variable with Laplace transformation $L_Z(s)$ and $\rho \in [0, 1]$, then there exist random variables Z_1, Z_2 each with univariate Laplace transform $L_Z(s)$ such that the Laplace transform of Z_1, Z_2 is given by:

$$L(s_1, s_2) = L_Z^{\rho}(s_1 + s_2) L_Z^{1-\rho}(s_1) L_Z^{1-\rho}(s_2) \tag{11}$$

If Z has a variance the $Corr(Z_1, Z_2) = \rho$.
The respective bivariate survival model is identifiable under mild regularity conditions on Z provided that $\rho > 0$. The case $\rho = 1$ is known as the shared frailty model.

The above equation (11) can be extended to multivariate case ($\rho > 0$) as below.

$$L(s_1, s_2, ..., s_k) \quad = \quad L_Z^{\rho}(s_1, s_2, ..., s_k) L_Z^{1-\rho}(s_1)....L_Z^{1-\rho}(s_k).$$

The case $\rho = 1$ leads to shared frailty. If $\rho = 0$, $Z_1, ...Z_k$ are mutually independent.

Let Z_i be the inverse Gaussian distributed with mean 1, variance σ^2, and Laplace transform

$$L(s_i, \sigma^2) = \exp[\frac{1-(1+2\sigma^2 s_i)^{\frac{1}{2}}}{\sigma^2}] \tag{12}$$

The bivariate Laplace transform for the correlated inverse Gaussian frailty model is given by

$$L(s_1, s_2, \sigma^2, \rho) = \exp\left[\rho\frac{1-(1+2\sigma^2(s_1+s_2))^{\frac{1}{2}}}{\sigma^2}\right]\exp\left[(1-\rho)\frac{1-(1+2\sigma^2 s_1)^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1+2\sigma^2 s_2)^{\frac{1}{2}}}{\sigma^2}\right] \tag{13}$$

where $Corr(Z_1, Z_2) = \rho$.

The correlated inverse Gaussian frailty model in the presence of covariates is characterized by the bivariate distribution function of the form:

$$F(t_{1j}, t_{2j}) = \exp\left[\rho\frac{1-(1+2\sigma^2\eta_j(M_{01}(t_{1j})+M_{02}(t_{2j})))^{\frac{1}{2}}}{\sigma^2}\right]\exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_j M_{01}(t_{1j}))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_j M_{02}(t_{2j}))^{\frac{1}{2}}}{\sigma^2}\right] \tag{14}$$

where $M_{01}(t_{1j})$ and $M_{02}(t_{2j})$ are the cumulative baseline hazard functions of the life time random variables T_{1j} and T_{2j} respectively.

The bivariate distribution function in the presence of covariates, when the frailty variable is degenerate is given by

$$F(t_{1j}, t_{2j}) = e^{-[(M_{01}(t_{1j})+M_{02}(t_{2j}))\eta_j]} \tag{15}$$

According to different assumptions on the baseline distributions we get different correlated inverse Gaussian frailty models.

4. Baseline Distributions

We present the modified inverse Weibull distribution, generalized log-logistic type I and generalized log-logistic type II as baseline distribution with the interesting properties.

4.1. Modified Inverse Weibull Distribution

The modified inverse Weibull distribution is more convenient for computational point of view for left censored data. The cumulative distribution function, the reversed hazard rate and the cumulative reversed hazard rate of the modified inverse Weibull are respectively as follows.

$$F(t) = exp\left(-\alpha t^{-\lambda}e^{-\gamma t}\right) \quad ; t > 0, \alpha > 0, \lambda > 0, \gamma > 0, \tag{16}$$

$$m(t) = \alpha e^{-\gamma t}t^{-1-\lambda}(\lambda + \gamma t). \tag{17}$$

$$M(t) = \alpha t^{-\lambda}e^{-\gamma t}, \tag{18}$$

When $\gamma = 0$, this distribution reduces to the inverse Weibull distribution. The reversed hazard rate of the modified inverse Weibull distribution is decreasing function of $t > 0$. For more details see Devendra *et al.* (2011).

4.2. Generalized Log-logistic Distribution

The log-logistic distribution is very useful in a wide variety of applications, especially in the analysis of survival data (O' Quigley and Struthers 1982; Bennett 1983; Cox and Snell 1989). The log-logistic distribution is very similar in shape to the log-normal distribution, however it has the advantage of having simple algebraic expressions for its survivor and hazard functions and a closed form for its distribution function. It is therefore more convenient than the log-normal distribution in handling censored data. However, due to the symmetry of the log-logistic distribution, it may be inappropriate for modeling censored survival data, especially for the cases where the hazard rate is skewed or heavily tailed. In order to overcome this, we use a generalization of the log-logistic distribution and refer to this as the generalized log-logistic distribution given in Mohammed *et al.*(1990). The generalized log-logistic distribution reflects the skewness and the structure of the heavy tail and generally shows some improvement over the log-logistic distribution.

Mohammed *et al.*(1990) show that the distribution function of generalized logistic is given by

$$F(x) = \frac{1}{\beta(m,n)} \int_0^{F_0(x)} u^{m-1}(1-u)^{n-1}du$$

where $\beta(m,n)$ is the complete beta function and

$$F_0(x) = (1+e^{-x})^{-1}, -\infty < x < \infty$$

is the logistic distribution function. We call $F(x)$ the generalized logistic distribution with parameters (m,n), and use the notation $X \sim GLD(m,n)$.
The logarithmic transformation $X = \gamma ln(\lambda T)$ applied to $GLD(m,1)$ to obtain the generalized log-logistic distribution $GLLD(m,1)$. The distribution function of T is

$$F(t) = (1+(\lambda t)^{-\gamma})^{-m}, t, m, \lambda > 0, \gamma \geq 1. \tag{19}$$

Similarly logarithmic transformation $X = \gamma ln(\lambda T)$ applied to $GLD(1,n)$ to obtain the generalized log-logistic distribution $GLLD(1,n)$. The distribution function of T is

$$F(t) = 1-(1+(\lambda t)^{\gamma})^{-n}, t, n, \lambda > 0, \gamma \geq 1. \tag{20}$$

A random variable T with c.d.f. as given by (19) and (20) are generalized log-logistic distribution with parameters $(m,1)$ and $(1,n)$ respectively. We call (19) as generalized log-logistic type I and (20) as generalized log-logistic type II.

Now rearranging the parameters, the cumulative distribution function of the generalized log-logistic distribution type I is

$$F(t) = \left(\frac{(\lambda t)^{\gamma}}{1+(\lambda t)^{\gamma}}\right)^{\alpha}. \tag{21}$$

The corresponding reversed hazard rate and cumulative reversed hazard rate are respectively as follows.

$$m(t) = \frac{\alpha\gamma}{t(1+(\lambda t)^{\gamma})}. \tag{22}$$

$$M(t) = \alpha ln\left(\frac{1+(\lambda t)^{\gamma}}{(\lambda t)^{\gamma}}\right) \tag{23}$$

Now rearranging the parameters, the cumulative distribution function of the generalized log-logistic distribution type II is

$$F(t) = 1 - (1+(\lambda t)^{\gamma})^{-\alpha}. \tag{24}$$

The corresponding reversed hazard rate and cumulative reversed hazard rate are respectively as follows.

$$m(t) = \frac{\alpha\gamma\lambda(\lambda t)^{-1+\gamma}(1+(\lambda t)^{\gamma})^{-1-\alpha}}{1-(1+(\lambda t)^{\gamma})^{-\alpha}}. \tag{25}$$

$$M(t) = -ln(1-(1+(\lambda t)^{\gamma})^{-\alpha}) \tag{26}$$

When $\alpha = 1$, this distribution reduces to log-logistic distribution. The reversed hazard rate of the generalized log-logistic distribution is decreasing function of $t > 0$.

5. Proposed Models

Substituting cumulative reversed hazard function for the modified inverse Weibull baseline distribution, generalized log-logistic type I and generalized log-logistic type II, we get following six models.

$$\begin{aligned} F(t_{1j},t_{2j}) &= \exp\left[\rho\frac{1-(1+2\sigma^2\eta_{0j}(\eta_{1j}\alpha_1 t_{1j}^{-\lambda_1}e^{-\gamma_1 t_{1j}}+\eta_{2j}\alpha_2 t_{2j}^{-\lambda_2}e^{-\gamma_2 t_{2j}}))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{1j}\alpha_1 t_{1j}^{-\lambda_1}e^{-\gamma_1 t_{1j}})^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{2j}\alpha_2 t_{2j}^{-\lambda_2}e^{-\gamma_2 t_{2j}})^{\frac{1}{2}}}{\sigma^2}\right] \end{aligned} \tag{27}$$

$$F(t_{1j},t_{2j}) = exp\left(-\eta_{0j}\left\{\eta_{1j}\alpha_1 t_{1j}^{-\lambda_1}e^{-\gamma_1 t_1}+\eta_{2j}\alpha_2 t_{2j}^{-\lambda_2}e^{-\gamma_2 t_2}\right\}\right) \tag{28}$$

$$\begin{aligned} F(t_{1j},t_{2j}) &= \exp\left[\rho\frac{1-(1+2\sigma^2\eta_{0j}(\eta_{1j}\alpha_1\ln(1+1/(\lambda_1 t_{1j})^{\gamma_1})+\eta_{2j}\alpha_2\ln(1+1/(\lambda_2 t_{2j})^{\gamma_2})))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{1j}\alpha_1\ln(1+1/(\lambda_1 t_{1j})^{\gamma_1}))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{2j}\alpha_2\ln(1+1/(\lambda_2 t_{2j})^{\gamma_2}))^{\frac{1}{2}}}{\sigma^2}\right] \end{aligned} \tag{29}$$

$$F(t_{1j},t_{2j}) = exp\left(-\eta_{0j}\left\{\eta_{1j}\alpha_1\left(ln(\frac{1+(\lambda_1 t_{1j})^{\gamma_1}}{(\lambda_1 t_{1j})^{\gamma_1}})\right)+\eta_{2j}\alpha_2\left(ln(\frac{1+(\lambda_2 t_{2j})^{\gamma_2}}{(\lambda_2 t_{2j})^{\gamma_2}})\right)\right\}\right) \tag{30}$$

$$\begin{aligned} F(t_{1j},t_{2j}) &= \exp\left[\rho\frac{1-(1+2\sigma^2\eta_{0j}(\eta_{1j}\ln(1-(1+1(\lambda_1 t_{1j})^{\gamma_1})^{-\alpha_1})+\eta_{2j}\ln(1-(1+1(\lambda_2 t_{2j})^{\gamma_2})^{-\alpha_2})))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{1j}\ln(1-(1+1(\lambda_1 t_{1j})^{\gamma_1})^{-\alpha_1}))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_{0j}\eta_{2j}\ln(1-(1+1(\lambda_2 t_{2j})^{\gamma_2})^{-\alpha_2}))^{\frac{1}{2}}}{\sigma^2}\right] \end{aligned} \tag{31}$$

$$F(t_{1j},t_{2j}) = exp\left(\eta_{0j}\left\{\eta_{1j}ln(1-(1+(\lambda_1 t_{1j})^{\gamma_1})^{-\alpha_1})+\eta_{2j}ln(1-(1+(\lambda_2 t_{2j})^{\gamma_2})^{-\alpha_2})\right\}\right) \tag{32}$$

Here onwards we call equations (27), (28), (29), (30), (31), and (32) as Model I, Model II, Model III, Model IV, Model V and Model VI respectively. Model I and Model II are the modified inverse Weibull baseline distribution with and without frailty, Model III and Model IV are the generalized log-logistic baseline distribution type I with and without frailty and likewise Model V and Model VI are the baseline with generalized log-logistic baseline distribution type II with and without frailty.

6. Likelihood Specification and Bayesian Estimation of Parameters

Suppose there are n individuals under study, whose first and second observed failure times are represented by (t_{1j},t_{2j}). Let c_{1j} and c_{2j} be the observed censoring times for the j^{th} individual $(j=1,2,3,...,n)$ for the first and the second recurrence times respectively. We use the left censoring scheme. We assume that the censoring time and the lifetimes of an individual are independently distributed.

The contribution of the bivariate lifetime random variable of the j^{th} individual to likelihood function is given by,

$$L_j(t_{1j},t_{2j}) = \begin{cases} f_1(t_{1j},t_{2j}), & ;\ t_{1j}>c_{1j}, t_{2j}>c_{2j}, \\ f_2(t_{1j},c_{2j}), & ;\ t_{1j}>c_{1j}, t_{2j}<c_{2j}, \\ f_3(c_{1j},t_{2j}), & ;\ t_{1j}<c_{1j}, t_{2j}>c_{2j}, \\ f_4(c_{1j},c_{2j}), & ;\ t_{1j}<c_{1j}, t_{2j}<c_{2j}. \end{cases}$$

and likelihood function is,

$$L(\underline{\psi},\underline{\beta},\theta) = \prod_{j=1}^{n_1} f_1(t_{1j},t_{2j})\prod_{j=1}^{n_2} f_2(t_{1j},c_{2j})\prod_{j=1}^{n_3} f_3(c_{1j},t_{2j})\prod_{j=1}^{n_4} f_4(c_{1j},c_{2j}) \tag{33}$$

where θ, $\underline{\psi}$ and $\underline{\beta}$ are respectively the frailty parameter (σ,ρ), the vector of baseline parameters and the vector of regression coefficients respectively. For without frailty model, likelihood function is

$$L(\underline{\psi},\underline{\beta}) = \prod_{j=1}^{n_1} f_1(t_{1j},t_{2j})\prod_{j=1}^{n_2} f_2(t_{1j},c_{2j})\prod_{j=1}^{n_3} f_3(c_{1j},t_{2j})\prod_{j=1}^{n_4} f_4(c_{1j},c_{2j}) \tag{34}$$

In equation (34) the frailty parameters θ and ρ are absent and in equation (33) they are present. The counts n_1, n_2, n_3 and n_4 be the numbers of individuals for which first and second failure times (t_{1j}, t_{2j}) lie in the ranges $t_{1j} > c_{1j}, t_{2j} > c_{2j}$; $t_{1j} > c_{1j}, t_{2j} < c_{2j}$; $t_{1j} < c_{1j}, t_{2j} > c_{2j}$ and $t_{1j} < c_{1j}, t_{2j} < c_{2j}$ respectively such that $n_1 + n_2 + n_3 + n_4 = n$ and

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{\partial^2 F(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}} \\ f_2(t_{1j}, c_{2j}) &= \frac{\partial F(t_{1j}, c_{2j})}{\partial t_{1j}} \\ f_3(c_{1j}, t_{2j}) &= \frac{\partial F(c_{1j}, t_{2j})}{\partial t_{2j}} \\ f_4(c_{1j}, c_{2j}) &= F(c_{1j}, c_{2j}) \end{aligned} \tag{35}$$

where $\eta_{oj} = e^{(\beta_0 X_0)}, \eta_{1j} = e^{(\beta_1 X_1)}$ and $\eta_{2j} = e^{(\beta_2 X_2)}$. Substituting cumulative reversed hazard rate $M_{01}(t_{1j})$, $M_{02}(t_{2j})$, reversed hazard rate $m_{01}(t_{1j})$, $m_{02}(t_{2j})$ and distribution function $F(c_{1j}, c_{2j})$ for six proposed models into the last relations we get the likelihood function given by equations (33) and (34) for all the six models.

Unfortunately computing the maximum likelihood estimators (MLEs) involves solving a eleven dimensional optimization problem for Model I, Model III and Model V and nine dimensional optimization problem for Model II, Model IV and Model VI. As the method of maximum likelihood fails to estimate the parameters due to convergence problem in the iterative procedure, so we use the Bayesian approach. The traditional maximum likelihood approach to estimation is commonly used in survival analysis, but it can encounter difficulties with frailty models. Moreover, standard maximum likelihood based inference methods may not be suitable for small sample sizes or situations in which there is heavy censoring (see Kheiri *et al.* (2007)). Thus, in our problem a Bayesian approach, which does not suffer from these difficulties, is a natural one, even though it is relatively computationally intensive. To estimate parameters of the model, the Bayesian approach is now popularly used, because computation of the Bayesian analysis become feasible due to advances in computing technology.

To estimate the parameters of the model, the Bayesian approach is now popularly used, because computation of the Bayesian analysis become feasible due to advances in computing technology. Several authors have discussed Bayesian approach for the estimation of parameters of the frailty models. Some of them are, Ibrahim *et al.*(2001) and references their in, Santos and Achcar (2010). Santos and Achcar (2010) considered parametric models with Weibull and generalized gamma distribution as baseline distributions and gamma, log-normal as frailty distributions. Ibrahim *et al.* (2001) and references therein considered Weibull model and piecewise exponential model with gamma frailty. They also considered positive stable frailty models.

The joint posterior density function of parameters for given failure times is obtained as,

$$\begin{aligned} \pi(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) &\propto L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \\ &\times g_1(\alpha_1) g_2(\lambda_1) g_3(\gamma_1) g_4(\alpha_2) g_5(\lambda_2) g_6(\gamma_2) g_7(\theta) \prod_{i=1}^{5} p_i(\boldsymbol{\beta}_i) \end{aligned}$$

where $g_i(.)$ $(i = 1, 2, \cdots, 7)$ indicates the prior density function with known hyper parameters of corresponding arguments for baseline parameters and frailty variance; $p_i(.)$ is prior density function for regression coefficient β_i; $\boldsymbol{\beta}_i$ represents a vector of regression coefficients except β_i, $i = 1, 2, \ldots, k$ and likelihood function $L(.)$ is given by equation (33) or (34). Here we assume that all the parameters are independently distributed.

To estimate the parameters of the model, we used Metropolis-Hastings algorithm and Gibbs sampler. We monitored the convergence of a Markov chain to a stationary distribution by Geweke test (Geweke 1992)and Gelman-Rubin Statistics (Gelman and Rubin, 1992). Trace plots, coupling from the past plots and sample autocorrelation plots are used to check the behaviour of the chain, to decide burn-in period and autocorrelation lag respectively.

Algorithm consists in successively obtaining a sample from the conditional distribution of each of the parameter given all other parameters of the model. These distributions are known as full conditional distributions. In our case full conditional distributions are not easy to integrate out. So full conditional distributions are obtained by considering that they are proportional to the joint distribution of the parameters of the model.

We have full conditional distribution of the parameter α_1 with frailty as,

$$\pi_1(\alpha_1 \mid \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \quad \propto \quad L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \cdot g_1(\alpha_1) \tag{36}$$

We have full conditional distribution of the parameter α_1 without frailty as,

$$\pi_1(\alpha_1 \mid \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \boldsymbol{\beta}) \quad \propto \quad L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \boldsymbol{\beta}) \cdot g_1(\alpha_1) \tag{37}$$

Similarly full conditional distributions for other parameters can be obtained.

To evaluate the performance of the Bayesian estimation procedure we carry out a simulation study. For the simulation purpose we have considered only one covariate X_0 which we assume to follow binomial distribution. The frailty variable Z_1 and Z_2 are assumed to have inverse Gaussian distribution with known variance and correlation ρ . Lifetimes (T_{1j}, T_{2j}) for j^{th} individual are conditionally independent for given frailty $Z_{1j} = z_{1j}$ and $Z_{2j} = z_{2j}$. We assume that $T_{ij}(i = 1, 2.; j = 1, 2, \cdots, n)$ follows one of the baseline distribution modified inverse Weibull distribution, Generalized log-logistic distribution type I and Generalized log-logistic distribution type II. As the Bayesian methods are time consuming, we generate only twenty five pairs of lifetimes.

A widely used prior for frailty parameters σ, are the gamma distribution $G(0.0001, 0.0001)$. In addition, we assume that the regression coefficients are normal with mean zero and large variance say 1000. Similar types of prior distributions are used in Ibrahim *et al.* (2001), Sahu *et al.* (1997) and Santos and Achcar (2010). So in our study we also use same non informative prior for frailty parameters σ, and regression coefficients $\beta_i, i = 1, .., 5$. Since we do not have any prior information about baseline parameters, $\lambda_1, \gamma_1, \alpha_1, \lambda_2, \gamma_2$ and α_2, prior distributions are assumed to be flat. We consider two different non-informative prior distributions for baseline parameters, one is $G(a_1, a_2)$ and another is $U(b_1, b_2)$. All the hyper-parameters a_1, a_2, b_1 and b_2 are known. Here $G(a, b)$ is the gamma distribution with the shape parameter a and the scale parameter b and $U(b_1, b_2)$ represents uniform distribution over the interval (b_1, b_2). For correlation parameter we use uniform distribution $U(0, 1)$. We use different value of baseline parameters for Model I, Model III and Model V, details are given in Table

1, 2 and 3. We assume the value of the hyper-parameters as $a_1 = 1, a_2 = 0.0001, b_1 = 0$ and $b_2 = 100$.

We run two parallel chains for model one using two sets of prior distributions with the different starting points using Metropolis-Hastings algorithm and Gibbs sampler based on normal transition kernels. We iterate both the chains for 100000 times. There is no effect of prior distribution on posterior summaries because the estimates of parameters are nearly the same and the convergence rate of Gibbs sampler for both the prior sets is almost the same. Also for both the chains the results were somewhat similar. For all models, the trace plots, the coupling from the past plots, the running mean plots and the sample autocorrelation plots for the simulation study are not provided due to lack of space. Table 1, 2 and 3 presents the estimates, the credible intervals of the parameters for the Model I, Model III and Model V based on the simulation study. These also contains the Gelman-Rubin (Gelman and Rubin, 1992) convergence statistic and the Geweke test (Geweke, 1992) for all the parameters of the Model I, Model III and Model V based on the simulation study. The Gelman-Rubin convergence statistic values are nearly equal to one and also the Geweke test values are quite small and the corresponding p-values are large enough to say that the chain attains stationary distribution. Simulated values of the parameters have the autocorrelation of lag k. So that every k^{th} iteration is selected as a sample from the posterior distribution.

7. Analysis of Australian Twin data

Duffy *et al.* (1990) considered Australian twin data which consist of information about the age at appendectomy of monozygotic (MZ) and dizygotic (DZ) twins. There were some pairs with missing age at onset and those are the left censored observations. Duffy *et al.* (1990) excluded these left censored observations in the analysis. It is therefore, appropriate to model common random effect by including those left censored observations, which can be done by developing frailty models using RHR. Accordingly, Sankaran and Gleeja (2011) introduced frailty as a common random effect that acts multiplicatively on reversed hazard rates, which is useful for the analysis of left censored data.

Now we apply the all six models to the Australian twin data given in Duffy *et a1.* (1990). The data consists of six zygote categories. We consider the subset of the data with zygote category 4. The data consists of males gender only and consist if 350 pair of twins with 9 and 11 censored in twin 1 and twin 2 respectively. An individual having age at onset less than 11 are considered as left censored observations. The data has information on the age at onset at appendectomy of twins. The genetic effect involved in the risk of appendectomy is the frailty variable. Here there is a common covariate age of twins for both T_1 and T_2 and one covariate each for T_1, T_2, i.e., presence or absence of appendectomy. To check goodness of fit of Australian twin data set, We obtain Kolmogorov-Smirnov(K-S) statistics and their p-values for T_1 and T_2. For Model I, Model III and Model V p-values of observe that p-values for Kolmogorov-Smirnov (K-S) statistics are provided in Table 4. Thus from p values of K-S test are quite high. We can say that there is no statistical evidence to the reject the hypothesis that data are from these three models.

As in case of simulation, here also we assume the same set of prior distributions. We run two parallel chains for all models using two sets of prior distributions with the different starting points using the Metropolis-Hastings algorithm and the Gibbs sampler based on

normal transition kernels. We iterate both the chains for 100000 times. As seen in simulation study here also we got nearly same estimates of parameters for both the set of priors, so estimates are not dependent on the different prior distributions. Convergence rates of Gibbs sampler for both the prior sets are almost the same. Also both the chains show somewhat similar results, so we present here the analysis for only one chain with $G(1, 0.0001)$ as prior for baseline parameters and $G(0.0001, 0.0001)$ as the prior for the frailty parameter σ^2. Due to lack of space we are presenting only for model one(trace plots and coupling from the past plots) for the parameters. Trace plots for all the parameters shows zigzag pattern which indicates that parameters move and mix more freely. Thus, it seems that the Markov chain has reached the stationary state. Burn in period is decided by using coupling from the past plot. However, a sequence of draws after burn-in period may have autocorrelation. Because of autocorrelation consecutive draws may not be random, but values at widely separated time points are approximately independent. So, a pseudo random sample from the posterior distribution can be found by taking values from a single run of the Markov chain at widely spaced time points (autocorrelation lag) after burn-in period. The autocorrelation of parameters become almost negligible after the certain lag.

The Gelman-Rubin convergence statistic values are nearly equal to one and the Geweke test statistic values are quite small and corresponding p-values are large enough to say the chains attains stationary distribution. The posterior mean and standard error with 95% credible intervals for baseline parameters, frailty parameter and regression coefficients are presented in Tables 5-10. The posterior summery of the Model I, Model II, Model III, Model IV, Model V and Model VI are given in Tables 5, 6, 7, 8, 9 and 10. Tables 5, 6, 7, 8, 9 and 10 presents estimates, credible intervals, Geweke test and Gelman-Rubin statistics for all the parameters of the Model I, Model II, Model III, Model IV, Model V and Model VI respectively, based on data. For Model I, Model III and Model V the estimates of frailty parameter σ are respectively 5.6081, 5.4875 and 4.7686. This shows that there is a heterogeneity between the pairs of twins. Bayes factor for Model I with Model II is 32.80, for Model III and Model IV is 298.41 and Model V with Model VI is 1704.12. This is also a Bayesian test based on Bayes factor for testing $\sigma^2 = 0$ against $\sigma^2 > 0$ and which supports the alternative hypothesis, i.e., models with frailty fits better. The credible interval of regression coefficient β_0 does not contain zero for all models except, Model VI. The credible interval of regression coefficient β_1 contain zero for all models except, Model III and Model V. The credible interval of regression coefficient β_2 contain zero for all models. Hence age is the significant covariate for Model I, Model II, Model III, Model IV and Model V. The convergence rate of Gibbs sampling algorithm does not depend on these choices of prior distributions in our proposed model for Australian twin data. The Geweke test values are near to zero and corresponding p-values are quite high and the Gelman-Rubin Statistics for all the parameters of all six models based on data are very close to one.

To compare six models we first use Aikaike information criteria (AIC), Bayesian information criteria (BIC) and deviance information criteria (DIC) values which are given in Table 11 and Bayes factor in Table 12. The AIC, BIC and DIC values for Model V is least among all six models. On the basis of AIC, BIC and DIC values Model V is the best among all six models. Similarly the Bayes factor show that models with frailty (Model I, Model III and Model V) are better than the models without frailty and Model V, the correlated inverse Gaussian frailty based on reversed hazard rate with generalised log-logistic type II baseline is the best and the frailty is significant.

8. Conclusions

Our main aim of the study is to examine the role of the bivariate correlated frailty model based on the reversed hazard rate in survival studies. For this we used the correlated inverse Gaussian frailty model with the modified inverse Weibull distribution, generalized log-logistic type I and generalized log-logistic type II as a baseline distribution and these models are compared with their baseline model based on reversed hazard rate. We also found that the correlated inverse Gaussian frailty models are better models as compared to their baseline model on the basis of AIC, BIC and DIC values for Australian twin data set. Bayes factor support the correlated frailty models.

Initially we thought to use the method of maximum likelihood to estimate the parameters but likelihood equations do not converge and the method of maximum likelihood fails to estimate the parameters so we used the Bayesian approach. In this study, the model is specified in a Bayesian framework and estimated with the MCMC algorithms. The estimates of the parameters are not dependent on the different prior distributions.

Two different chains were run for the proposed models from different starting points using the Metropolis-Hastings algorithm within Gibbs sampler. We have provided 100,000 iterations to perform the simulation study. Estimates were calculated after discarding a burn-in interval for each chain. The quality of convergence was checked by Gelman-Rubin statistics. The values of the Gelman-Rubin statistics in this case are quite close to one and also the Geweke test values are small with large p-values. Thus the sample can be considered to have arisen from stationary distribution and descriptive statistics can be seen as valid estimates of unknown parameters. The simulation results indicate that the performance of the Bayesian estimation method is quite satisfactory. Bayes factor is used to test the frailty parameter $\sigma^2 = 0$ and it is observed that the frailty parameter is highly significant in all frailty models. From Table 12 it is clear that the models with frailty fit better than without frailty models and Model V is best among the all six models. Age is the significant for all the models except Model VI.

The choice of the best model for Australian twin data is based on AIC, BIC, DIC and Bayes factor values. We found that Model V is a best Model on the basis of AIC, BIC, DIC and Bayes factor values. The age is the significant covariate for all models except Model IV. Correlated inverse Gaussian frailty models(Model I, Model III and Model V) are better than their baseline model. We also compare with correlated gamma frailty models suggested by Hanagal and Pandey (2017) and observe that correlated inverse Gaussian frailty based on reversed hazard rate with generalized log-logistic type II baseline performs better and more suitable than the correlated gamma frailty models proposed by Hanagal and Pandey (2017) for Australian twin data set, with left censored observations. The methods discussed in this paper may be extended into other frailty models and correlated frailty models with different baseline distributions, using the Bayesian approach, provided the models fit to the data.

Acknowledgements

I thank the referee for the valuable suggestions and comments.

References

Barlow, R. E., Marshal, A. W. and Proschan, F. (1963). Properties of the probability distribution with monotone hazard rate. *Annals of Mathematical Statistics*, **34**, 375-389.

Bennett, S. (1983). "Log-logistic regression model for survival data". *Applied Statistics*, **32(2)**, 165-171.

Block, H. W., Savits, T. H. and Singh, H. (1998)." On the reversed hazard rate function", *Probability in the Engineering and Informational Sciences*, **12**, 69-90.

Clayton, D. G. (1978). A model for association in bivariate life tables and its applications to epidemiological studies of familial tendency in chronic disease incidence. *Biometrica*, **65**, 141-151.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of Royal Statistical Society, Ser., A*, **148**, 82-117.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman and Hall, New York.

Crowder, M. (1985). A distributional model for repeated failure time measurements. *Journal of Royal Statistical Society, Ser. B*, **47**, 447-452.

Duffy, D. L., Martin, N. G. and Mathews, J. D. (1990). Appendectomy in Australian twins. *Australian Journal of Human Genetics*, **47(3)**, 590-92.

Gelman, A. and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. *In Bayesian Statistics 4 (J. M. Bernardo, J. 0.Berger, A. P. Dawid and A. F. M. Smith, eds.)* . Oxford Univ. Press. pp. 625-632.

Genest, C. and Mackay, J. (1986). Joy of Copulas: Bivariate distributions with uniform marginals. *The American Statistician*, **40(4)**, 280-283.

Geweke, J. (1992). "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." *In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith)*, Oxford: Oxford University Press, pp. 169-193.

Hanagal, D. D. (2006). A gamma frailty regression model in bivariate survival data. *IAPQR Transactions*, **31**, 73-83.

Hanagal, D. D. (2007). Gamma frailty regression models in mixture distributions. *Economic Quality Control*, **22(2)**, 295-302.

Hanagal, D. D. (2011). *Modeling Survival Data Using Frailty Models*. Chapman & Hall/CRC, Boca Rotan.

Hanagal, D. D. (2017). Frailty Models in Public Health. *Handbook of Statistics, Elsevier Publishers; Amsterdam*, **37(B)**, 209-247.

Hanagal, D. D. (2019). *Modeling Survival Data Using Frailty Models. Second Edition*. Springer Nature, Singapore.

Hanagal, D. D. and Bhambure, S. M. (2014). Shared inverse Gaussian frailty model based on reversed hazard rate for modeling Australian twin data. *Journal of Indian Society for Probability and Statistics*, **15**, 9-37.

Hanagal, D. D. and Bhambure, S. M. (2015). Comparison of shared gamma frailty models using Bayesian approach. *Model Assisted Statistics & Applications*, **10**, 25-41.

Hanagal, D. D. and Bhambure, S. M. (2016). Modeling bivariate survival data using shared inverse Gaussian frailty model. *Communications in Statistics, Theory & Methods*, **45(17)**, 4969-4987.

Hanagal, D. D. and Dabade, A. D. (2013). Modeling of inverse Gaussian frailty model for bivariate survival data. *Communications in Statistics, Theory & Methods*, **42(20)**, 3744-3769.

Hanagal, D. D. and Pandey, A. (2014a). Inverse Gaussian shared frailty for modeling kidney infection data. *Advances in Reliability*, **1**, 1-14.

Hanagal, D. D. and Pandey, A. (2014b). Gamma shared frailty model based on reversed hazard rate for bivariate survival data. *Statistics & Probability Letters*, **88**, 190-196.

Hanagal, D. D. and Pandey, A. (2015a). Gamma frailty models for bivarivate survival data. *Journal of Statistical Computation and Simulation*, **85(15)**, 3172-3189.

Hanagal, D. D. and Pandey, A. (2015b). Inverse Gaussian shared frailty models with generalized exponential and generalized inverted exponential as baseline distributions. *Journal of Data Science*, **13(2)**, 569-602.

Hanagal, D. D. and Pandey, A. (2016). Inverse Gaussian shared frailty models based on reversed hazard rate. *Model Assisted Statistics and Applications*, **11**, 137-151.

Hanagal, D. D. and Pandey, A. (2017). Correlated Gamma Frailty Models for Bivariate Survival Data Based on Reversed Hazard Rate. *International Journal of Data Science*, **2(4)**, 301-324.

Hanagal, D. D. and Pandey, A. and Sankaran, P. G. (2017a). Shared frailty model based on reversed hazard rate for left censoring data. *Communications in Statistics, Simulation and Computation*, **46(1)**, 230-243.

Hanagal, D. D., Pandey, A. and Ganguly, A. (2017b). Correlated gamma frailty models for bivariate survival data. *Communications in Statistics, Simulation and Computation*, **46(5)**, 3627-3644.

Hanagal, D. D. and Sharma, R. (2013). Modeling heterogeneity for bivariate survival data by shared gamma frailty regression model. *Model Assisted Statistics and Applications*, **8**, 85-102.

Hanagal, D. D. and Sharma, R. (2015a). Bayesian inference in Marshall-Olkin bivariate exponential shared gamma frailty regression model under random censoring. *Communications in Statistics, Theory and Methods*, **44(1)**, 24-47.

Hanagal, D. D. and Sharma, R. (2015b). Comparison of frailty models for acute leukaemia data under Gompertz baseline distribution. *Communications in Statistics, Theory & Methods*, **44(7)**, 1338-1350.

Hanagal, D. D. and Sharma, R. (2015c). Analysis of bivariate survival data using shared inverse Gaussian frailty model. *Communications in Statistics, Theory & Methods*, **44(7)**, 1351-1380.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, **73**, 671-678.

Hougaard, P. (1984). Life table methods for heterogeneous populations. *Biometrika*, **71(1)**, 75-83.

Hougaard, P. (1987). Modelling Multivariate Survival. *Scandinavian Journal of Statistics*, **14(4)**, 291-304.

Iachine, I. A. (1995a). Correlated frailty concept in the analysis of bivariate survival data. *Bachelor project, Department of Mathematics and Computer Science, Odense University*, Denmark.

Iachine, I. A. (1995b). Parameter estimation in the bivariate correlated frailty model with observed covariates via the EM-algorithm. *Working Paper Series: Population Studies of Aging 16, CHS, Odense University*, Denmark.

Ibrahim, J. G., Ming-Hui C. and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, Verlag.

Jeffreys, H. (1961). *Theory of Probability. 3rd edn.*, Oxford: Oxford University Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factor. *Journal of the American Statistical Association*, **90(430)**, 773-795.

Kheiri, S., Kimber, A. and Meshkani M. R. (2007). Bayesian analysis of an inverse Gaussian correlated frailty model. *Computational Statistics and Data Analysis*, **51**, 5317-5326.

Mohammed, A. El-Saidi, Singh, Karan P. and Alfred A. Bartolucci, (1990). A Note on a Characterization of the Generalized Log-Logistic Distribution. *Environmetrics*, **1(4)**,337-342.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, **19**, 25-43.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of Royal Statistical Society, Series B*, **44**, 414-422.

Oakes, D. (1989). Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association*, **84(406)**, 487-493.

O'Quigley, J. and L. Struthers (1982), "Survival models based upon the logistic and loglogistic distributions". *Computer Programs in Biomedical Researches*, **15**, 3-12.

Pickles, A., Crouchley, R., Simonoff, E., Eaves, L., Meyer, J., Rutter, M., Hewitt, J. and Silberg, J. (1994). Survival models for developmental genetic data: age of onset of puberty and antisocial behavior in twins. *Genetic Epidemiology*, **11**, 155-170.

Raftey, A. E. (1994). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Biometrika*, **83(2)**, 251-266.

Sankaran P. G. and Gleeja V. L. (2011). On proportional reversed hazards frailty models, *Metron*, **69**, 151-173.

Sahu, S. K., Dey, D. K., Aslanidou, H. and Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Life Time Data Analysis*, **3**, 123-137.

Santos, C. A. and Achcar, J. A.(2010). A Bayesian analysis for multivariate survival data in the presence of covariates. *Journal of Statistical Theory and Applications*, **9**, 233-253.

Schwarz. G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6(2)**, 461-464.

Shaked, M. and Shantikumar, J. G. (1994). *Stochastic Orders and Their Applications*. Academic Press, New York.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002): Bayesian measure of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583-639.

Vaupel, J. W. (1991). Relatives risks: Frailty models of life history data. *Theoretical Population Biology*, **37 (1)**, 220-234.

Vaupel, J. W., Yashin, A. I., Hauge, M., Harvald, B., Holm, N., and Liang X. (1991). Survival analysis in genetics: Danish twins data applied to gerontological question. *J. P. Klein and P. K. Goel (eds.), Survival Analysis: State of the Art*, Kluwer Academic Publisher. pp. 121-138.

ANNEXURE

Table 1: Baseline Distribution Modified inverse Weibull Distribution Model I with Correlated Inverse Gaussian Frailty (Simulation for Model I)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 2500; autocorrelation lag = 250							
α_1 (2.0)	1.9894	0.1319	1.7806	2.2365	-0.00670	0.4973	0.9999
λ_1 (1.5)	1.5197	0.0188	1.4725	1.5510	-0.00085	0.4996	1.0003
γ_1 (2.5)	2.5306	0.1482	2.2461	2.7781	0.00052	0.5002	1.0020
α_2 (2.2)	2.2126	0.0370	2.1228	2.2726	-0.00321	0.4987	1.0007
λ_2 (2.5)	2.5197	0.0186	2.4721	2.5474	-0.02074	0.4917	1.0332
γ_2 (3.0)	3.0537	0.1413	2.7424	3.2960	-0.00574	0.4977	1.0006
σ (2.0)	2.0660	0.0544	1.8817	2.1360	0.00084	0.5003	1.0031
ρ (0.7)	0.7349	0.0324	0.6458	0.7785	-0.00578	0.4976	1.0047
β (0.50)	0.5131	0.0343	0.4157	0.5812	-0.00229	0.4990	1.0059

Table 2: Baseline Distribution Generalized Logistic Distribution Type I Model-III with Correlated Inverse Gaussian Frailty(Simulation for Model III)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 7000; autocorrelation lag = 350							
α_1 (2.0)	1.8692	0.0528	1.7779	2.0937	0.00261	0.5010	0.9999
λ_1 (1.5)	1.5621	0.0211	1.5032	1.5786	0.00171	0.5006	1.0031
γ_1 (2.5)	2.3454	0.0836	2.2347	2.6174	-0.01263	0.4949	1.0099
α_2 (2.2)	2.0152	0.0472	1.9713	2.2819	0.00531	0.5021	1.0138
λ_2 (2.5)	2.5667	0.0425	2.4663	2.6287	0.00511	0.5020	1.0043
γ_2 (2.5)	2.4307	0.1117	2.2773	2.6411	-0.00644	0.4974	1.0015
σ (0.20)	0.2304	0.0085	0.2091	0.2396	0.01255	0.5050	1.0071
ρ (0.7)	0.7686	0.0745	0.6106	0.8548	0.00563	0.5022	1.0001
β (0.50)	0.4879	0.0149	0.4702	0.5105	-0.02020	0.4919	1.0005

Table 3: Baseline Distribution Generalized Logistic Distribution Type II Model-V with Correlated Inverse Gaussian Frailty(Simulation for Model V)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 7000; autocorrelation lag = 250							
α_1 (2.0)	2.0512	0.1433	1.7726	2.2631	-0.00394	0.4984	1.0012
λ_1 (1.5)	1.5015	0.0218	1.4812	1.5492	0.00149	0.5005	1.0027
γ_1 (3.5)	3.5009	0.1216	3.3806	3.6401	0.00061	0.5002	1.0009
α_2 (2.2)	2.1858	0.1226	1.9576	2.4178	-0.01473	0.4941	0.9999
λ_2 (2.5)	2.5055	0.0212	2.4715	2.5318	0.00056	0.5002	1.0004
γ_2 (3.5)	3.4905	0.1519	3.4567	3.6781	-0.00213	0.4991	1.0097
σ (0.2)	0.21811	0.0198	0.1801	0.2191	-0.00431	0.4982	1.0047
ρ (0.7)	0.7125	0.1011	0.6128	0.7867	-0.00015	0.4999	1.0055
β (0.50)	0.4888	0.0402	0.4557	0.5549	-0.00171	0.4993	1.0000

Table 4: p-values of K-S Statistics for Goodness of Fit Test for Australian Twin Data Set

Distribution	Recurrence time First	Second
Model I	0.57402	0.59688
Model III	0.85443	0.7794
Model V	0.99977	0.99787

Table 5: Posterior Summary for Australian Twin Data Set (Model I)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 7500; autocorrelation lag = 1300							
α_1	44.1762	2.7887	39.12322	48.7418	0.01771	0.5070	1.0068
λ_1	0.4632	0.0311	0.4111	0.5071	-0.00162	0.4993	1.0341
γ_1	0.1227	0.0042	0.1106	0.1312	-0.00224	0.4991	1.0034
α_2	41.1201	2.9285	35.816	46.3785	0.00857	0.5034	1.0192
λ_2	0.4574	0.0217	0.4161	0.4989	0.00243	0.5009	1.0041
γ_2	0.2011	0.0035	0.1913	0.2112	-0.01221	0.4951	1.0066
ρ	0.9294	0.0414	0.8424	0.9978	-0.00069	0.4997	1.0099
σ	5.6081	0.0651	5.4172	5.7571	-0.01036	0.4958	1.0054
β_0	0.0209	0.0023	0.0133	0.0304	-0.01873	0.4925	1.0000
β_1	-0.0742	0.0641	-0.2287	0.1041	-0.00320	0.4987	1.0086
β_2	-0.0312	0.0204	-0.0564	0.0161	-0.00684	0.4972	1.0045

Table 6: Posterior Summary for Australian Twin Data Set (Model II)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 12000; autocorrelation lag = 180							
α_1	11.44908	0.5554	10.47060	12.34491	-0.00330	0.4986	1.0075
λ_1	0.06949	0.0112	0.04948	0.08727	0.01716	0.5068	1.0043
γ_1	0.10227	0.0036	0.09499	0.10941	-0.00179	0.4992	1.0065
α_2	10.43929	0.5275	9.44735	11.34131	-0.00460	0.4981	1.0008
λ_2	0.07101	0.0106	0.05109	0.08880	-0.00714	0.4971	1.0031
γ_2	0.09919	0.0038	0.09192	0.10693	9.27e-05	0.5001	0.9999
β_0	0.00575	0.0020	0.00152	0.00950	0.003835	0.5015	1.0000
β_1	-0.01649	0.0715	-0.14075	0.12942	-0.007237	0.4971	1.0008
β_2	0.06323	0.1238	-0.17651	0.29295	0.004567	0.5018	1.0000

Table 7: Posterior Summary for Australian Twin Data Set (Model III)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 7500; autocorrelation lag = 1100							
α_1	18.5175	0.2150	18.0612	18.8641	0.0015	0.5006	1.0004
λ_1	0.0701	0.0012	0.0676	0.0722	-0.0259	0.4896	1.0095
γ_1	23.1861	0.2119	22.6789	23.4689	-0.0056	0.4977	1.0000
α_2	18.0016	0.5215	17.1468	18.1801	0.0022	0.5008	1.0003
λ_2	0.0801	0.0012	0.0771	0.0823	-0.0254	0.4898	1.0149
γ_2	24.1014	0.2182	23.6952	24.4461	-0.0043	0.4982	1.0026
ρ	0.8941	0.0151	0.8721	0.9078	0.0026	0.5010	0.9999
σ	5.7845	0.1155	5.5526	5.9101	-0.0061	0.4975	1.0039
β_0	0.8465	0.0290	0.8161	0.8722	-0.0235	0.4906	1.0083
β_1	-0.0507	0.0277	-0.0971	-0.0052	0.0042	0.5017	1.0134
β_2	-0.0143	0.0314	-0.0426	0.0413	0.0113	0.5045	1.0321

Table 8: Posterior Summary for Australian Twin Data Set (Model IV)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 6500; autocorrelation lag = 200							
α_1	1.4806	0.2172	1.08298	1.87900	-0.00716	0.4971	1.0091
λ_1	0.0475	0.0031	0.04162	0.05341	-0.00838	0.4966	1.0008
γ_1	1.3022	0.1518	3.29435	3.93994	0.00624	0.5024	1.0017
α_2	0.0456	0.0026	1.00322	1.56740	-0.00976	0.4961	1.0096
λ_2	3.6265	0.1957	0.04062	0.05151	-0.00932	0.4962	1.0088
γ_2	3.6135	0.1680	3.25814	3.99208	0.01088	0.5043	1.0018
β_0	0.0059	0.0026	0.00068	0.01158	-0.00174	0.4993	1.0005
β_1	8.9e-06	0.0024	-0.00437	0.00442	-0.00050	0.4997	1.0054
β_2	0.0592	0.1285	-0.20118	0.29163	-7.76e-05	0.4999	1.0134

Table 9: Posterior Summary for Australian Twin Data Set (Model V)

Parameter (value)	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 6500; autocorrelation lag = 300							
α_1	0.2817	0.0032	0.2781	0.2877	-0.00066	0.4997	1.0002
λ_1	0.0701	0.0025	0.0585	0.0724	-0.00274	0.4989	1.0004
γ_1	55.4383	1.1412	52.9768	57.2222	0.00761	0.5030	1.0026
α_2	0.1051	0.0031	0.0891	0.1108	0.00678	0.5027	1.0069
λ_2	0.0706	0.0012	0.0687	0.0728	-0.00166	0.4993	1.0012
γ_2	58.6274	1.6105	55.7728	61.0344	-0.00435	0.4983	1.0057
ρ	0.8824	0.0242	0.8461	0.9165	0.00041	0.5001	0.9999
σ	4.7686	0.0505	4.5495	4.8869	0.00315	0.5012	0.9999
β_0	0.0785	0.0057	0.0751	0.0903	0.00317	0.5012	1.0002
β_1	-0.0412	0.0202	-0.0819	-0.0051	0.01608	0.5064	1.0018
β_2	-0.0214	0.0247	-0.0615	0.0221	-0.00370	0.4985	1.0136

Table 10: Posterior Summary for Australian Twin Data Set (Model VI)

Parameter (value)	**Estimate**	**Standard Error**	**Lower Credible Limit**	**Upper Credible Limit**	**Geweke values**	**p values**	**Gelman & Rubin values**
burn in period = 6500; autocorrelation lag = 200							
α_1	0.6047	0.1027	0.4205	0.8341	0.01169	0.5047	1.00
λ_1	0.0475	0.0026	0.0428	0.0526	-0.00679	0.4972	1.03
γ_1	5.1896	0.5964	4.1191	6.4377	-0.01175	0.4953	1.00
α_2	0.6736	0.0885	0.5048	0.8412	0.01642	0.5065	1.00
λ_2	0.0463	0.0029	0.0406	0.0524	-0.00860	0.4965	1.00
γ_2	4.7336	0.4456	3.9453	5.7410	-0.01335	0.4946	1.01
β_0	0.0042	0.0041	-0.0041	0.0119	0.00415	0.5016	1.00
β_1	-0.0013	0.0239	-0.0441	0.0452	0.01221	0.5048	1.00
β_2	0.0481	0.1225	-0.1985	0.2838	-0.00571	0.4977	1.01

Table 11: AIC, BIC and DIC Comparison

Model	**AIC**	**BIC**	**DIC**
Model- I	5155.713	5188.813	5113.985
Model- II	5384.161	5426.847	5375.313
Model- III	5071.699	5082.212	5057.809
Model- IV	5355.809	5396.766	5351.894
Model- V	5016.714	5018.908	5003.065
Model- VI	5781.328	5901.931	5935.093

Table 12: Bayes Factors for Four Models

-	M12	M31	M14	M51	M16	M32	M42	M52
Bayes Factor	32.80	302.08	5.59	336.81	1274.5	301.2	27.40	338.49
-	M26	M34	M53	M36	M54	M46	M56	-
Bayes Factor	1268.83	298.41	32.79	1582.82	316.8	1271.11	1704.12	-

$$Mij = 2 * ln(\frac{Ii}{Ij})$$

Statistics and Applications {ISSN 2454 -7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 113-139

Efficient Use of Non-Sensitive Auxiliary Variable under Scrambled Randomized Response Mechanism for Estimating Sensitive Population Mean in Successive Sampling

Housila P. Singh and Preeti Patidar
School of Studies in Statistics, Vikram University, Ujjain, (M.P.), India.

Received: 01 July 2020; Revised: 28 February 2021; Accepted: 02 March 2021

Abstract

This paper addresses the problem of estimation of population mean of a sensitive variable under investigation using scrambled randomized response mechanism in presence of non-sensitive auxiliary variable at current move in two occasion successive sampling. The proposed estimator is studied under scrambled randomized response models. The detail properties of the suggested estimators have been provided. To measure the scrambled model effect the envisaged estimators are compared with direct estimators. Optimum replacement policy has been elaborated. Numerical study is carried out to demonstrate the applicability of the propounded estimators and hence appropriate recommendations are given.

Key words: Scrambled response mechanism; Successive sampling; Non-sensitive auxiliary variable; Bias; Mean squared error.

AMS Classification: 62D05

1. Introduction

Surveys covering human population associated with sensitive issues, for instance, drug addiction, induced abortion, HIV infection status, excessive gambling, incidence of domestic violence, illegitimacy of offspring, drinking and driving, social security frauds, tax evasion, substance abuse, alcoholism, illegal income and AIDS *etc*. need to be addressed in social, psychological, socioeconomic and biometric research. In such cases employing the derived method of interview, people do not respond truthfully on even refuse to respond owing to social stigma and/ or fear. Under such circumstances, to gather valid and reliable data, protect respondent confidentiality and avoid unacceptable rate of non-response, randomized response procedures pioneered by Warner (1965) may be employed. Later Horvitz *et al.* (1967), Greenberg *et al.* (1969), Chaudhari and Mukerjee (1988), Kuk (1990), Mangat and Singh (1990), Christofides (2003), Mangat (1994), Arnab (2011) and Chaudhari and Christofides (2013) introduced various other devices for obtaining information on sensitive questions.

Pollock and Bek (1976) and Eichhorn and Hayre (1983) have made initial efforts to take sword of scrambled response technique ahead. Later various authors including Singh and Joardar (1996), Bar-Lev *et al.* (2004), Saha (2007), Diana and Perri (2012), Gjestvang and Singh (2009), Odumade and Singh (2009), Singh and Mathur (2005), Singh and Kim (2007),

Corresponding Author: Preeti Patidar
E-mail: p.patidarstatujn@gmail.com

Tarray and Singh (2015), Arcos *et al.* (2015), Singh and Gorey (2017), and many more have discussed the problem of estimating the population mean of a sensitive variable under scrambled randomized response model.

It is to be mentioned that the above work done for single time survey associated with sensitive character analysis; instead, these issues need to be tracked constantly over time so that reflection of actual scenario in society associated with sensitive issues as well as changed level of sensitivity of issues with respect to time may be understood in a better way [see Priyanka and Trisandhya (2018)]. Interesting aspect of the scrambled response mechanism is that it can be used to protect the anonymity of individuals who have provided sensitive information. In such situations, the use of successive sampling scheme can be attractive alternative to improve the estimators of level at a point in time or to measure the change between two time points. Arnab and Singh (2013, pp. 2499-2500) have given the following examples well suited to the above mentioned situations: (i) A police department may be interested to know the average number of rapes in a large city during a particular year or a change in the number of rapes over a period of couple of years; (ii) A university administration may be interested to know the total amount of a particular drug used by students in a particular year, and after educating students about the adverse effect of drug use on health and society, if there is any significant change in the total drug use on campus or not; (iii) A social organization could be interested to know the proportion of those people who truly quit the drug after attending a lecture or seminar or after taking a medication.

Jessen (1942) first introduced the successive sampling procedure on two occasions to estimate the mean on the most recent (current) occasion. Later several authors including Patterson (1950), Narain (1953), Singh (1968), Ghangurde and Rao (1969), Sen (1973), Okafor and Arnab (1987), Biradar and Singh (2001), Singh and Priyanka (2008), Singh *et al.* (2008), Singh and Vishwakarma (2007, 2009), Singh and Pal (2017) *etc*. have paid their attention toward the estimation of mean on current occasion using successive sampling. Arnab and Singh (2013), and Yu *et al.* (2015), have used randomized response technique to deal with sensitive issues on successive occasion. Singh *et al.* (2017) applied scramble response procedure using Patterson (1950) method to tackle sensitive issues on successive occasion. Assuming non-sensitive additional auxiliary information is available at both occasions, Priyanka *et al.* (2017) and Priyanka and Trisandhya (2018, 2019) have employed both randomized and scramble response procedure to cope up with the studies related to sensitive issues on successive occasions. For example, we consider a situation, where an investigator is interested in estimating the average monthly expenditure on drug usage by undergraduate students in the current year 2016 (*i.e.* at second occasion) designated as the study variable y, then the auxiliary variable x may be taken as the average monthly expenditure on drug usage by undergraduate students in the year 2015 (*i.e.* at first occasion) and the average monthly pocket money of undergraduate students from all sources in the year 2015 may be taken as a non-sensitive additional auxiliary variable z. Here non-sensitive auxiliary data are available at both occasions. Hence this led authors to propose a class of estimators for estimating sensitive population mean of a sensitive variable at current occasion in two occasions successive sampling using non-sensitive auxiliary information. To deal with sensitive issues, randomized response technique due to Gjestvang and Singh (2009) has been applied. The detail properties of the suggested class of estimators have been discussed. Numerical illustration is given in support of the present study.

2. Survey Strategies and Analysis

2.1. Sampling procedure

Let $\Omega = (\Omega_1, \Omega_2, ..., \Omega_N)$ be a finite population of size *N*, which has been sampled over two occasions to estimate the population mean of sensitive variable at current occasion. It is supposed that the units of the population are unchanged over two occasions *i.e.* the sampling frame remain the same there by meaning is that no new units are added or deleted from the population. The character under investigation is sensitive variable designated by $x(y)$ on the first (second) occasion and *z* is a non-sensitive auxiliary variable available at both occasions. At the first occasion, a sample of *n* units is drawn from the population Ω by simple random sampling without replacement (SRSWOR). However, at the second occasion considering the partial overlap case, two independent samples are selected; one is matched sample of size $m = n\lambda$ drawn as subsample from the sample of size *n* and another is unmatched simple random sample of size $u = (n-m) = n\mu$ selected afresh at the second occasion so that the sample size at both the occasions is same (*i.e. n*). The sensitive variable $x(y)$ on the first (second) occasion are perturbed to $g(h)$ with the aid of scrambling variable *W*. The scrambling variable *W* so considered as it may follow any distribution. The following notations are considered further

$\mu\left(=\frac{u}{n}\right)$: Fraction of sample drawn afresh at current occasion,

$\lambda\left(=\frac{m}{n}=1-\mu\right)$: Fraction of samples matched from previous occasion,

$\overline{X}, \overline{Y}, \overline{Z}, \overline{G}, \overline{H}, \overline{W}$: Population means of variables *x, y, z, g, h* and *w* respectively,

$\overline{h}_u, \overline{g}_m, \overline{h}_m, \overline{g}_n$: Sample means of the variate based on sample sizes shown in suffices,

$\overline{z}_u, \overline{z}_m, \overline{z}_n$: Sample means of non-sensitive auxiliary variate *z* based on sample sizes shown in suffices,

$\rho_{yx}, \rho_{xz}, \rho_{yz}, \rho_{gh}, \rho_{hz}, \rho_{gz}$: Correlation coefficient between the variables depicted in suffices,

C_x, C_y, C_z: Coefficient of variation of variables depicted in suffices,

S_x^2, S_y^2, S_z^2 : Population mean square of variability *x, y, z* respectively,

$\sigma_x^2, \sigma_y^2, \sigma_z^2, \sigma_w^2$: Population variance of *x, y, z* and *w* respectively,

Note that the scrambling variable *W* such that $E(W) = \overline{W}$ and $V(W) = \sigma_w^2$.

2.2. Randomized response technique on successive occasions

For estimating the population mean (or) total of a sensitive variable Gjestvang and Singh (2009) suggested a randomized response model (*say* M_{GS}). In this paper Gjestvang and Singh (2009) randomized response model has been modified to be applied on successive occasions.

Let α *and* β be two known positive real numbers. Consider a deck of cards in which p is the proportion of cards bearing the statement: Multiply scrambling variable *W* with α and add to the real value of the sensitive variable $x(y)$ at first (second) move and $(1-p)$ be the

proportion of cards bearing the statement: Multiply scrambling variable W with β and subtract it from the real value of the sensitive variable $x(y)$ at first (second) move. Let $p=\beta/(\alpha+\beta)$ be known. In this procedure each respondent is asked to draw one each secretly (confidentially) and report the scrambled response according at first (second) move accordingly. Using above randomization device, response given by j^{th} respondent on the first and second moves, respectively are described as

$$G_j=\begin{cases}x_j+\alpha W \ \ with\ probability\ \ p\\ x_j-\beta W \ \ with\ probability\ (1-p)\end{cases}$$ and

$$H_j=\begin{cases}y_j+\alpha W \ \ with\ probability\ \ p\\ y_j-\beta W \ \ with\ probability\ (1-p)\end{cases}$$

Therefore applying M_{GS} on two successive occasions, the sensitive variable $x(y)$ are switched to $g(h)$ and are given by

$$G=(X+\alpha W)p+(X-\beta W)(1-p)$$

and

$$H=(Y+\alpha W)p+(Y-\beta W)(1-p)$$

such that $\overline{Y}=\overline{H}$. (1)

$$\rho_{hg}=\frac{\rho_{yx}\sigma_y\sigma_x}{\sqrt{\{\sigma_y^2+\alpha\beta(\sigma_w^2+\overline{w}^2)\}}\sqrt{\{\sigma_x^2+\alpha\beta(\sigma_w^2+\overline{w}^2)\}}},\ \rho_{hz}=\frac{\rho_{yz}\sigma_y}{\sqrt{\{\sigma_y^2+\alpha\beta(\sigma_w^2+\overline{w}^2)\}}},\ \rho_{gz}=\frac{\rho_{xz}\sigma_x}{\sqrt{\{\sigma_x^2+\alpha\beta(\sigma_w^2+\overline{w}^2)\}}}.$$

Remark 1: Strategy is to obtain suitable estimator of population mean of coded response variable $\overline{H}$ on current occasion and substituting the same in (1) to obtain the relevant estimator for sensitive population mean $\overline{Y}$ at current occasion.

2.3. Design of the Class of Suggested Estimators

For estimating the population mean of perturbed variable H on the second (current) occasion, we have suggested two classes of estimators where one class of estimators D_u based on unmatched sample (or afresh sample) of size u on the current (second) occasion and others class of estimators based on the matched sample of size m (which is common to both the occasions).

2.3.1. Class of estimators based on unmatched sample on the second occasion using information on $(\bar{z}_u,\overline{Z})$ of non-sensitive auxiliary variable z

The usual ratio and product-type estimators can be ramified to estimate the population mean of coded response variable. The following estimators based on sample of size u drawn afresh at current occasion for estimating the population mean of switched variable H on current (second) move can be considered

$$d_1=\bar{h}_u(\overline{Z}/\bar{z}_u),\qquad d_2=\bar{h}_u(\bar{z}_u/\overline{Z}),$$

$$d_3=\bar{h}_u\exp\left(\frac{\overline{Z}-\bar{z}_u}{\overline{Z}+\bar{z}_u}\right),\qquad d_4=\bar{h}_u\exp\left(\frac{\bar{z}_u-\overline{Z}}{\bar{z}_u+\overline{Z}}\right),$$

$$d_5 = \bar{h}_u\left[\alpha\left(\frac{\bar{Z}}{\bar{z}_u}\right)+(1-\alpha)\left(\frac{\bar{z}_u}{\bar{Z}}\right)\right],$$

$$d_6 = \bar{h}_u\left[\alpha_1\exp\left(\frac{\bar{Z}-\bar{z}_u}{\bar{Z}+\bar{z}_u}\right)+(1-\alpha_1)\exp\left(\frac{\bar{z}_u-\bar{Z}}{\bar{z}_u+\bar{Z}}\right)\right],$$

$$d_7 = \bar{h}_u\left(\bar{z}_u/\bar{Z}\right)^{\theta_1}, \qquad d_8 = \bar{h}_u\exp\left\{\frac{\delta_1\left(\bar{Z}-\bar{z}_u\right)}{\bar{Z}+\bar{z}_u}\right\},$$

$$d_9 = \bar{h}_u\frac{\bar{Z}}{\left\{\alpha_2\bar{Z}+(1-\alpha_2)\bar{z}_u\right\}}, \qquad d_{10} = \bar{h}_u\left(\bar{Z}/\bar{z}_u\right)^{1/2},$$

$$d_{11} = \bar{h}_u\exp\left\{\frac{\alpha_3\left(\bar{z}_u-\bar{Z}\right)}{\bar{Z}}\right\}, \qquad d_{12} = \bar{h}_u\left[2-\left(\frac{\bar{z}_u}{\bar{Z}}\right)^{\theta_2}\right],$$

$$d_{13} = \bar{h}_u\left[2-\exp\left\{\frac{\theta_3\left(\bar{z}_u-\bar{Z}\right)}{\bar{z}_u+\bar{Z}}\right\}\right], \quad d_{14} = \bar{h}_u\left[2-\left(\frac{\bar{z}_u}{\bar{Z}}\right)^{\theta_2}.\exp\left\{\frac{\theta_3\left(\bar{z}_u-\bar{Z}\right)}{\bar{z}_u+\bar{Z}}\right\}\right],$$

$$d_{15} = \bar{h}_u\exp\left\{\frac{\alpha\left(\bar{Z}-\bar{z}_u\right)}{\bar{Z}+(a_1-1)\bar{z}_u}\right\}, \qquad d_{16} = \bar{h}_u+k\left(\bar{z}_u-\bar{Z}\right),$$

$$d_{17} = \bar{h}_u+\beta_{hz}\left(\bar{z}_u-\bar{Z}\right), \qquad d_{18} = \left\{\bar{h}_u+k\left(\bar{z}_u-\bar{Z}\right)\right\}\exp\left\{\frac{\delta_1\left(\bar{Z}-\bar{z}_u\right)}{\bar{Z}+\bar{z}_u}\right\},$$

$$d_{19} = \left\{\bar{h}_u+k\left(\bar{z}_u-\bar{Z}\right)\right\}\left(\frac{\bar{Z}}{\bar{z}_u}\right)^{\theta_1},$$

etc., where $\alpha,\alpha_1,\alpha_2,\alpha_3,k,\theta_1,\theta_2,\theta_3,\delta,\delta_1$ are suitably chosen constants to be determined such that mean squared errors (*MSEs*) of $d_i\left(i=5to9\right)$, $d_j\left(j=11to16\right)$, $d_l\left(l=18,19,20\right)$are (maybe) minimized, $a\left(\neq 0\right)$ and *b* are real numbers or the values of the parameters C_z,S_z, coefficient of skewness $\beta_1(z)$, coefficient of kurtosis $\beta_2(z)$ and $\Delta=\left(\beta_2(z)-\beta_1(z)-1\right)$ *etc*. associated with additional non-sensitive auxiliary variable z, for instance, see Upadhyaya and Singh (1999).

We propose a class of estimators of population mean of sensitive characteristic based on unmatched sample of size *u*, by following Srivastava (1980). When the population mean $\bar{Z}$ of the auxiliary variable *z* is known, we define a class of estimators for population mean of sensitive characteristic as

$$D_{u1} = D\left(\bar{h}_u,\bar{z}_u\right), \tag{2}$$

where $D\left(\bar{h}_u,\bar{z}_u\right)$ is a function of $\bar{h}_u$ and $\bar{z}_u$ such that $D\left(\bar{H},\bar{Z}\right)=\bar{H}\Rightarrow D_1\left(\bar{H},\bar{Z}\right)=1$; $D_1\left(\bar{H},\bar{Z}\right)$ being the first order partial derivative of the function $D\left(\bar{h}_u,\bar{z}_u\right)$ at the point and the function $D\left(\bar{h}_u,\bar{z}_u\right)$ also satisfies the following conditions

(i) The point $\left(\bar{h}_u,\bar{z}_u\right)$ assume values in a bounded, closed convex subset, T, of the two dimensional real space containing the point $\left(\bar{H},\bar{Z}\right)$.

(ii) The function $D\left(\bar{h}_u,\bar{z}_u\right)$ is continuous and bounded in T.

(iii) The first and second order partial derivatives of $D\left(\bar{h}_u,\bar{z}_u\right)$ exist and are continuous and bounded in T.

Assuming that the population size is sufficiently large so that finite population correction (fpc) term can be ignored, the bias and *MSE* of D_{u1} to the first degree of approximation (fda) are respectively given by

$$B(D_{u1}) = \frac{1}{2u}\left[S_z^2 D_{22}(\overline{H},\overline{Z}) + 2S_h S_z \rho_{hz} D_{12}(\overline{H},\overline{Z})\right], \tag{3}$$

$$MSE(D_{u1}) = \frac{1}{u}\left[S_h^2 + S_z^2 D_2^2(\overline{H},\overline{Z}) + 2S_h S_z \rho_{hz} D_2(\overline{H},\overline{Z})\right], \tag{4}$$

where $D_2(\overline{H},\overline{Z}), D_{12}(\overline{H},\overline{Z})$ and $D_{22}(\overline{H},\overline{Z})$ are the first and second order partial derivatives of the function $D(\overline{h}_u, \overline{z}_u)$ at the point $(\overline{H},\overline{Z})$.

Any parametric function $D(\overline{h}_u, \overline{z}_u)$ satisfying conditions (i)-(iii) can be an acceptable estimator of population mean of a sensitive variable at current move. The class of such estimators is very large.

It can be easily identified that the estimators $d_i\,(i = 1 to 20)$ are members of the suggested class of estimators D_{u1} given by (2). Thus the biases and mean squared errors of the estimators d_1 to d_{20} can be easily obtained from (3) and (4) just by putting the values of ($D_{12}(\overline{H},\overline{Z})$, $D_{22}(\overline{H},\overline{Z})$) and $D_2(\overline{H},\overline{Z})$ in (3) and (4) respectively.
The $MSE(D_{u1})$ at (4) is minimized for

$$D_2(\overline{H},\overline{Z}) = -\rho_{hz}\, S_h / S_z = -\beta_{hz}, \tag{5}$$

where $\beta_{hz} = S_{hz}/S_z^2$ is the population regression coefficient of *h* on *z,* and

$$S_{hz} = \frac{1}{N-1}\sum_{i=1}^{N}(h_i - \overline{H})(z_i - \overline{Z}).$$

Thus, the resulting minimum *MSE* of D_{u1} is given by

$$MSE_{\min}(D_{u1}) = (S_h^2/u)(1-\rho_{hz}^2). \tag{6}$$

Thus, we established the following theorem.

Theorem 1: Up to terms of order u^{-1},

$$MSE_{\min}(D_{u1}) \geq (S_h^2/u)(1-\rho_{hz}^2)$$

with equality holding if $D_2(\overline{H},\overline{Z}) = -\beta_{hz}$.

2.3.2. Class of estimators based on unmatched sample of size *u* utilizing information on $(\overline{z}_u, \overline{Z}, S_z^2)$ of non-sensitive auxiliary variable *z*

Das and Tripathi (1978) and Srivastava and Jhajji (1980, 1981) have advocated that in many survey situations of practical importance, information on population variance (σ_z^2) / mean square S_z^2 is also known along with population mean $\overline{Z}$. Thus, utilizing the knowledge

on $\left(\bar{z}_u, \bar{Z}, S_z^2\right)$ we define a class of estimators for population mean $\bar{H}$ of coded response variable at current (second) occasion (move) in two occasion successive sampling as

$$D_{u2} = F\left(\bar{h}_m, \bar{z}_u^*, s_{zu}^{*2}\right), \tag{7}$$

where $\bar{z}_u^* = \bar{z}_u / \bar{Z}$, $s_{zu}^{*2} = s_{zu}^2 / S_z^2$ and $F(.)$ is a function of $\left(\bar{h}_m, \bar{z}_u^*, s_{zu}^{*2}\right)$such that $F\left(\bar{H},1,1\right) = \bar{H}$ $\Rightarrow F_1\left(\bar{H},1,1\right) = 1, F_1\left(\bar{H},1,1\right)$ being the first order partial derivative of the function $F(.)$ at the point $\left(\bar{H},1,1\right)$.

The function $F(.)$ at (7) also satisfies certain regularity conditions like those given in Srivastava and Jhajji (1980, 1981).

The Bias and *MSE* of D_{u2} to the fda, ignoring fpc term, are respectively given by

$$B\left(D_{u2}\right) = \frac{1}{2u}\left[C_z^2 F_{22}(S) + (\lambda_{004} - 1)F_{33}(S) + 2\rho_{hz}S_h C_z F_{12}(S) + 2\lambda_{012}S_h F_{13}(S) + 2\lambda_{003}C_z F_{23}(S)\right], \tag{8}$$

$$\begin{aligned} MSE\left(D_{u2}\right) = \frac{1}{u}\Big[&S_h^2 + C_z^2 F_2^2(S) + (\lambda_{004} - 1)F_3^2(S) + 2\rho_{hz}S_h C_z F_2(S) \\ &+ 2\lambda_{012}S_h F_3(S) + 2\lambda_{003}C_z F_2(S)F_3(S)\Big] \end{aligned}, \tag{9}$$

where $\lambda_{rst} = \dfrac{\mu_{rst}}{\mu_{200}^{r/2}\ \mu_{020}^{s/2}\ \mu_{002}^{t/2}}$, $\mu_{rst} = \dfrac{1}{N}\sum_{i=1}^{N}\left(g_i - \bar{G}\right)^r\left(h_i - \bar{H}\right)^s\left(z_i - \bar{Z}\right)^t$, (r,s,t) being non-negative integers, $\left(F_2(S), F_3(S)\right)$ and $\{F_{12}(S), F_{13}(S), F_{22}(S),\ F_{23}(S), F_{33}(S)\}$ are the first and second order partial derivatives of the function $F(.)$ at the point $S = \left(\bar{H},1,1\right)$.

Differentiating (9) partially with respect to ($F_2(S)$ and $F_3(S)$) and equating them to zero we have

$$\begin{bmatrix} C_z^2 & C_z\lambda_{003} \\ C_z\lambda_{003} & (\lambda_{004} - 1) \end{bmatrix}\begin{bmatrix} F_2(S) \\ F_3(S) \end{bmatrix} = \begin{bmatrix} -\rho_{hz}S_h C_z \\ -\lambda_{012}S_h \end{bmatrix}. \tag{10}$$

After simplification of (10) we get the optimum values of $F_2(S)$ and $F_3(S)$ respectively as

$$F_2(S) = \frac{S_h\left[\lambda_{003}\lambda_{012} - (\lambda_{004} - 1)\rho_{hz}\right]}{C_z\left(\lambda_{004} - \lambda_{003}^2 - 1\right)} = F_{20}(S)\ , say \tag{11}$$

$$F_3(S) = \frac{S_h\left[\lambda_{003}\rho_{hz} - \lambda_{012}\right]}{\left(\lambda_{004} - \lambda_{003}^2 - 1\right)} = F_{30}(S), say. \tag{12}$$

Thus, the resulting minimum *MSE* of D_{u2} is given by

$$MSE_{\min}\left(D_{u2}\right) = \frac{S_h^2}{u}\left[1 - \rho_{hz}^2 - \frac{\left(\lambda_{012} - \lambda_{003}\rho_{hz}\right)^2}{\left(\lambda_{004} - \lambda_{003}^2 - 1\right)}\right]. \tag{13}$$

Now we state the following theorem.

Theorem 2: Up to the fda,

$$MSE_{\min}\left(D_{u2}\right) \geq \frac{S_h^2}{u}\left[1-\rho_{hz}^2-\frac{\left(\lambda_{012}-\lambda_{003}\rho_{hz}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}\right]$$

with equality holding if $F_2(S)=F_{20}(S)$, $F_3(S)=F_{30}(S)$.

The class of estimators D_{u2} is very large. The following estimators are given below

$$D_{u2(1)}=\bar{h}_u \exp\left\{\frac{\beta\left(1-\bar{z}_u^*\right)}{1+\left(a_1-1\right)\bar{z}_u^*}\right\}\exp\left\{\frac{\gamma\left(1-s_{zu}^{*2}\right)}{1+\left(a_2-1\right)s_{zu}^{*2}}\right\}, \qquad D_{u2(2)}=\bar{h}_u\frac{\left\{1+\alpha\left(\bar{z}_u^*-1\right)\right\}}{\left\{1+\beta\left(s_{zu}^{*2}-1\right)\right\}},$$

$$D_{u2(3)}=\bar{h}_u\left[1+\alpha\left(\bar{z}_u^*-1\right)+\beta\left(s_{zu}^{*2}-1\right)\right]^{-1}, \qquad D_{u2(4)}=\bar{h}_u\left[1+\alpha\left(\bar{z}_u^*-1\right)\right]\left(\frac{S_{zu}^2}{S_z^2}\right)^{\beta},$$

etc. are members of the proposed class of estimators D_{u2}, where $(\alpha,\beta,\gamma,a_1,a_2)$ are suitably chosen constants. The bias and *MSE* of the estimators $D_{u2(j)}; j=1 to 4$ can be easily obtained from (8) and (9) just by putting the suitable values of $F_2(S)$, $F_3(S)$, $F_{12}(S)$, $F_{13}(S)$, $F_{22}(S)$, $F_{23}(S)$ and $F_{33}(S)$.

It is to be mentioned that the estimator like $D_{u2(1)}$ has been suggested by Priyanka and Trisandhya (2019). The bias and *MSE* of $D_{u2(1)}$ can be easily obtained by putting

$$\left[F_{12}(S)=-\frac{\beta\overline{H}}{a_1},\ F_{13}(S)=-\frac{\gamma\overline{H}}{a_2},\ F_{22}(S)=\overline{H}\left(\frac{2\beta}{a_1}-\frac{2\beta}{a_1^2}+\frac{\beta^2}{a_1^2}\right),\ F_{23}(S)=-\frac{\beta\gamma\overline{H}}{a_1a_2},\right.$$

$$\left.F_{33}(S)=\overline{H}\left(\frac{2\gamma}{a_2}-\frac{2\gamma}{a_2^2}+\frac{\gamma^2}{a_2^2}\right)\right] \text{ and } \left[F_2(S)=-\frac{\beta\overline{H}}{a_1}, F_3(S)=-\frac{\gamma\overline{H}}{a_2}\right] \text{ in (8) and (9) respectively.}$$

From (6) and (13) we have

$$MSE_{\min}\left(D_{u1}\right)-MSE_{\min}\left(D_{u2}\right)=\frac{S_h^2}{u}\cdot\frac{\left(\lambda_{012}-\lambda_{003}\rho_{hz}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}\geq 0 \tag{14}$$

Thus, the class of estimators D_{u2} is more efficient than D_{u1} provided $\lambda_{012}\neq\lambda_{003}\rho_{hz}$. For this situation $\lambda_{012}=\lambda_{003}\rho_{hz}$, both the classes of estimators D_{u1} and D_{u2} are equally efficient. We should also add here that if the variables (*h, z*) have bivariate normal distribution, then there is no advantage of using the estimator D_{u2}. In such case, it is worth advisable to pick up the estimators belonging to the class of estimators D_{u1}.

2.3.3. Class of estimators based on matched sample at current (second) occasion using information on $\left(\bar{g}_m,\bar{g}_n,\bar{z}_m,\bar{z}_n,\bar{Z}\right)$

In successive sampling it is tradition to use information gathered on first occasion as auxiliary information in addition to additional non-sensitive auxiliary variable for improving the precision of estimates at current move. The estimator based on matched sample of size *m* for population mean $\overline{H}$ at current move due to Priyanka and Trisandhya (2019) are

$$t_{m1} = \bar{h}_m \exp\left\{\frac{\theta(1-\phi)}{1+(b_1-1)\phi}\right\}\exp\left\{\frac{\delta(1-\xi)}{1+(b_2-1)\xi}\right\},$$

$$t_{m2} = \bar{h}_m \exp\left\{\frac{\theta(1-\phi)}{1+(b_1-1)\phi}\right\}\exp\left\{\frac{\delta(1-\xi)}{1+(b_2-1)\xi}\right\}\exp\left\{\frac{\eta(1-v)}{1+(b_3-1)v}\right\},$$

where $\phi = g_m/g_n$, $\xi = \bar{z}_n/\bar{Z}$, $v = s_{zn}^2/S_z^2$ *and* $(\theta,\delta,\eta,b_i;i=1,2,3)$ are suitably chosen constants. It is to be noted that the estimators t_{mi} , $i=1,2$ are not utilizing the information on matched sample for additional non-sensitive auxiliary variable *z* (i.e. information on ($\bar{z}_m$, s_{zm}^2)) while information on ($\bar{z}_m$, s_{zm}^2) associated with additional non-sensitive auxiliary variable *z* can be made available easily. This led authors to propose classes of estimators utilizing information on ($\bar{z}_m$, s_{zm}^2) along with $\left(\bar{g}_m, \bar{g}_n, , \bar{z}_n s_{zn}^2, \bar{Z}, S_z^2\right)$ of non-sensitive auxiliary variable *z*.

We propose a class of estimators of population mean $\bar{H}$ at current (second) occasion, by following Srivastava (1971, 1980) as

$$J_{m1} = J\left(\bar{h}_m, \phi, \psi, \xi\right), \tag{15}$$

where *J*(.) is a function of $\left(\bar{h}_m, \phi, \psi, \xi\right)$with $\phi = g_m/g_n$ $\psi = \bar{z}_m/\bar{z}_n$, $\xi = \bar{z}_n/\bar{Z}$ such that

$$J(Q) = J\left(\bar{H},1,1,1\right) = \bar{H} \;\; \Rightarrow J_1(Q) = 1; \tag{16}$$

$J_1(Q)$ being the first order partial derivative of the function $J(.)$ at the point $Q = \left(\bar{H},1,1,1\right)$ and satisfies certain regularity conditions similar to these given in Srivastava (1971,1980). A large number of estimators may be identified as member of the class J_{m1} at (15). The following are some examples

$$J_{m1(1)} = \bar{h}_m \phi^{\alpha_1}, \psi^{\alpha_2} \xi^{\alpha_3},$$

$$J_{m1(2)} = \bar{h}_m \exp\left\{\frac{\alpha_1(1-\phi)}{(1+\phi)}\right\}\exp\left\{\frac{\alpha_2(1-\psi)}{(1+\psi)}\right\}\exp\left\{\frac{\alpha_3(1-\xi)}{(1+\xi)}\right\},$$

$$J_{m1(3)} = \left\{\bar{h}_m + \beta_{hg}(1-\phi)\right\}, \psi^{\alpha_2}\xi^{\alpha_3}, \quad J_{m1(4)} = \left\{\bar{h}_m + \beta_{hg}(1-\phi) + \beta_{hz}(1-\psi)\right\}\xi^{\alpha_3},$$

$$J_{m1(5)} = \bar{h}_m\left[1 + \alpha_1(1-\phi) + \alpha_2(1-\psi) + \alpha_3(1-\xi)\right],$$

etc., where $\alpha_i's, \eta_i's$, $(i=1,2,3)$ are suitably chosen constants.

To the fda, ignoring fpc term, the bias and *MSE* of the class of estimators J_{m1} are respectively given by

$$B(J_{m1}) = \frac{1}{2}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{C_g^2 J_{22}(Q) + C_z^2 J_{33}(Q) + 2\rho_{hg}S_h C_g J_{12}(Q) + 2\rho_{hz}S_h C_z J_{13}(Q) + 2\rho_{gz}C_g C_z J_{23}(Q)\right\}\right.$$
$$\left. + \frac{1}{n}\left\{C_z^2 J_{44}(Q) + 2\rho_{hz}S_h C_z J_{14}(Q)\right\}\right], \tag{17}$$

$$MSE(J_{m1}) = \left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{S_h^2 + C_g^2 J_2^2(Q) + C_z^2 J_3^2(Q) + 2\rho_{hg} S_h C_g J_2(Q) + 2\rho_{hz} S_h C_z J_3(Q) + 2\rho_{gz} C_g C_z J_2(Q) J_3(Q)\right\} + \frac{1}{n}\left\{S_h^2 + C_z^2 J_4^2(Q) + 2\rho_{hz} S_h C_z J_4(Q)\right\}\right], \tag{18}$$

where and $J_{jl}(Q)\ ((j,l)=1,2,3,4)$; are first and second order partial derivatives of the function $J(.)$ at the point Q.

The $MSE(J_{m1})$ at (18) is minimized for

$$J_2(Q) = \frac{S_h(\rho_{hz}\rho_{gz}-\rho_{hg})}{C_g(1-\rho_{gz}^2)} = J_{20}(Q)\ ,\ J_3(Q) = \frac{S_h(\rho_{hg}\rho_{gz}-\rho_{hz})}{C_z(1-\rho_{gz}^2)} = J_{30}(Q)\ ,\ J_4(Q) = -\frac{S_h\rho_{hz}}{C_z} = J_{40}(Q)\ ,\ \Big\}(say) \tag{19}$$

Thus, the resulting minimum *MSE* of J_{m1} is given by

$$MSE_{\min}(J_{m1}) = S_h^2\left[\frac{1}{m}(1-R_{h.gz}^2)+\frac{1}{n}(R_{h.gz}^2-\rho_{hz}^2)\right], \tag{20}$$

where $$R_{h.gz}^2 = \frac{(\rho_{hg}^2+\rho_{hz}^2-2\rho_{hz}\rho_{gz}\rho_{hg})}{(1-\rho_{gz}^2)}.$$

Now, we state the following theorem:

Theorem 3: Up to the fda,

$$MSE_{\min}(J_{m1}) \geq S_h^2\left[\frac{1}{m}(1-R_{h.gz}^2)+\frac{1}{n}(R_{h.gz}^2-\rho_{hz}^2)\right]$$
$$= S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)(1-R_{h.gz}^2)+\frac{1}{n}(1-\rho_{hz}^2)\right]$$

with equality holding if $J_i(Q) = J_{i0}(Q)\ i = 2,3\,4$.

2.3.3.1. Class of estimators based on matched sample of size *m* at current occasion using information on (ϕ,ξ)

If the information on (h_m,ϕ,ξ) is used to estimate the population mean $\overline{H}$ at current move, then following the procedure adopted by Srivastava (1971), we define a class of estimators as

$$J_{m2} = J^*(\overline{h}_m,\phi,\xi), \tag{21}$$

where $J(.)$ is a function of $(\overline{h}_m,\phi,\ \xi)$ such that $J^*(\overline{H},1,1,) = \overline{H} \Rightarrow J_1^*(H,1,1) = 1.\ J_1^*(\overline{H},1,1,\ )$ being the first and second order partial derivatives of the function about the point $(\overline{H},1,1)$; and satisfies certain regularity conditions similar to these given in Srivastava (1971).

To the fda, ignoring fpc term, the bias and *MSE* of the class of estimators J_{m2} are respectively given by

$$B(J_{m2}) = \frac{1}{2}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{C_g^2 J_{22}^*(Q^*) + 2\rho_{hg} S_h C_g J_{12}^*(Q^*)\right\} + \frac{1}{n}\left\{C_z^2 J_{33}^*(Q^*) + 2\rho_{hz} S_h C_z J_{13}^*(Q^*)\right\}\right], \tag{22}$$

$$MSE(J_{m2}) = \left[\begin{array}{l}\left(\frac{1}{m}-\frac{1}{n}\right)\left\{S_h^2 + C_g^2 J_{22}^{*2}(Q^*) + 2\rho_{hg} S_h C_g J_2^*(Q^*)\right\} \\ + \frac{1}{n}\left\{S_h^2 + C_z^2 J_3^{*2}(Q^*) + 2\rho_{hz} S_h C_z J_3^{*2}(Q^*)\right\}\end{array}\right] \tag{23}$$

where $J_i^*(Q^*)\,(i = 2, 3, 4)$ and $J_{jl}^*(Q^*)\,((j,l) = 1,2\,3,4)$; are first and second order partial derivatives of the function $J(.)$ at the point $Q^* = (\overline{H}, 1, 1)$.

The $MSE(J_{m2})$ at (23) is minimized when

$$J_2^*(Q^*) = -\frac{S_h \rho_{hg}}{C_g} = J_{20}^*(Q^*), \; say \tag{24}$$

$$J_3^*(Q^*) = -\frac{S_h \rho_{hz}}{C_z} = J_{30}^*(Q^*), \; say \tag{25}$$

Thus, the resulting minimum *MSE* of J_{m2} is given by

$$MSE_{\min}(J_{m2}) = S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{hg}^2\right) + \frac{1}{n}\left(1-\rho_{hz}^2\right)\right]. \tag{26}$$

Now, we state the following theorem.

Theorem 4: To the fda,

$$MSE_{\min}(J_{m2}) \geq S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{hg}^2\right) + \frac{1}{n}\left(1-\rho_{hz}^2\right)\right]$$

with equality holding if $J_2^*(Q^*) = J_{20}^*(Q^*)$ and $J_3^*(Q^*) = J_{30}^*(Q^*)$.

The class of estimators J_{m2} is very large. In addition to Priyanka and Trisandhya (2019) estimator t_{m1}, the following estimators

$$J_{m2(1)} = \bar{h}_m\, \phi^{\alpha_1} \xi^{\alpha_2}, \qquad J_{m2(2)} = \bar{h}_m \exp\left\{\frac{\alpha_1(1-\phi)}{(1+\phi)}\right\} \exp\left\{\frac{\alpha_2(1-\xi)}{(1+\xi)}\right\},$$

$$J_{m2(3)} = \bar{h}_m \frac{\{1+\alpha_1(\phi-1)\}}{\{1+\alpha_2(\xi-1)\}}, \qquad J_{m2(4)} = \bar{h}_m \phi^{\alpha_1} \exp\left\{\frac{\alpha_2(1-\xi)}{(1+\xi)}\right\},$$

etc. are the members of the proposed class of estimators J_{m2}, where (α_1, α_2) are suitably chosen constants. The bias and *MSE* of the estimators $t_{m1}, J_{m2(j)}; j = 1 to 4$ can be obtained easily just by putting the values of derivatives $\left(J_{12}^*(Q^*), J_{13}^*(Q^*), J_{22}^*(Q^*), J_{33}^*(Q^*), J_{23}^*(Q^*)\right)$ and $\left(J_2^*(Q^*),\; J_3^*(Q^*)\right)$ in (22) and (23) respectively.

We also note that the proposed class of estimators J_{m2} is a member of class of estimators J_{m1} at (15).

From (20) and (16) we have

$$MSE_{\min}(J_{m2}) - MSE_{\min}(J_{m1}) = S_h^2\left(\frac{1}{m}-\frac{1}{n}\right)\rho_{hg.z}^2\left(1-\rho_{hg}^2\right) \geq 0 \tag{27}$$

where $\rho_{hg.z} = \dfrac{\left(\rho_{hg} - \rho_{hz}\rho_{gz}\right)}{\sqrt{\left(1-\rho_{hz}^2\right)\left(1-\rho_{gz}^2\right)}}$ is the partial correlation coefficient between *h* and *g*.

It follows from (27) that the proposed class of estimator J_{m1} is more efficient than the estimator J_{m2}.

2.3.4. Class of estimators based on matched sample at current move using information on $\left(\bar{h}_m, \bar{g}_m, \bar{z}_m, s_{zm}^2, \bar{g}_n, \bar{z}_n s_{zn}^2, \bar{Z}, S_z^2\right)$

It is to be noted that the estimator t_{m2} due to Priyanka and Trisandhya (2019) utilizing information on $\left(\bar{Z}, S_z^2\right)$ based on matched sample of size *m* can be further generalized as

$$t_{m3} = \bar{h}_m \exp\left\{\frac{\theta_1(1-\phi)}{1+(b_1-1)\phi}\right\}\exp\left\{\frac{\theta_2(1-\psi)}{1+(b_2-1)\psi}\right\}\exp\left\{\frac{\theta_3(1-\xi)}{1+(b_3-1)\xi}\right\}\exp\left\{\frac{\theta_4(1-w)}{1+(b_4-1)w}\right\}\exp\left\{\frac{\theta_5(1-v)}{1+(b_5-1)v}\right\}, \tag{28}$$

where

$\phi = g_m/g_n, \psi = \bar{z}_m/\bar{z}_n\ \xi = \bar{z}_n/\bar{Z}$, $v = s_{zn}^2/\bar{Z}$, $w = s_{zm}^2/s_{zn}^2$, $v = s_{zn}^2/S_z^2$

and $\theta_{i's}, b_i's\,(i = 1, to 5)$

are suitably chosen constants.

Keeping the class of estimators (28) in view and adopting the same procedure as adopted by Srivastava and Jhajji (1981) we define a class of estimators of sensitive population mean $\bar{H}$ of coded response variable *h* based on the matched sample of size *m* at current move as

$$J_{m3} = L\left(\bar{h}_m, \phi, \psi, \xi, w, v\right), \tag{29}$$

where *L*(.) is a function of $\left(\bar{h}_m, \phi, \psi, \xi, w, v\right)$ such that $L(B) = \bar{H} \Rightarrow L_1(B) = 1$, $L_1(B)$ being the first order partial derivative of the function $L(.)$ at $B = \left(\bar{H}, 1, 1, 1, 1, 1\right)$and also satisfies certain regularity conditions similar to these given in Srivastava and Jhajji (1981).

To the fda, ignoring fpc term, the bias and *MSE* of the class of estimators J_{m3} are respectively given by

$$\begin{aligned} B(J_{m3}) = \frac{1}{2}\Bigg[\left(\frac{1}{m}-\frac{1}{n}\right)\Big\{&C_g^2 L_{22}(B) + C_z^2 L_{33}(B) + (\lambda_{004}-1)L_{55}(B) + 2\rho_{hg}S_h C_g L_{12}(B) \\ &+ 2\rho_{hz}S_h C_z L_{13}(B) + 2\lambda_{012}S_h L_{15}(B) + 2\rho_{gz}C_g C_z L_{23}(B) + 2\lambda_{102}C_g L_{25}(B), \\ &+ 2\lambda_{003}C_z L_{35}(B)\Big\} + \frac{1}{n}\Big\{C_z^2 L_{44}(B) + (\lambda_{004}-1)L_{66}(B) + 2\rho_{hz}S_h C_z L_{14}(B) \\ &+ 2\lambda_{012}S_h L_{16}(B) + 2\lambda_{003}C_z L_{46}(B)\Big\}\Bigg] \end{aligned} \tag{30}$$

$$MSE(J_{m3}) = \frac{1}{2}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\{S_h^2 + C_g^2 L_2^2(B) + C_z^2 L_3^2(B) + (\lambda_{004}-1)L_5^2(B) + 2\rho_{hg}S_h C_g L_2(B) + 2\rho_{hz}S_h C_z L_3(B) + 2\lambda_{012}S_h L_5(B) + 2\rho_{gz}C_g C_z L_2(B)L_3(B) + 2\lambda_{102}C_g L_2(B)L_5(B) + 2\lambda_{003}C_z L_3(B)L_5(B)\} + \frac{1}{n}\{S_h^2 + C_z^2 L_4^2(B) + (\lambda_{004}-1)L_6^2(B) + 2\rho_{hz}S_h C_z L_4(B) + 2\lambda_{012}S_h L_6(B) + 2\lambda_{003}C_z L_4(B)L_6(B)\}\right] \tag{31}$$

where $L_i(B)\,(i = 2\ 3,4,5,6)$ and $L_{ij}(B)\,((i,j) = 1,2\ 3,4,5)$; are the first and second order partial derivatives of the function $L(\bar{h}_m, \phi, \psi, \xi, w, v)$ at the point B.

The $MSE(J_{m3})$ at (31) is minimized for

$$L_2(B) = \frac{S_h \delta_1^*}{C_g \delta^*} = L_{20}(B)\ , say, \tag{32}$$

$$L_3(B) = \frac{S_h \delta_2^*}{C_z \delta^*} = L_{30}(B)\ , say\ , \tag{33}$$

$$L_4(B) = \frac{S_h\left[\lambda_{003}\lambda_{012} - (\lambda_{004}-1)\rho_{hz}\right]}{C_z\left(\lambda_{004} - \lambda_{003}^2 - 1\right)} = L_{40}(B)\ , say\ , \tag{34}$$

$$L_5(B) = \frac{S_h \delta_3^*}{\delta^*} = L_{50}(B)\ , say\ , \tag{35}$$

$$L_6(B) = \frac{S_h\left[\lambda_{003}\rho_{hz} - \lambda_{012}\right]}{\left(\lambda_{004} - \lambda_{003}^2 - 1\right)} = L_{60}(B), say\ , \tag{36}$$

where
$$\delta^* = \left[\left(\lambda_{004} - \lambda_{003}^2 - 1\right) - \rho_{gz}\left\{\rho_{gz}(\lambda_{004}-1) - \lambda_{102}\lambda_{003}\right\} + \lambda_{102}\left(\rho_{gz}\lambda_{003} - \lambda_{102}\right)\right]$$
$$\delta_1^* = \left[\rho_{hg}\left(\lambda_{004} - \lambda_{003}^2 - 1\right) - \rho_{gz}\left\{\rho_{hz}(\lambda_{004}-1) - \lambda_{012}\lambda_{003}\right\} + \lambda_{102}\left(\rho_{hz}\lambda_{003} - \lambda_{012}\right)\right]$$

$$\delta_2^* = \left[\left\{\rho_{hz}(\lambda_{004}-1) - \lambda_{012}\lambda_{003}\right\} - \rho_{hg}\left\{\rho_{gz}(\lambda_{004}-1) - \lambda_{102}\lambda_{003}\right\} + \lambda_{102}\left\{\lambda_{012}\rho_{gz} - \lambda_{102}\rho_{hz}\right\}\right]$$
$$\delta_3^* = \left[\left(\lambda_{012} - \rho_{hg}\lambda_{003}\right) - \rho_{gz}\left(\rho_{gz}\lambda_{012} - \lambda_{102}\rho_{hz}\right) + \rho_{hg}\left(\rho_{gz}\lambda_{003} - \lambda_{102}\right)\right].$$

Substitution of (32) to (36) in (31) yields the minimum *MSE* of J_{m3} as

$$MSE_{\min}(J_{m3}) = S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{1 - R_{h.gz}^2 - \frac{\rho_1^2}{Q_1\left(1-\rho_{gz}^2\right)}\right\} + \frac{1}{n}\left\{1 - \rho_{hz}^2 - \frac{\left(\lambda_{003}\rho_{hz} - \lambda_{012}\right)^2}{\left(\lambda_{004} - \lambda_{003}^2 - 1\right)}\right\}\right], \tag{37}$$
$$\rho_1 = \left[\rho_{hg}\left(\lambda_{102} - \rho_{gz}\lambda_{003}\right) + \rho_{hz}\left(\lambda_{003} - \rho_{gz}\lambda_{102}\right) + \lambda_{012}\left(1-\rho_{gz}^2\right)\right],$$
$$Q_1 = \left[(\lambda_{004}-1)\left(1-\rho_{gz}^2\right) - \lambda_{102}\left(\lambda_{102} - \rho_{gz}\lambda_{003}\right) - \lambda_{003}\left(\lambda_{003} - \rho_{gz}\lambda_{102}\right)\right].$$

Now, we state the following theorem.

Theorem 5: To the fda,

$$MSE_{\min}(J_{m3}) \geq S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{1 - R_{h.gz}^2 - \frac{\rho_1^2}{Q_1\left(1-\rho_{gz}^2\right)}\right\} + \frac{1}{n}\left\{1 - \rho_{hz}^2 - \frac{\left(\lambda_{003}\rho_{hz} - \lambda_{012}\right)^2}{\left(\lambda_{004} - \lambda_{003}^2 - 1\right)}\right\}\right],$$

with equality holding if $L_j(B)=L_{j0}(B)$, $j=2,3,4,5,6$.

The class of estimators J_{m3} at (29) is very large. The following estimators:

$$J_{m3(1)}=\bar{h}_m\,\phi^{\alpha_1}\,\psi^{\alpha_2}\xi^{\alpha_3}w^{\alpha_4}v^{\alpha_5},$$
$$J_{m3(2)}=\bar{h}_m\left[1+\alpha_1(\phi-1)+\alpha_2(\psi-1)+\alpha_3(\xi-1)+\alpha_4(w-1)+\alpha_5(v-1)\right]^{-1},$$
$$J_{m3(3)}=\bar{h}_m\left[2-\phi^{\alpha_1},\psi^{\alpha_2}\xi^{\alpha_3}w^{\alpha_4}v^{\alpha_5}\right],$$

etc. are the members of the suggested class of estimators J_{m3}, where $\alpha_i's\,(i=1to5)$ are suitably chosen constants. The bias and *MSE* of the estimators, $J_{m3(j)};\,j=1to3$ and t_{m3} at (28) can be obtained easily from (30) and (31) just by putting the values of derivatives.

Keeping the form of Priyanka and Trisandhya (2019) the estimator t_{m2} and motivated by Srivastava and Jhajji (1981) we define a subclass of estimators J_{m3} of the class of estimators J_{m2} for the population mean $\overline{H}$ of the coded response at current move as

$$J_{m4}=L^*\left(\bar{h}_m,\phi,\xi,\,v\right), \tag{38}$$

where $L^*(.)$is a function of $\left(\bar{h}_m,\phi,\xi,\,v\right)$ such that $L^*\left(B^*\right)=\overline{H},\ \Rightarrow L_1^*\left(B^*\right)=1,\ L_1^*\left(B^*\right)$ being the first order partial derivative of the function $L^*(.)$ at the point $B^*=\left(\overline{H},1,1,1\right)$ and also satisfies
certain regularity conditions similar to these given in Srivastava and Jhajji (1981).

To the fda, ignoring fpc term, the bias and *MSE* of the class of estimators J_{m4} are respectively given by

$$B(J_{m4})=\frac{1}{2}\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{C_g^2L_{22}^*\left(B^*\right)+2\rho_{hg}S_hC_gL_{12}^*\left(B^*\right)\right\}+\frac{1}{n}\left\{C_z^2L_{33}^*\left(B^*\right)+(\lambda_{004}-1)L_{44}^*\left(B^*\right)\right.\right.$$
$$\left.\left.+2\lambda_{003}C_zL_{34}^*\left(B^*\right)+2\rho_{hz}S_hC_zL_{13}^*\left(B^*\right)+2\lambda_{012}S_hL_{14}^*\left(B^*\right)\right\}\right], \tag{39}$$

$$MSE(J_{m4})=\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left\{S_h^2+C_g^2L_2^{*2}\left(B^*\right)+2\rho_{hg}S_hC_gL_2^*\left(B^*\right)\right\}+\frac{1}{n}\left\{S_h^2+C_z^2L_3^{*2}\left(B^*\right)\right.\right.$$
$$\left.\left.+(\lambda_{004}-1)L_4^{*2}\left(B^*\right)+2\lambda_{003}C_zL_3^*\left(B^*\right)L_4^*\left(B^*\right)+2\rho_{hz}S_hC_zL_3^*\left(B^*\right)+2\lambda_{012}S_hL_4^*\left(B^*\right)\right\}\right], \tag{40}$$

where $L_2^*\left(B^*\right),L_3^*\left(B^*\right),L_4^*\left(B^*\right)\ and\ L_{ij}^*\left(B^*\right),((i,j)=1,2,3,4)$ are the first and second order partial derivatives of the function $L^*\left(\bar{h}_m,\phi,\,\xi,v\right)$ at the point $B^*=\left(\overline{H},1,1,1\right)$.

The $MSE(J_{m4})$ at (40) is minimized for

$$\left.\begin{aligned} L_2^*(B^*) &= -\rho_{hg}\frac{S_h}{C_g} = L_{20}^*(B^*), say \\ L_3^*(B^*) &= \frac{S_h\left[(\lambda_{004}-1)\rho_{hz}-\lambda_{003}\lambda_{012}\right]}{C_z\left(\lambda_{004}-\lambda_{003}^2-1\right)} = L_{30}^*(B^*), say \\ L_4^*(B^*) &= \frac{S_h\left[\lambda_{012}-\lambda_{003}\rho_{hz}\right]}{\left(\lambda_{004}-\lambda_{003}^2-1\right)} = L_{40}^*(B^*), say \end{aligned}\right\}. \quad (41)$$

Thus, the resulting minimum *MSE* of J_{m4} is given by

$$MSE_{\min}(J_{m4}) = S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{gz}^2\right)+\frac{1}{n}\left\{1-\rho_{hz}^2-\frac{\left(\lambda_{003}\rho_{hz}-\lambda_{012}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}\right\}\right]. \quad (42)$$

Thus, we arrived at the following theorem.

Theorem 6: To the fda,

$$MSE_{\min}(J_{m4}) \geq S_h^2\left[\left(\frac{1}{m}-\frac{1}{n}\right)\left(1-\rho_{gz}^2\right)+\frac{1}{n}\left\{1-\rho_{hz}^2-\frac{\left(\lambda_{003}\rho_{hz}-\lambda_{012}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}\right\}\right],$$

with equality holding if $L_j^*(B^*) = L_{j0}^*(B^*)$, $j=2,3,4$; where $L_{j0}^*(B^*)$ is given by (41), $j=2,3,4$.

The class of estimators J_{m3} is very large. The following estimators

$$\begin{aligned} J_{m4(1)} &= \bar{h}_m \phi^{\alpha_1}\xi^{\alpha_2}v^{\alpha_3}, \\ J_{m4(2)} &= \bar{h}_m\left[1+\alpha_1(\phi-1)+\alpha_2(\xi-1)+\alpha_3(v-1)\right]^{-1}, \\ J_{m4(3)} &= \bar{h}_m\left[2-\phi^{\alpha_1}\xi^{\alpha_2}v^{\alpha_3}\right], \\ J_{m4(4)} &= \left[\bar{h}_m+\alpha_1(1-\phi)+\alpha_3(1-\xi)+\alpha_3(1-v)\right], \end{aligned}$$

etc. are the members of the suggested class of estimators J_{m4}, where $\alpha_i's\,(i=1to3)$ are suitably chosen constants. The bias and *MSE* of the estimators can easily be obtained from (29) and (40) just by putting the values of derivatives.

From (37) and (42) we have

$$MSE_{\min}(J_{m4}) - MSE_{\min}(J_{m3}) = \left(\frac{1}{m}-\frac{1}{n}\right)\frac{S_h^2}{\left(1-\rho_{hg}^2\right)}\left[\left(\rho_{hg}\rho_{gz-}\rho_{hz}\right)^2+\frac{\rho_1^2}{Q_1}\right] \geq 0. \quad (43)$$

It follows from (43) that the proposed class of estimator J_{m3} is more efficient than the estimator J_{m4}, and hence better than the Priyanka and Trisandhya (2019) -type estimator t_{m2}.

3. Combined Classes of Estimators

Taking the convex linear combination of class of estimators $D_{ui}\,(i=1,2)$ and $J_{mj}\,(j=1to4)$based sample of size *u* and *m* respectively, the final estimator for the population

mean $\overline{H}$ of coded response at current occasion in two occasion successive sampling is defined by

$$T_{ij} = \Phi_{ij} D_{ui} + (1 - \Phi_{ij}) J_{mj}, \tag{44}$$

where $D_{u1}, D_{u2}, J_{m1}, J_{m2}, J_{m3}$, and J_{m4} are respectively defined in (2), (7), (15), (21), (29) and (38) and $\Phi_{ij} \in [0,1]$ is a scalar quantity to be chosen suitably.

Theorem 7: Bias of the combined class of estimators T_{ij} to the fda, ignoring fpc term, is obtained as

$$B(T_{ij}) = \Phi_{ij} B(D_{ui}) + (1 - \Phi_{ij}) B(J_{mj}), \tag{45}$$

where $i = 1,2$ and $j = 1 to 4$.

Proof is simple so omitted.

Putting the values of $B(D_{u1}), B(D_{u2}), B(j_{m1}), B(j_{m2}), B(j_{m3})$, and $B(j_{m4})$ as respectively defined in (3), (8), (17), (22), (30) and (39) in the above equation, we get the expression for the bias of the class of estimators T_{ij} in (45).

Theorem 8: The mean squared error of the class of estimators T_{ij} is given by

$$MSE(T_{ij}) = \Phi_{ij}^2 MSE_{\min}(D_{ui}) + (1 - \Phi_{ij})^2 MSE_{\min}(J_{mj}), \tag{46}$$

where $\Phi_{ij}'s \ \{i = 1,2 \ and \ j = 1 to 4\}$ are constants to be determined such that mean squared errors of $T_{ij}'s$ are minimum.

Proof: The mean squared error of the class of estimators T_{ij} is derived as

$$\begin{aligned} MSE(T_{ij}) &= E(T_{ij} - \overline{H})^2 = E[\Phi_{ij}(D_{ui} - \overline{H}) + (1 - \Phi_{ij})(J_{mj} - \overline{H})]^2, \\ &= \Phi_{ij}^2 E(D_{ui} - \overline{H})^2 + (1 - \Phi_{ij})^2 E(J_{mj} - \overline{H})^2 + 2\Phi_{ij}(1 - \Phi_{ij}) E\{(D_{ui} - \overline{H})(J_{mj} - \overline{H})\}, \\ &= [\Phi_{ij}^2 MSE(D_{ui}) + (1 - \Phi_{ij})^2 MSE(J_{mj}) + 2\Phi_{ij}(1 - \Phi_{ij}) Cov(D_{ui}, J_{mj})]. \end{aligned} \tag{47}$$

The minimum *MSE*'*s* of D_{ui} $(i = 1,2)$ and J_{mj} $(j = 1 to 4)$ are given by (6), (13), (20), (26), (37) and (42) respectively and as the two sets of estimators $\{D_{u1}, D_{u2}\}$ and $\{J_{m1}, J_{m2}, J_{m3}, J_{m4}\}$ are based on two non-overlapping samples of sizes *u* and *m,* respectively, therefore $Cov(D_{ui}, J_{mj}) = 0$. Thus inserting $MSE_{\min}(D_{ui})$ and $MSE_{\min}(J_{mj})$ in place of $MSE(D_{ui})$ and $MSE(J_{mj})$ respectively and $Cov(D_{ui}, J_{mj}) = 0$ in (47), we get the *MSE* of T_{ij} as in (46).

3.1. Minimum *MSE* of the suggested combined class of estimators T_{ij}

Differentiating (46) with respect to Φ_{ij} and equating them to zero, we get the optimum value of Φ_{ij} as

$$\Phi_{ij(opt)} = \frac{MSE_{\min}\left(J_{mj}\right)}{MSE_{\min}\left(D_{ui}\right)+MSE_{\min}\left(J_{mj}\right)},\ \left(i=1,2\ ;\ j=1to4\right). \tag{48}$$

Inserting the value of $\Phi_{ij(opt)}$ from equation (48) in (46), we get the minimum *MSE* of classes of estimators T_{ij} as

$$MSE_{\min}\left(T_{ij}\right) = \frac{MSE_{\min}\left(D_{ui}\right)* MSE_{\min}\left(J_{mj}\right)}{MSE_{\min}\left(D_{ui}\right)+MSE_{\min}\left(J_{mj}\right)},\left(i=1,2\ ;\ j=1to4\right). \tag{49}$$

Putting the value of $MSE_{\min}\left(D_{ui}\right)$ and $MSE_{\min}\left(J_{mj}\right)$ from (6), (13), (20), (26), (37) and (42) respectively in (49), the simplified values of $MSE_{\min}\left(T_{ij}\right)$ are obtained as

$$MSE_{\min}\left(T_{11}\right) = \frac{S_h^2}{n}\frac{\left(1-\rho_{hg}^2\right)\left(1-\mu_{11}\rho_{hg.z}^2\right)}{\left(1-\mu_{11}^2\rho_{hg.z}^2\right)}, \tag{50}$$

$$MSE_{\min}\left(T_{12}\right) = \frac{S_h^2}{n}\frac{\left(1-\rho_{hg}^2\right)\left[\left(1-\rho_{hg}^2\right)-\mu_{12}\left(\rho_{hg}^2-\rho_{hz}^2\right)\right]}{\left[\left(1-\rho_{hg}^2\right)-\mu_{12}^2\left(\rho_{hg}^2-\rho_{hz}^2\right)\right]}, \tag{51}$$

$$MSE_{\min}\left(T_{13}\right) = \frac{S_h^2}{n}\frac{A_{11}\left(B_{13}-\mu_{13}C_{13}\right)}{\left[A_{13}-\mu_{13}\left(A_{11}-B_{13}\right)-\mu_{13}^2C_{13}\right]}, \tag{52}$$

$$MSE_{\min}\left(T_{14}\right) = \frac{S_h^2}{n}\frac{A_{11}\left(B_{13}+\mu_{14}C_{14}\right)}{\left[A_{11}-\mu_{14}C_{13}+\mu_{14}^2C_{14}\right]}, \tag{53}$$

$$MSE_{\min}\left(T_{21}\right) = \frac{S_h^2}{n}\frac{A_{11}B_{13}\left(1-\mu_{21}\rho_{hg.z}^2\right)}{\left[B_{13}+\mu_{21}C_{13}-\mu_{21}^2\rho_{hg.z}^2A_{11}\right]}, \tag{54}$$

$$MSE_{\min}\left(T_{22}\right) = \frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[A_{11}-\mu_{22}\left(\rho_{hg}^2-\rho_{hz}^2\right)\right]}{\left[A_{11}-C_{13}+\mu_{22}C_{13}-\mu_{22}^2\left(\rho_{hg}^2-\rho_{hz}^2\right)\right]}, \tag{55}$$

$$MSE_{\min}\left(T_{23}\right) = \frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[A_{11}-\mu_{23}C_{23}\right]}{\left[\left(A_{11}-C_{13}\right)-\mu_{22}^2C_{23}\right]}, \tag{56}$$

$$MSE_{\min}\left(T_{24}\right) = \frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[\left(A_{11}-C_{13}\right)-\mu_{24}\left(\rho_{hg}^2-\rho_{hz}^2-C_{13}\right)\right]}{\left[\left(A_{11}-C_{13}\right)-\mu_{24}^2\left(\rho_{hg}^2-\rho_{hz}^2-C_{13}\right)\right]}, \tag{57}$$

where $A_{11}=\left(1-\rho_{hg}^2\right)$, $\rho_{hg.z}=\dfrac{\left(\rho_{hg}-\rho_{gz}\rho_{hz}\right)}{\sqrt{\left(1-\rho_{hg}^2\right)}\sqrt{\left(1-\rho_{gz}^2\right)}}$, $A_{13}=\left\{1-R_{h.gz}^2-\dfrac{\rho_1^2}{Q_1\left(1-\rho_{gz}^2\right)}\right\}$,

$$B_{13}=\left\{1-\rho_{hz}^2-\frac{\left(\lambda_{003}\rho_{hz}-\lambda_{012}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}\right\}=\left(A_{11}-C_{13}\right),$$

$$C_{13}=\left(A_{11}-B_{13}\right)=\frac{\left(\lambda_{003}\rho_{hz}-\lambda_{012}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)},$$

$$C_{14}=\left[\frac{\left(\lambda_{003}\rho_{hz}-\lambda_{012}\right)^2}{\left(\lambda_{004}-\lambda_{003}^2-1\right)}+d_{11}\right],\ d_{11}=\left(\rho_{hz}^2-\rho_{hg}^2\right),$$

$$C_{23}=\left\{\rho_{h.g.z}^2\left(1-\rho_{hz}^2\right)+\frac{\rho_1^2}{Q_1\left(1-\rho_{gz}^2\right)}-C_{13}\right\}.$$

3.2. Optimum rotation rate

It is observed from (50) to (57) that $MSE_{\min}\left(T_{ij}\right),\ (i=1,2;\ j=1,2,3,4)$ is the function of μ_{ij} which is rotation rate or the fraction of sample to be drawn afresh at current occasion. As less the sample need to be selected afresh, less is the total cost of the survey so to estimate population mean with maximum precision and minimum cost $MSE_{\min}\left(T_{ij}\right)$ at (50)-(57) have been minimized with respect to μ_{ij}. The optimum values μ_{ij} have been derived as

$$\hat{\mu}_{11}=\frac{1}{\left[1+\sqrt{\left(1-\rho_{h.g.z}^2\right)}\right]}, \tag{58}$$

$$\hat{\mu}_{12}=\frac{\sqrt{\left(1-\rho_{h.z}^2\right)}}{\left[\sqrt{\left(1-\rho_{h.z}^2\right)}+\sqrt{\left(1-\rho_{h.g}^2\right)}\right]}, \tag{59}$$

$$\hat{\mu}_{13}=\min\left[\frac{B_{13}-\sqrt{\left(B_{13}-C_{13}C_{23}\right)}}{C_{13}},\frac{B_{13}+\sqrt{\left(B_{13}-C_{13}C_{23}\right)}}{C_{13}}\right]\in(0,1), \tag{60}$$

$$\hat{\mu}_{14}=\min\left[\frac{-B_{13}-\sqrt{\left(B_{13}\left(B_{13}-C_{13}\right)-A_{11}C_{14}\right)}}{C_{14}},\frac{-B_{13}+\sqrt{\left(B_{13}\left(B_{13}-C_{13}\right)-A_{11}C_{14}\right)}}{C_{14}}\right]\in(0,1), \tag{61}$$

$$\hat{\mu}_{21}=\min\left[\frac{A_{11}-\sqrt{A_{11}B_{13}\left(1-\rho_{h.g.z}^2\right)}}{\rho_{h.g.z}^2},\frac{A_{11}+\sqrt{A_{11}B_{13}\left(1-\rho_{h.g.z}^2\right)}}{\rho_{h.g.z}^2}\right]\in(0,1), \tag{62}$$

$$\hat{\mu}_{22}=\min\left[\frac{A_{11}-\sqrt{\left(A_{11}C_{13}-d_{11}B_{13}\right)}}{d_{11}},\frac{A_{11}+\sqrt{\left(A_{11}C_{13}-d_{11}B_{13}\right)}}{d_{11}}\right]\in(0,1), \tag{63}$$

$$\hat{\mu}_{23}=\min\left[\frac{A_{11}-\sqrt{\left(A_{11}^2-B_{13}C_{23}\right)}}{C_{23}},\frac{A_{11}+\sqrt{\left(A_{11}^2-B_{13}C_{23}\right)}}{C_{23}}\right]\in(0,1), \tag{64}$$

$$\hat{\mu}_{24}=\min\left[\frac{B_{13}-\sqrt{B_{13}\left(A_{11}-d_{11}\right)}}{\left(d_{11}-C_{13}\right)},\frac{B_{13}+\sqrt{B_{13}\left(A_{11}-d_{11}\right)}}{\left(d_{11}-C_{13}\right)}\right]\in(0,1), \tag{65}$$

Inserting $\hat{\mu}_{11},\hat{\mu}_{12}$ admissible value of $\hat{\mu}_{13},\hat{\mu}_{14},\hat{\mu}_{21},\hat{\mu}_{22},\hat{\mu}_{23}\ and\ \hat{\mu}_{24}$ from (58)-(65) respectively in (50) - (57) we get the optimum values of $MSE_{\min}\left(T_{ij}\right),\ (i=1,2;\ j=1,2,3,4)$ as

$$MSE_{\min}\left(T_{11}\right)_{opt}=\frac{S_h^2\left(1-\rho_{hz}^2\right)\left(1+\sqrt{\left(1-\rho_{hg.z}^2\right)}\right)}{2n}, \tag{66}$$

$$MSE_{\min}\left(T_{12}\right)_{opt}=\frac{S_h^2\sqrt{\left(1-\rho_{hz}^2\right)}\left[\sqrt{\left(1-\rho_{hz}^2\right)}+\sqrt{\left(1-\rho_{hg}^2\right)}\right]}{2n}, \tag{67}$$

$$MSE_{\min}\left(T_{13}\right)_{opt}=\frac{S_h^2}{n}\frac{A_{11}\left(B_{13}-\hat{\mu}_{13}C_{13}\right)}{\left[A_{13}-\hat{\mu}_{13}\left(A_{11}-B_{13}\right)-\hat{\mu}_{13}^2C_{13}\right]}, \tag{68}$$

$$MSE_{\min}\left(T_{14}\right)_{opt}=\frac{S_h^2}{n}\frac{A_{11}\left(B_{13}+\hat{\mu}_{14}C_{14}\right)}{\left[A_{11}-\hat{\mu}_{14}C_{13}+\hat{\mu}_{14}^2C_{14}\right]}, \tag{69}$$

$$MSE_{\min}\left(T_{21}\right)_{opt}=\frac{S_h^2}{n}\frac{A_{11}B_{13}\left(1-\hat{\mu}_{21}\rho_{hg.z}^2\right)}{\left[B_{13}+\hat{\mu}_{21}C_{13}-\hat{\mu}_{21}^2\rho_{hg.z}^2A_{11}\right]}, \tag{70}$$

$$MSE_{\min}\left(T_{22}\right)_{opt}=\frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[A_{11}-\hat{\mu}_{22}d_{11}\right]}{\left[\left(A_{11}-C_{13}\right)+\hat{\mu}_{22}C_{13}-\hat{\mu}_{22}^2d_{11}\right]}, \tag{71}$$

$$MSE_{\min}\left(T_{23}\right)_{opt}=\frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[A_{11}-\hat{\mu}_{23}C_{23}\right]}{\left[\left(A_{11}-C_{13}\right)-\hat{\mu}_{22}^2C_{23}\right]}, \tag{72}$$

$$MSE_{\min}\left(T_{24}\right)_{opt}=\frac{S_h^2}{n}\frac{\left(A_{11}-C_{13}\right)\left[\left(A_{11}-C_{13}\right)-\hat{\mu}_{24}\left(d_{11}-C_{13}\right)\right]}{\left[\left(A_{11}-C_{13}\right)-\hat{\mu}_{24}^2\left(d_{11}-C_{13}\right)\right]}. \tag{73}$$

4. Performances of the Suggested Classes of Estimators

For examining the relevance and utility of the information used on non-sensitive auxiliary variable with the proposed classes of estimators, we have considered a class of estimators where no additional non-sensitive auxiliary information is used, known as modified general class of successive sampling estimators.

4.1. Modified general class of estimators based on matched sample of size *m*

Following the procedure adopted by Srivastava (1971,1980) we consider the following class of estimators of the population mean $\overline{H}$ of coded response variable on the current (second) occasion as

$$t_G=G\left(\overline{h}_m,\phi\right), \tag{74}$$

where $G(.)$ is a function of $\left(\overline{h}_m,\phi=\overline{g}_m/\overline{g}_n\right)$ such that

$$G\left(\overline{H},1\right)=\overline{H}\Rightarrow G_1\left(\overline{H},1\right)=1, \tag{75}$$

$G_1\left(\overline{H},1\right)$ being the first order partial derivative of the function $G\left(\overline{h}_m,\phi\right)$ at the point $\left(\overline{H},1\right)$and satisfies certain regularity conditions similar to these given in Srivastava (1971,1980).

To the fda, ignoring fpc term, the bias and *MSE* of t_G are respectively given by

$$B\left(t_G\right)=\frac{1}{2}\left(\frac{1}{m}-\frac{1}{n}\right)\left[C_g^2G_{22}\left(\overline{H},1\right)+2\rho_{hg}S_hC_gG_{12}\left(\overline{H},1\right)\right], \tag{76}$$

$$MSE\left(t_G\right)=\left[\frac{1}{m}S_h^2+\left(\frac{1}{m}-\frac{1}{n}\right)\left\{C_g^2G_2^2\left(\overline{H},1\right)+2\rho_{hg}S_hC_gG_2\left(\overline{H},1\right)\right\}\right], \tag{77}$$

where $G_2(\overline{H},1)=\left.\frac{\partial G(.)}{\partial \phi}\right|_{(\overline{H},1)}$, $G_{12}(\overline{H},1)=\left.\frac{\partial^2 G(.)}{\partial \phi \partial \overline{h}_m}\right|_{(\overline{H},1)}$ and $G_{22}(\overline{H},1)=\left.\frac{\partial^2 G(.)}{\partial^2 \phi}\right|_{(\overline{H},1)}$.

The *MSE* of t_G at (77) is minimum when

$$G_2(\overline{H},1)=-\rho_{hg}(S_h/C_g). \tag{78}$$

Substitution (78) in (77) yields the minimum *MSE* of the class of estimators t_G as

$$MSE_{\min}(t_G)=S_h^2\left[\frac{1}{m}(1-\rho_{hg}^2)+\frac{1}{n}\rho_{hg}^2\right] \tag{79}$$

which is equal to the minimum *MSE* of the difference estimator

$$t_{dm}=\overline{h}_m+\beta_{hg}(\overline{g}_n-\overline{g}_m), \tag{80}$$

where β_{hg} is the population regression coefficient of h on g.

We note that the class of estimators t_G at (74) is very vast. The following estimators (including t_{dm})

$$t_{G1}=\overline{h}_m\phi^{\alpha_1},\quad t_{G2}=\overline{h}_m(2-\phi^{\alpha_1}),\quad t_{G3}=\overline{h}_m\exp\left\{\frac{\alpha_1(\overline{g}_m-\overline{g}_n)}{(\overline{g}_m+\overline{g}_n)}\right\},$$

etc. are the members of the class of estimators t_G at (74). The bias and *MSE* of the estimators $t_{Gj}\ (j=1to3)$ and t_{dm} can be easily obtained from (76) and (77) just by putting the suitable values of derivatives $G_2(\overline{H},1)$, $G_{12}(\overline{H},1)$ and $G_{22}(\overline{H},1)$.

Now we state the following theorem.

Theorem 9: Up to the first order of approximation,

$$MSE_{\min}(t_G)\geq S_h^2\left[\frac{1}{m}(1-\rho_{hg}^2)+\frac{1}{n}\rho_{hg}^2\right]$$

with equality holding if $G_2(\overline{H},1)=-\rho_{hg}(S_h/C_g)$.

4.2. Combined class of estimators

We consider the following combined classes of estimators for population mean response $\overline{H}$ of coded response variable at current (second) move at

$$T_C=\Phi\overline{h}_u+(1-\Phi)t_G, \tag{81}$$

where $\Phi\in[0,1]$ is unknown constant.

We note that the class of estimators

$$T_J=\Phi_{(1)}\overline{h}_u+(1-\Phi_{(1)})[\overline{h}_m+k_1(\overline{g}_n-\overline{g}_m)], \tag{82}$$

due to Priyanka and Trisandhya (2019) is a member of the class of estimators T_C at (81), $\Phi_{(1)} \in [0,1]$ is unknown constant and k_1 is a suitably chosen scalar.

The *MSE* of T_C is given by

$$MSE(T_C) = \Phi^2 MSE(\bar{h}_u) + (1-\Phi)^2 MSE(t_G), \tag{83}$$

as the term $Cov(\bar{h}_u, t_G) = 0$.

Replacing $MSE(t_G)$ by its $MSE_{\min}(t_G)$ in (83) we have

$$MSE(T_C) = \Phi^2 [MSE(\bar{h}_u) + MSE_{\min}(t_G)] - 2\Phi MSE_{\min}(t_G) + MSE_{\min}(t_G), \tag{84}$$

which is minimum when

$$\Phi = \frac{MSE_{\min}(t_G)}{MSE(\bar{h}_u) + MSE_{\min}(t_G)}, \tag{85}$$

and thus the resulting minimum *MSE* of class of estimators T_C is given by

$$MSE_{\min}(T_C) = \frac{S_h^2}{n} \frac{(1-\mu_J \rho_{hg}^2)}{(1-\mu_J^2 \rho_{hg}^2)} . \tag{86}$$

Expression (86) is optimized for

$$\hat{\mu}_J = \frac{1}{\{1+\sqrt{(1-\rho_{hg}^2)}\}} . \tag{87}$$

Thus the optimum value of $MSE_{\min}(T_C)$ is

$$MSE_{\min}(T_C)_{opt} = \frac{S_h^2 \{1+\sqrt{(1-\rho_{hg}^2)}\}}{2n} . \tag{88}$$

4.3. Theoretical comparison of the estimators $T_{11} = \{\Phi_{11} D_{u1} + (1-\Phi_{11}) J_{m1}\}$ and $T_{12} = \{\Phi_{12} D_{u1} + (1-\Phi_{12}) J_{m2}\}$ with the estimators $T_C = \{\Phi \bar{h}_u + (1-\Phi) t_G\}$, [or with the estimators $T_J = \Phi_{(1)} \bar{h}_u + (1-\Phi_{(1)}) [\bar{h}_m + k_1 (\bar{g}_n - \bar{g}_m)]$]

From (66), (67) and (68) we have

$$MSE_{\min}(T_C)_{opt} - MSE_{\min}(T_{11})_{opt} = \frac{S_h^2}{2n} \left(\rho_{hg}^2 + \frac{N_1}{D_1} \right) > 0, \tag{89}$$

$$MSE_{\min}(T_C)_{opt} - MSE_{\min}(T_{12})_{opt} = \frac{S_h^2}{2n} \left[\rho_{hz}^2 + \sqrt{(1-\rho_{hg}^2)} \{1-\sqrt{(1-\rho_{hz}^2)}\} \right] > 0, \tag{90}$$

$$MSE_{\min}(T_{12})_{opt} - MSE_{\min}(T_{11})_{opt} = \frac{S_h^2}{2n} \frac{\sqrt{(1-\rho_{hz}^2)} (\rho_{hz} - \rho_{hg}\rho_{gz})}{(1-\rho_{gz}^2) D_2} > 0, \tag{91}$$

where $N_1 = [\rho_{hz}^2 (1-R_{h.gz}^2) + (\rho_{hz} - \rho_{hg}\rho_{gz})^2]$,

$D_1 = (1-\rho_{gz}^2) [\sqrt{(1-\rho_{hz}^2)} + (1-\rho_{hz}^2) \{\sqrt{(1-\rho_{hg.z}^2)}\}]$, and $D_2 = [\sqrt{(1-\rho_{hg}^2)} + \sqrt{(1-\rho_{hz}^2)} \sqrt{(1-\rho_{hg.z}^2)}]$.

From (89), (90) and (91) we have the inequality

$$MSE_{\min}\left(T_{11}\right)_{opt} < MSE_{\min}\left(T_{12}\right)_{opt} < MSE_{\min}\left(T_{C}\right)_{opt}. \tag{92}$$

It follows from (92) that the proposed estimator T_{11} is more efficient than the estimators T_{12} and T_C. Theoretical Comparison among the estimators $T_{13}, T_{14}, T_{21}, T_{22}, T_{23}$ and T_{24} are tedious, therefore we have not made the comparison among these estimators.

5. Estimators of Sensitive Population Mean at Current (Second) Move under Model M_G

The population mean $\overline{H}$ of the coded response variable *h* at current occasion in (1) is replaced by its estimators $T_{ij}, T_C\ (i=1,2;\ j=1,2,3,4)$ given in (44) and (81) respectively, to derive the estimators $\hat{\bar{Y}}_{ij}$ and $\hat{\bar{Y}}_C$ for sensitive population mean which are given below

$$\hat{\bar{Y}}_{11} = T_{11} = \left\{\Phi_{11}D_{u1} + \left(1-\Phi_{11}\right)J_{m1}\right\}, \qquad \hat{\bar{Y}}_{12} = T_{12} = \left\{\Phi_{12}D_{u1} + \left(1-\Phi_{12}\right)J_{m2}\right\},$$
$$\hat{\bar{Y}}_{13} = T_{13} = \left\{\Phi_{13}D_{u1} + \left(1-\Phi_{13}\right)J_{m3}\right\}, \qquad \hat{\bar{Y}}_{14} = T_{14} = \left\{\Phi_{14}D_{u1} + \left(1-\Phi_{14}\right)J_{m4}\right\},$$
$$\hat{\bar{Y}}_{21} = T_{21} = \left\{\Phi_{21}D_{u2} + \left(1-\Phi_{21}\right)J_{m1}\right\}, \qquad \hat{\bar{Y}}_{22} = T_{22} = \left\{\Phi_{22}D_{u2} + \left(1-\Phi_{22}\right)J_{m2}\right\},$$
$$\hat{\bar{Y}}_{23} = T_{23} = \left\{\Phi_{23}D_{u2} + \left(1-\Phi_{23}\right)J_{m3}\right\}, \qquad \hat{\bar{Y}}_{24} = T_{24} = \left\{\Phi_{24}D_{u2} + \left(1-\Phi_{24}\right)J_{m4}\right\},$$
$$\hat{\bar{Y}}_C = T_C = \left\{\Phi\bar{h}_u + \left(1-\Phi\right)t_G\right\}.$$

5.1. Numerical illustration

To have tangible idea about the performance of the suggested estimators $\hat{\bar{Y}}_{11}$ and $\hat{\bar{Y}}_{12}$ (for the sake of convenience we have considered only two estimators $\hat{\bar{Y}}_{11}$ and $\hat{\bar{Y}}_{12}$ for purpose of comparison), we have considered artificial parametric values

$$\sigma_x^2 = 6,\ \sigma_y^2 = 2,\ \sigma_z^2 = 2,\ \rho_{yx} = 0.6820,\ \rho_{yz} = 0.7520,\ \rho_{xz} = 0.650,\ \overline{Y} = 5.00,\ \overline{X} = 4.50, \overline{Z} = 5.50$$

Here we suppose that $W \sim F(10,5)$ for which $\overline{W} = 1.6667, \sigma_w = 2.6874$.

The optimum values of fraction of sample to be drawn afresh at current (second) occasion (move) and percent relative efficiencies (*PRE*s) have been computed by using the following formulae

$$\hat{\mu}_{11} = \left[\frac{1}{1+\sqrt{\left(1-\rho_{h.g.z}^2\right)}}\right], \tag{93}$$

$$\hat{\mu}_{12} = \left[\frac{\sqrt{\left(1-\rho_{h.z}^2\right)}}{\sqrt{\left(1-\rho_{h.z}^2\right)}+\sqrt{\left(1-\rho_{h.g}^2\right)}}\right], \tag{94}$$

$$PRE\left(\hat{\bar{Y}}_{11}, \hat{\bar{Y}}_C\right) = \frac{\left[1+\sqrt{\left(1-\rho_{h.g}^2\right)}\right]}{\left(1-\rho_{h.z}^2\right)\left[1+\sqrt{\left(1-\rho_{h.g.z}^2\right)}\right]} * 100, \tag{95}$$

$$PRE\left(\hat{\bar{Y}}_{12},\hat{\bar{Y}}_{C}\right)=\frac{\left[1+\sqrt{\left(1-\rho_{h.g}^{2}\right)}\right]}{\sqrt{\left(1-\rho_{h.z}^{2}\right)}\left[\sqrt{\left(1-\rho_{h.z}^{2}\right)}+\sqrt{\left(1-\rho_{h.g}^{2}\right)}\right]}*100, \tag{96}$$

Findings are given in Table 5.1 and 5.2.

Table 5.1: Optimum Values of $\hat{\mu}_{11}$ (in bracket) and *PRE* of $\hat{\bar{Y}}_{11}$ with respect to $\hat{\bar{Y}}_{C}$.

α / β	**0.01**	**0.05**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**	**1.50**	**2.00**
0.01	**207.13** (0.5201)	**206.63** (0.52)	**206.02** (0.5199)	**203.64** (0.5194)	**201.36** (0.5189)	**199.18** (0.5185)	**197.1** (0.518)	**196.09** (0.5178)	**191.35** (0.5169)	**187.06** (0.516)
0.05	**206.63** (0.52)	**204.22** (0.5195)	**201.36** (0.5189)	**191.35** (0.5169)	**183.16** (0.5152)	**176.33** (0.5139)	**170.56** (0.5127)	**167.99** (0.5122)	**157.56** (0.5101)	**149.94** (0.5086)
0.1	**206.02** (0.5199)	**201.36** (0.5189)	**196.09** (0.5178)	**179.6** (0.5145)	**167.99** (0.5122)	**159.37** (0.5105)	**152.72** (0.5092)	**149.94** (0.5086)	**139.56** (0.5065)	**132.82** (0.5052)
0.3	**203.64** (0.5194)	**191.35** (0.5169)	**179.6** (0.5145)	**152.72** (0.5092)	**139.56** (0.5065)	**131.75** (0.505)	**126.58** (0.5039)	**124.6** (0.5035)	**118.05** (0.5021)	**114.37** (0.5014)
0.5	**201.36** (0.5189)	**183.16** (0.5152)	**167.99** (0.5122)	**139.56** (0.5065)	**128.09** (0.5042)	**121.91** (0.5029)	**118.05** (0.5021)	**116.61** (0.5019)	**112.01** (0.501)	**109.48** (0.5005)
0.7	**199.18** (0.5185)	**176.33** (0.5139)	**159.37** (0.5105)	**131.75** (0.505)	**121.91** (0.5029)	**116.88** (0.5019)	**113.82** (0.5013)	**112.7** (0.5011)	**109.11** (0.5005)	**107.14** (0.5003)
0.9	**197.1** (0.518)	**170.56** (0.5127)	**152.72** (0.5092)	**126.58** (0.5039)	**118.05** (0.5021)	**113.82** (0.5013)	**111.28** (0.5008)	**110.35** (0.5007)	**107.37** (0.5003)	**105.74** (0.5001)
1	**196.09** (0.5178)	**167.99** (0.5122)	**149.94** (0.5086)	**124.6** (0.5035)	**116.61** (0.5019)	**112.7** (0.5011)	**110.35** (0.5007)	**109.48** (0.5005)	**106.73** (0.5002)	**105.23** (0.5001)
1.5	**191.35** (0.5169)	**157.56** (0.5101)	**139.56** (0.5065)	**118.05** (0.5021)	**112.01** (0.501)	**109.11** (0.5005)	**107.37** (0.5003)	**106.73** (0.5002)	**104.7** (0.5001)	**103.6** (0.5)
2	**187.06** (0.516)	**149.94** (0.5086)	**132.82** (0.5052)	**114.37** (0.5014)	**109.48** (0.5005)	**107.14** (0.5003)	**105.74** (0.5001)	**105.23** (0.5001)	**103.6** (0.5)	**102.75** (0.5)

Table 5.2: Optimum Values of $\hat{\mu}_{12}$ (in bracket) and *PRE* of $\hat{\bar{Y}}_{12}$ with respect to $\hat{\bar{Y}}_{C}$.

α / β	**0.01**	**0.05**	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**	**1.0**	**1.50**	**2.00**
0.01	**188.8** (0.4741)	**188.42** (0.4742)	**187.96** (0.4743)	**186.15** (0.4748)	**184.41** (0.4752)	**182.74** (0.4757)	**181.14** (0.4761)	**180.36** (0.4763)	**176.68** (0.4773)	**173.33** (0.4781)
0.05	**188.42** (0.4742)	**186.59** (0.4747)	**184.41** (0.4752)	**176.68** (0.4773)	**170.26** (0.479)	**164.84** (0.4804)	**160.2** (0.4816)	**158.12** (0.4821)	**149.58** (0.4843)	**143.23** (0.4859)
0.1	**187.96** (0.4743)	**184.41** (0.4752)	**180.36** (0.4763)	**167.44** (0.4797)	**158.12** (0.4821)	**151.08** (0.4839)	**145.57** (0.4853)	**143.23** (0.4859)	**134.43** (0.4879)	**128.61** (0.4892)
0.3	**186.15** (0.4748)	**176.68** (0.4773)	**167.44** (0.4797)	**145.57** (0.4853)	**134.43** (0.4879)	**127.68** (0.4894)	**123.14** (0.4902)	**121.39** (0.4905)	**115.51** (0.4914)	**112.17** (0.4918)
0.5	**184.41** (0.4752)	**170.26** (0.479)	**158.12** (0.4821)	**134.43** (0.4879)	**124.48** (0.49)	**118.99** (0.4909)	**115.51** (0.4914)	**114.21** (0.4915)	**110.02** (0.4921)	**107.73** (0.4925)
0.7	**182.74** (0.4757)	**164.84** (0.4804)	**151.08** (0.4839)	**127.68** (0.4894)	**118.99** (0.4909)	**114.46** (0.4915)	**111.67** (0.4918)	**110.64** (0.492)	**107.4** (0.4926)	**105.67** (0.4934)
0.9	**181.14** (0.4761)	**160.2** (0.4816)	**145.57** (0.4853)	**123.14** (0.4902)	**115.51** (0.4914)	**111.67** (0.4918)	**109.35** (0.4922)	**108.51** (0.4923)	**105.87** (0.4933)	**104.48** (0.4942)
1	**180.36** (0.4763)	**158.12** (0.4821)	**143.23** (0.4859)	**121.39** (0.4905)	**114.21** (0.4915)	**110.64** (0.492)	**108.51** (0.4923)	**107.73** (0.4925)	**105.32** (0.4936)	**104.05** (0.4945)
1.5	**176.68** (0.4773)	149.58 (0.4843)	**134.43** (0.4879)	**115.51** (0.4914)	**110.02** (0.4921)	**107.4** (0.4926)	**105.87** (0.4933)	**105.32** (0.4936)	**103.62** (0.4949)	**102.74** (0.4959)
2	**173.33** (0.4781)	**143.23** (0.4859)	**128.61** (0.4892)	**112.17** (0.4918)	**107.73** (0.4925)	**105.67** (0.4934)	**104.48** (0.4942)	**104.05** (0.4945)	**102.74** (0.4959)	**102.08** (0.4968)

It is observed from Tables 5.1 and 5.2 that

(i) the suggested class of estimators $\hat{\bar{Y}}_{11}$ and $\hat{\bar{Y}}_{12}$ perform better than the class of estimators $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$ which does not utilize information on additional auxiliary variable '*z*' in terms of optimum fraction of sample to be drawn afresh and also in terms of percent relative efficiency.

(ii) the PRE's of the proposed estimators $\hat{\bar{Y}}_{11}$ and $\hat{\bar{Y}}_{12}$ decreases with increasing values of (α, β).

(iii) the larger gain in efficiency is observed by using the proposed classes of estimators $\hat{\bar{Y}}_{11}$ and $\hat{\bar{Y}}_{12}$ over the class of estimators $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$ when the value of (α, β) are small.

(iv) the gain in efficiency by using the proposed class of estimators $\hat{\bar{Y}}_{11}$ over $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$ is larger as compared to the class of estimators $\hat{\bar{Y}}_{12}$ over $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$.

It is to be mentioned that a practical choice of α and β, fixed by the experience of the experimenter from repeated surveys can always provide better results than the class of estimators $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$.

Remark 2: The procedure outlined in this paper can be also applied to the randomized response models mentioned in Priyanka and Trisandhya (2019) {see Arcos *et al.* (2015) and Odumade and Singh (2008) *etc*} to get the efficient estimators of the population mean at current (second) move using information on additional non-sensitive auxiliary variable at both the occasion in two occasion successive sampling.

6. Conclusion

This article presents some classes of estimators for estimating the population mean at current (second) occasion in two occasions successive sampling using information on an additional non-sensitive auxiliary variable in presence of randomized response model. The properties of the suggested classes are studied under randomized response models. Optimum replacement policies have been elaborated. It has been demonstrated that the proposed classes of estimators are better than the class of estimators which does not utilize non-sensitive auxiliary information. Numerical illustration is given in support of the present study. It has been shown that there is appreciable gain in efficiency by using the proposed classes of estimators over the class of estimators $\hat{\bar{Y}}_C \left(or\ \hat{\bar{Y}}_J\right)$. Thus the proposed study is recommended for its use in practice.

Acknowledgement

Authors are thankful to the learned referee and the Editor – in - Chief for their valuable suggestions regarding improvement of the paper.

References

Arnab, R. (2011). Alternative estimators for randomized response techniques in multi-character surveys. *Communication in Statistics-Theory and Method*, **40**, 1839-1848.

Arnab, R. and Singh, S. (2013). Estimation of mean of sensitive characteristics for successive sampling. *Communication in Statistics-Theory and Method*, **42**, 2499-2524.

Arcos, A., Rueda, M. and Singh, S. (2015). A generalized approach to randomized response for quantitative variables. *Quality and Quantity*, **49(3)**, 1239-1256.

Bar- Lev, S. K., Bobivitch, K. and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika,* **60**, 255 – 260.

Biradar, R. S. and Singh, H. P. (2001). Successive sampling using auxiliary information on both the occasions. *Calcutta Statistical Association Bulletin,* **51 (203-204),** 243-251.

Chaudhuri, A. and Christofides, T. C. (2013). Indirect questioning in sample surveys. *Heidelberg: Springer*.

Chaudhari, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York, USA: Marcel-Dekker.

Christofides, T. C. (2003). A generalized randomized response technique. *Metrika* **57**, 195, 200.

Das, A. K. and Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhya,* **40**,134-148.

Diana, G. and Perri, P. F. (2012). Calibration-based approach to sensitive data: a simulation study. *Journal of Applied Statist*ics, **39**, 53–65.

Eichhorn, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data*. Journal of Statistical Planning and Inference,* **7**, 307-316.

Ghangurde, P. D. and Rao, J. N. K. (1969). Some results on sampling over two occasions. *Sankhya,* **A31**, 463–472.

Greenberg, B., Abul- Ela, A., Simmons, W. R. and Horvitz, D. G. (1969). The unreleased question randomized response: Theoretical framework. *Journal of American Statistical Association,* **64 (5)**, 29-539.

Gjestvang, C. R. and Singh, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statist*ics*.* **36(12),** 1361-1367.

Horvitz, D. G., Shah, B. V. and Simmons, W. R. (1967). The unrelated question randomized response model. Proceedings of Social Statistical section, *Journal of American Statistical Association,* **326**, 65-72.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Road Bulletin no.* **304**, 1-104,Ames, USA.

Kuk, A. Y. C. (1990). Asking sensitive question indirectly. *Biometrika*, **77**, 439-442.

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, **56**, 93–95.

Mangat, N. S. and Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, **77**, 439–442.

Narain, R. D. (1953). On the recurrence formula in sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, **5**, 96–99.

Okafor, F. C. and Arnab, R. (1987). Some strategies of two-stage sampling for estimating population ratio over two occasions. *Australian journal of statistics,* **29(2)**.128–142.

Odumade, O. and Singh, S. (2008). Generalized forced quantitative randomized response model: A unified approach. *Journal of Indian Society of Agricultural Statistics*, **62(3)**, 244- 252.

Odumade, O. and Singh, S. (2009). Improved bar-lev, bobovitch, and boukai randomized response models. *Communication in Statistics-Simulation and Computation*, **38(3),** 473-502.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*: *Series B (Methodological)*, **12 (2)**, 241–55.

Priyanka, K., Trisandhya, P. and Mittal, R. (2017). Dealing sensitive characters on successive occasions through a general class of estimators using scrambled response technique. *Metron,* DOI 10.1007/s 40300-017-0131-1.

Priyanka, K. and Trisandhya, P. (2018). A composite class of estimators using scrambled response mechanism for sensitive population mean in successive sampling. *Communication in Statistics - Theory and Meth*od, https:// doi. Org/ 10.1080/03610926.2017.1422762.

Priyanka, K. and Trisandhya, P. (2019). Some classes of estimators for sensitive population mean on successive moves. *Journal of Statistical Theory and Practice,* https:// doi. Org/ 10.1007/ s42519-018-0008-5.

Pollock, K. H. and Bek, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of American Statistical Association***, 71(356),** 884-886.

Saha, A. (2007). A simple randomized response technique in complex surveys. *Metron*, **65(1),** 59-66.

Sen, A. R. (1973). Some theory of sampling on successive occasions. *Australian Journal of Statistics,* **15 (2)**, 105-110.

Singh, D. (1968). Estimates in successive sampling using a multi-stage design. *Journal of the American Statistical Association,* **63**, 99-112.

Singh. H. P. and Gorey, S. M. (2017). Efficient estimation of population mean of sensitive variable in presence of scrambled responses, *Communication in Statistics- Theory and Methods*, **46(19)**, 9557-9565.

Singh, S. and Joarder, A. (1996). Regression analysis using scrambled analysis. *Australian Journal of Statistics*, **38(2)**, 201 – 211.

Singh, S. and Kim, J. M. (2007). A pseudo-empirical log-likelihood estimator using scrambled responses. *Statistical and Probability Letter,* **81**, 345–351.

Singh, S., Kim, J. M. and Grewal, I. S. (2008). Imputing and jackknifing scrambled responses. *Metron,* **66(2),** 183-204.

Singh, G. N. and Priyanka, K. (2008). On the use of several auxiliary variates to improve the precision of estimates at current occasion. *Journal of Indian Society of Agricultural Statistics,* **62(3),** 253–265.

Singh, G. N., Suman, S., Khetan, M. and Paul, C. (2017). Some estimation procedures of sensitive character using scrambled response techniques in successive sampling. *Communication in Statistics- Theory and Methods,* **46 (6)**, 2249-2269.

Singh, H. P. and Mathur, N. (2005). Estimation of population mean when coefficient of variation is known using scrambled response technique. *Journal of Statistical Planning and Inference,* **131,** 135–144.

Singh, H. P. and Pal, S. K. (2017). A review of chain- type estimators and a new chain-type multivariate regression approach for population mean under two-occasion successive sampling. *Proceedings of the National Academy of Sciences, India Section A- Physical Science,* 87, 31-56

Singh, H. P. and Vishwakarma, G. K. (2007). A general class of estimators in successive sampling. *Metron,* **65(2)**, 201-207.

Singh, H. P. and Vishwakarma, G. K. (2009). A general procedure for estimating population mean in successive sampling. *Communication in Statistics- Theory and Methods,* **38(2)**, 293-308.

Srivastava, S. K. (1971). A generalized estimator for population mean of finite population using multi-auxiliary information. *Journal of American Statistical Association*, **66**, 404-407.

Srivastava, S. K. (1980). A class of estimators using auxiliary information in sample surveys. *The Canadian Journal of Statistics*, **8**,253-254.

Srivastava, S. K. and Jhajji, H. S. (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhya C*, **42,** 87-96.

Srivastava, S. K. and Jhajji, H. S. (1981). A class of estimating of the population mean in survey sampling using auxiliary information. *Biometrika,* **68**,341-343.

Tarray, T. A. and Singh, H. P. (2015). Some improved additive randomized response models utilizing higher order moments ratios of scrambling variable. *Model Assisted Statistics and Applications,* **10(4),** 361–383.

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

Warner, S. L. (1965). Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, **60**, 63-69.

Yu, B., Jin, Z., Tian, J. and Gao, G. (2015). Estimation of sensitive proportion by randomized response data in successive sampling. *Computational and Mathematical Methods in Medicines,* DOI: 10.1155/2015/172918.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 141–148

Some Recursive Constructions of α– Resolvable Group Divisible Designs

Shyam Saurabh[1] and Kishore Sinha[2]
[1]*Ranchi University, Ranchi, India*
[2] *Formerly at Birsa Agricultural University, Ranchi, India and*
#201 Maitry Residency, Kalkere Main Road, Bangalore– 560043.

Received: 10 January 2021; Revised: 04 March 2021; Accepted: 10 March 2021

Abstract

Some recursive constructions of *α*– resolvable group divisible designs with λ_1=0 from certain group divisible designs with λ_1=0 are presented here. In this process some non–isomorphic solutions of group divisible designs are also obtained. A group divisible design with λ_1=0 is used in the construction of Group divisible codes, optimal codes over a cyclic group and LDPC codes, see Ge (2007), Chee *et al.* (2008) and Xu *et al.* (2019). Transversal designs are special classes of such designs.

Keywords: Group divisible designs and codes; Uniform frames; *α*– resolvable designs; Optimal and LDPC *codes*.

1. Introduction

1.1. Group divisible design

In statistical design theory, a *Group divisible* (*GD*) *design* is defined as an arrangement of v (= mn; m, $n \geq 2$; m groups of n treatments) treatments into b blocks such that each block contains k ($<v$) distinct treatments, each treatment occurs r times and any pair of distinct treatments which are first associates occur together in λ_1 blocks and in λ_2 blocks if they are second associates. Furthermore, if $r-\lambda_1=0$ then the GD design is singular; if $r-\lambda_1>0$ and $rk-v\lambda_2=0$ then it is semi-regular (*SR*); and if $r-\lambda_1>0$ and $rk-v\lambda_2>0$, the design is regular (*R*).

In Combinatorial design theory; a GD design with index λ is a triple $(\mathcal{V}, \mathcal{G}, \mathcal{B})$ where

(i) $\mathcal{V}$ is a finite set of elements and $\mathcal{G}$ is a set of subsets of $\mathcal{V}$, called groups, which partition $\mathcal{V}$;

(ii) $\mathcal{B}$ is a set of subsets of $\mathcal{V}$, called blocks, such that every pair of elements from distinct groups occurs in exactly λ blocks and $|G \cap B| \leq 1$ for all $G \in \mathcal{G}, B \in \mathcal{B}$.

If all the blocks of the GD design have the same size k and all the groups have the same size n then the GD design is uniform and it is known as (k, λ) – GD design of type n^m for some positive integer m.

Clearly a (k, λ) – GD design of type n^m is a GD design with λ_1=0 (in Statistical design theory). A GD design with λ_1=0 is used in the construction of Group divisible codes, optimal codes over a cyclic group and LDPC codes, see Ge (2007), Chee *et al.* (2008) and Xu *et al.* (2019). A semi- regular GD design with λ_1=0 and $k = m$ is also known as a transversal design in combinatorial design theory.

Corresponding Author: Kishore Sinha
Email: kishore.sinha@gmail.com

1.2. Partial resolution classes and frames

Suppose b blocks of a block design D (v, b, r, k) can be divided into $t = r/\alpha$ classes, each of size $\beta = v\alpha/k$ such that in each class of β blocks every treatment of D is replicated α times. Then these t classes are known as α– resolution (or parallel) classes and the design is called an α– resolvable design. When α=1 the design is said to be resolvable and the classes are called resolution classes.

Let $\mathcal{V}$ be the set of treatments, $\mathcal{G}$ be a set of subsets of $\mathcal{V}$ (called groups), each of size n which partitions $\mathcal{V}$ and $\mathcal{B}$ be the set of subsets of $\mathcal{V}$, called blocks of a GD design. A partial resolution class is a collection of blocks such that every treatment of $\mathcal{V}\backslash G, G \in \mathcal{G}$ occurs exactly once and the treatments of G do not occur.

A uniform (k, λ) – frame of index λ is a GD design with parameters v, b, r, k, $\lambda_1 = 0$, $\lambda_2 = \lambda$, m, n such that

(i) the block set $\mathcal{B}$ can be partitioned into a family $\mathcal{R}$ of partial resolution classes, and

(ii) each $R \in \mathcal{R}$ can be associated with a group $G \in \mathcal{G}$ so that R contains every treatment of $\mathcal{V}\backslash G$ exactly once.

Such frame is of type n^m where n is the size of the each group of the GD design. For details see Furino *et al.* (1996; pp. 27– 28) and Ge and Miao (2007).

A comprehensive coverage of constructions of GD designs may be found in Dey (1986, 2010), Raghavarao (1971), Raghavarao and Padgett (2005) and Saurabh *et al.* (2021). Some construction methods of α– resolvable partially balanced incomplete block designs may be found in Sinha and Dey (1982), Kadowaki and Kageyama (2009) and Saurabh and Sinha (2020). Here, some recursive constructions of α– resolvable group divisible designs with λ_1=0 from certain group divisible designs with λ_1=0 are presented. In this process some non–isomorphic solutions of group divisible designs are also obtained. All the group divisible designs constructed here have λ_1=0. *SRX* numbers are from Clatworthy (1973).

2. Recursive Constructions

Theorem 1: The existence of a resolvable SRGD design with parameters

$$v, b = nr, r, k, \lambda_1 = 0, \lambda_2 = \lambda, m, n \tag{1}$$

implies the existence of another resolvable SRGD design with parameters

$$v^* = v + pn, b^* = n^p b, r^* = n^p r, k^* = k + p, \lambda_1^* = 0, \lambda_2^* = n^p \lambda, m^* = m + p, n^* = n \tag{2}$$

p is a positive integer.

Proof: Let $R^1, R^2, \ldots, R^r$ be the resolution classes of the SRGD design with parameters (1). Let $B_1^i, B_2^i, \ldots, B_n^i$ be arbitrarily chosen blocks in its i^{th} resolution class and $\theta_1, \theta_2, \ldots, \theta_n$ be the new treatments other than the v treatments of the GD design. We form b/r resolution classes $R_1^i, R_2^i, \ldots, R_{b/r}^i$ corresponding to a resolution class R^i of the SRGD design with parameters (1) as follows:

R_1^i	R_2^i	...	$R_{b/r}^i$
$B_1^i \cup \{\theta_1\}$	$B_1^i \cup \{\theta_2\}$	$\cdots$	$B_1^i \cup \{\theta_n\}$
$B_2^i \cup \{\theta_2\}$	$B_2^i \cup \{\theta_1\}$	$\cdots$	$B_2^i \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B_n^i \cup \{\theta_n\}$	$B_n^i \cup \{\theta_{n-1}\}$	$\cdots$	$B_n^i \cup \{\theta_1\}$

This process is continued for all the resolution classes of the SRGD design. New treatments are added once only in each block. The union of these new resolution classes generates the blocks of another resolvable SRGD design with parameters:

$$v^* = v + n, b^* = nb, r^* = b = nr, k^* = k + 1, \lambda_1^* = 0, \lambda_2^* = \lambda n, m^* = m + 1, n^* = n.$$

Further, by repeated application of this process p (p a positive integer) times we will get a resolvable SRGD design with parameters (2).

Example 1: Consider the following resolution classes of *SR*23: $v = b = 9$, $r = k = 3$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = n = 3$

R^1	R^2	R^3
1, 2, 3	1, 5, 9	1, 6, 8
4, 5, 6	2, 6, 7	2, 4, 9
7, 8, 9	3, 4, 8	3, 5, 7

Then using Theorem 1, the resolution classes of *SR*43: $v = 12$, $r = 9$, $k = 4$, $b = 27$, $\lambda_1 = 0$, $\lambda_2 = 3$, $m = 4$, $n = 3$ are

R_1^1	R_2^1	R_3^1	R_1^2	R_2^2	R_3^2
1, 2, 3, 10	1, 2, 3, 11	1, 2, 3, 12	1, 5, 9, 10	1, 5, 9, 11	1, 5, 9, 12
4, 5, 6, 11	4, 5, 6, 12	4, 5, 6, 10	2, 6, 7, 11	2, 6, 7, 12	2, 6, 7, 10
7, 8, 9, 12	7, 8, 9, 10	7, 8, 9, 11	3, 4, 8, 12	3, 4, 8, 10	3, 4, 8, 11

R_1^3	R_2^3	R_3^3
1, 6, 8, 10	1, 6, 8, 11	1, 6, 8, 12
2, 4, 9, 11	2, 4, 9, 12	2, 4, 9, 10
3, 5, 7, 12	3, 5, 7, 10	3, 5, 7, 11

Example 2: Consider the following resolution classes of *SR*36: $v = b = 8$, $r = k = 4$, $\lambda_1 = 0$, $\lambda_2 = 2$, $m = 4$, $n = 2$

R^1	R^2	R^3	R^4
1, 2, 3, 4	1, 2, 7, 8	1, 3, 6, 8	1, 4, 6, 7
5, 6, 7, 8	3, 4, 5, 6	2, 4, 5, 7	2, 3, 5, 8

Then using Theorem 1, the resolution classes of *SR*54: $v = 10$, $r = 8$, $k = 5$, $b = 16$, $\lambda_1 = 0$, $\lambda_2 = 4$, $m = 5$, $n = 2$ are

R_1^1	R_2^1	R_1^2	R_2^2	R_1^3	R_2^3
1, 2, 3, 4, 9 5, 6, 7, 8, 10	1, 2, 3, 4, 10 5, 6, 7, 8, 9	1, 2, 7, 8, 9 3, 4, 5, 6, 10	1, 2, 7, 8, 10 3, 4, 5, 6, 9	1, 3, 6, 8, 9 2, 4, 5, 7, 10	1, 3, 6, 8, 10 2, 4, 5, 7, 9

R_1^4	R_2^4
1, 4, 6, 7, 9 2, 3, 5, 8, 10	1, 4, 6, 7, 10 2, 3, 5, 8, 9

Theorem 2: The existence of an *n*– resolvable SRGD design with parameters (1) implies the existence of another *n*- resolvable SRGD design with parameters

$$v^* = v + pn, b^* = n^p b, r^* = n^p r, k^* = k + p, \lambda_1^* = 0, \lambda_2^* = n^p \lambda, m^* = m + p, n^* = n. \quad (3)$$

p is a positive integer.

Proof: Since the SRGD design with parameters (1) is *n*-resolvable, the number of resolution classes is r/n and the number of blocks in each resolution class is $bn/r = vn/k = n^2$. Let $R^1, R^2, \ldots, R^{r/n}$ be the resolution classes of SRGD design with parameters (1). Let $B_1^i, B_2^i, \ldots, B_{n^2}^i$ be arbitrarily chosen blocks in its i^{th} resolution class and $\theta_1, \theta_2, \ldots, \theta_n$ be the new treatments distinct from the *v* treatments of the SRGD design. We construct *n* resolution classes corresponding to a resolution class R^i of the SRGD design as follows:

R_1^i	R_2^i	...	R_n^i
$B_1^i \cup \{\theta_1\}$	$B_1^i \cup \{\theta_2\}$	$\cdots$	$B_1^i \cup \{\theta_n\}$
$B_2^i \cup \{\theta_1\}$	$B_2^i \cup \{\theta_2\}$	$\cdots$	$B_2^i \cup \{\theta_n\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B_n^i \cup \{\theta_1\}$	$B_n^i \cup \{\theta_2\}$	$\cdots$	$B_n^i \cup \{\theta_n\}$
$B_{n+1}^i \cup \{\theta_2\}$	$B_{n+1}^i \cup \{\theta_1\}$	$\cdots$	$B_{n+1}^i \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B_{2n}^i \cup \{\theta_2\}$	$B_{2n}^i \cup \{\theta_1\}$		$B_{2n}^i \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$B_{n^2-1}^i \cup \{\theta_n\}$	$B_{n^2-1}^i \cup \{\theta_{n-1}\}$	$\cdots$	$B_{n^2-1}^i \cup \{\theta_1\}$
$B_{n^2}^i \cup \{\theta_n\}$	$B_{n^2}^i \cup \{\theta_{n-1}\}$	$\cdots$	$B_{n^2}^i \cup \{\theta_1\}$

We continue this process for all the resolution classes of the SRGD design with parameters (1). New treatments are added once only in each block. The union of these new resolution classes generates the blocks of another *n*-resolvable SRGD design with parameters:

$$v^* = v + n, b^* = bn, r^* = b = nr, k^* = k + 1, \lambda_1^* = 0, \lambda_2^* = n\lambda, m^* = m + 1, n^* = n.$$

Further, by repeated application of this process *p* (*p* a positive integer) times we will get an *n*-resolvable SRGD design with parameters (3).

Example 3: Consider the following 2– resolvable solution of *SR*66: $v = 12$, $b = 8$, $r = 4$, $k = 6$, $\lambda_1 = 0$, $\lambda_2 = 2$, $m = 6$, $n = 2$:

R^1	R^2
1, 2, 3, 4, 5, 6	1, 3, 8, 5, 10, 12
5, 6, 7, 8, 9, 10	2, 4, 5, 7, 9, 12
1, 2, 9, 10, 11, 12	1, 4, 6, 8, 9, 11
3, 4, 7, 8, 11, 12	2, 3, 6, 7, 10, 11

Then using Theorem 2, a 2– resolvable solution of *SR*82: $v = 14$, $r = 8$, $k = 7$, $b = 16$, $\lambda_1 = 0$, $\lambda_2 = 4$, $m = 7$, $n = 2$ is obtained as:

R_1^1	R_2^1	R_1^2	R_2^2
1, 2, 3, 4, 5, 6, 13	1, 2, 3, 4, 5, 6, 14	1, 3, 8, 5, 10, 12, 14	1, 3, 8, 5, 10, 12, 13
5, 6, 7, 8, 9, 10, 13	5, 6, 7, 8, 9, 10, 14	2, 4, 5, 7, 9, 12, 14	2, 4, 5, 7, 9, 12, 13
1, 2, 9, 10, 11, 12, 14	1, 2, 9, 10, 11, 12, 13	1, 4, 6, 8, 9, 11, 13	1, 4, 6, 8, 9, 11, 14
3, 4, 7, 8, 11, 12, 14	3, 4, 7, 8, 11, 12, 13	2, 3, 6, 7, 10, 11, 13	2, 3, 6, 7, 10, 11, 14

Remark 1: Clatworthy (1973) reported a resolvable and a 4-resolvable solution for *SR*82 while the solution presented here is 2-resolvable. Hence the present solution is non– isomorphic.

The following Table lists n-resolvable ($n \geq 1$) solutions of some SRGD designs using Theorems 1 and 2 with $p = 1$:

Table 1: SRGD Designs

No.	Original design	Derived design	Source
1	*SR*1, Resolvable	*SR*19, Resolvable	Th. 1
2	*SR*6, Resolvable	*SR*25, Resolvable	Th. 1
3	*SR*23, Resolvable	*SR*43, Resolvable	Th. 1
4	*SR*36, Resolvable and 2– resolvable	*SR*54, Resolvable and 2– resolvable	Th. 1 & 2
5	*SR*52, 2– resolvable	*SR*69, 2– resolvable	Th. 2
6	*SR*66, 2– resolvable	*SR*82, 2– resolvable, Non- isomorphic	Th. 2

Theorem 3: The existence of a uniform (k, λ) – frame with parameters: $v = b$, $r = k$, $\lambda_1 = 0$, $\lambda_2 = \lambda$, m, n implies the existence of a resolvable GD design with parameters

$$v^* = v, b^* = nv, r^* = n(k+1), k^* = k+1, \lambda_1^* = 0, \lambda_2^* = \lambda n\,(k+1)/(k-1)\,, m^* = m, \quad n^* = n. \tag{4}$$

where n is equal to the number of blocks in partial resolution class of a uniform (k, λ) – frame.

Proof: Let $R^1, R^2, \ldots, R^t$ be the partial resolution classes. Let $B_1^i, B_2^i, \ldots, B_n^i$ be arbitrarily chosen blocks and $\{\theta_1, \theta_2, \ldots, \theta_n\}$ be the missing group in i^{th} partial resolution class of the uniform (k, λ) – frame. We form n resolution classes corresponding to a partial resolution class R^i as follows:

R_1^i	R_2^i	...	R_n^i
$B_1^i \cup \{\theta_1\}$	$B_1^i \cup \{\theta_2\}$	$\cdots$	$B_1^i \cup \{\theta_n\}$
$B_2^i \cup \{\theta_2\}$	$B_2^i \cup \{\theta_1\}$	$\cdots$	$B_2^i \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$B_n^i \cup \{\theta_n\}$	$B_n^i \cup \{\theta_{n-1}\}$	$\cdots$	$B_n^i \cup \{\theta_1\}$

We continue this process for all the partial resolution classes of an (k, λ) – frame. One of the n treatments from missing groups are added once only in each block. The union of these new resolution classes generates the blocks of a resolvable GD design with parameters (4).

Example 4: Consider a (3, 1) – frame of type 2^4 whose partial resolution classes are:

Partial Resolution Classes	R^1	R^2	R^3	R^4
groups	{1, 5}	{2, 4}	{3, 6}	{7, 8}
blocks	{2, 6, 7}	{1, 6, 8}	{1, 4, 7}	{1, 2, 3}
	{3, 4, 8}	{3, 5, 7}	{2, 5, 8}	{4, 5, 6}

Then using Theorem 3, we obtain a resolvable GD design with parameters *SR*39: $v = 8$, $r = 8$, $k = 4$, $b = 16$, $\lambda_1 = 0$, $\lambda_2 = 4$, $m = 4$, $n = 2$ whose resolution classes are:

R_1^1	R_2^1	R_1^2	R_2^2	R_1^3	R_2^3
{1, 2, 6, 7}	{2, 5, 6, 7}	{1, 2, 6, 8}	{1, 4, 6, 8}	{1, 3, 4, 7}	{1, 4, 6, 7}
{3, 4, 5, 8}	{1, 3, 4, 8}	{3, 4, 5, 7}	{2, 3, 5, 7}	{2, 5, 6, 8}	{2, 3, 5, 8}

R_1^4	R_2^4
{1, 2, 3, 7}	{1, 2, 3, 8}
{4, 5, 6, 8}	{4, 5, 6, 7}

Theorem 4: The existence of a nonresolvable SRGD design with parameters (1) implies the existence of another *r*–resolvable SRGD design with parameters

$$v^* = v + n, b^* = n^2 r, r^* = nr, k^* = k + 1, \lambda_1^* = 0, \lambda_2^* = \lambda n, m^* = m + 1, n^* = n. \quad (5)$$

Proof: Let $B_1^i, B_2^i, \ldots, B_{nr}^i$ be arbitrarily chosen blocks of the nonresolvable SRGD design with parameters (1) and $\theta_1, \theta_2, \ldots, \theta_n$ be the new treatments other than v treatments of the SRGD design. We constitute an *r*– resolvable solution of a GD design with parameters (5) whose blocks are given as follows:

R_1	R_2	...	R_n
$B_1 \cup \{\theta_1\}$	$B_1 \cup \{\theta_2\}$	$\cdots$	$B_1 \cup \{\theta_n\}$
$B_2 \cup \{\theta_1\}$	$B_2 \cup \{\theta_2\}$		$B_2 \cup \{\theta_n\}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$B_r \cup \{\theta_1\}$	$B_r \cup \{\theta_2\}$		$B_r \cup \{\theta_n\}$
$B_{r+1} \cup \{\theta_2\}$	$B_{r+1} \cup \{\theta_1\}$	$\cdots$	$B_{r+1} \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$		$\vdots$
$B_{2r} \cup \{\theta_2\}$	$B_{2r} \cup \{\theta_1\}$	$\cdots$	$B_{2r} \cup \{\theta_{n-1}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$B_{nr-1} \cup \{\theta_n\}$	$B_{nr-1} \cup \{\theta_{n-1}\}$	$\cdots$	$B_{nr-1} \cup \{\theta_1\}$
$B_{nr} \cup \{\theta_n\}$	$B_{nr} \cup \{\theta_{n-1}\}$	$\cdots$	$B_{nr} \cup \{\theta_1\}$

New treatments are added once only in each block.

When *r*=*n* in Theorem 4, by the repeated application of the process in Theorem 2 we get:

Corollary 1: The existence of a nonresolvable SRGD design with parameters (1) implies the existence of another *n*– resolvable SRGD design with parameters:

$v^* = v + (p+1)n, b^* = n^{p+3}, r^* = n^{p+2}, k^* = k + p + 1, \lambda_1^* = 0, \lambda_2^* = \lambda n^{p+1}, m^* = m + p + 1, n^* = n$; p is a positive integer. (6)

Example 5: Consider a SRGD design *SR*41: $v = 12$, $b = 9$, $r = 3$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$, $m = 4$, $n = 3$ whose blocks are given as:

{1, 2, 3, 4}, {4, 5, 7, 10}, {4, 6, 9, 11}, {1, 6, 7, 8}, {2, 5, 8, 11}, {3, 8, 9, 10}, {1, 10, 11, 12}, {2, 7, 9, 12}, {3, 5, 6, 12}.

Then using Theorem 4, a 3-resolvable solution of *SR*57: $v = 15$, $b = 27$, $r = 9$, $k = 5$, $\lambda_1 = 0$, $\lambda_2 = 3$, $m = 5$, $n = 3$ is given as:

R_1	R_2	R_3
{1, 2, 3, 4, 13}	{1, 2, 3, 4, 14}	{1, 2, 3, 4, 15}
{4, 5, 7, 10, 13}	{4, 5, 7, 10, 14}	{4, 5, 7, 10, 15}
{4, 6, 9, 11, 13}	{4, 6, 9, 11, 14}	{4, 6, 9, 11, 15}
{1, 6, 7, 8, 14}	{1, 6, 7, 8, 15}	{1, 6, 7, 8, 13}
{2, 5, 8, 11, 14}	{2, 5, 8, 11, 15}	{2, 5, 8, 11, 13}
{3, 8, 9, 10, 14}	{3, 8, 9, 10, 15}	{3, 8, 9, 10, 13}
{1, 10, 11, 12, 15}	{1, 10, 11, 12, 13}	{1, 10, 11, 12, 14}
{2, 7, 9, 12, 15}	{2, 7, 9, 12, 13}	{2, 7, 9, 12, 14}
{3, 5, 6, 12, 15}	{3, 5, 6, 12, 13}	{3, 5, 6, 12, 14}

Since $r = n = 3$ here, following Corollary 1 for $p = 1$ we will get a SRGD design with parameters: $v = 18$, $b = 81$, $r = 27$, $k = 6$, $\lambda_1 = 0$, $\lambda_2 = 9$, $m = 6$, $n = 3$.

Remark 2: Clatworthy (1973) reported a resolvable solution for *SR*57 whereas a 3-resolvable solution is obtained here for the same. Hence the present solution is non– isomorphic.

Remark 3: The association scheme of the derived GD design in the Theorems 1, 2 and 4 is obtained by adjoining a new row: $mn+1, mn+2, \ldots, n(m+1)$ to the $m \times n$ association scheme of the original GD design.

Acknowledgement

The authors are thankful to an anonymous referee and Editor-in-Chief for valuable suggestions in improving the paper.

References

Chee, M. Y., Ge, G and Ling, C. H. A. (2008). Group Divisible Codes and Their Application in the Construction of Optimal Constant-Composition Codes of Weight Three. *IEEE Transactions on Information Theory*, **54**, 3552– 3564.

Clatworthy, W. H. (1973). Tables of two-associate-class partially balanced designs. *National Bureau of Standards* (*U.S.*), *Applied Mathematics*, Series **63**.

Dey, A. (1986). *Theory of Block Designs.* Wiley Eastern, New Delhi.

Dey, A. (2010). *Incomplete Block Designs*. Hindustan Book Agency, New Delhi.

Furino, S., Miao, Y. and Yin, J. (1996). Frames and resolvable designs, Uses constructions and existence. CRC Press, Boca Raton.

Ge, G. (2007). *Group Divisible Designs*. In Colbourn, C. J. and Dinitz, J. H. (Eds) Handbook of Combinatorial Designs, Second edition, Chapman & Hall/CRC, New York, 255–260.

Ge, G. and Miao, Y. (2007). *PBDs, Frames, and Resolvability*. In Colbourn, C. J. and Dinitz, J. H. (Eds) Handbook of Combinatorial Designs, Second edition, Chapman & Hall/CRC, New York, 261–265.

Kadowaki, S. and Kageyama, S. (2009). Existence of affine α-resolvable PBIB designs with some constructions. *Hiroshima Mathematical Journal*, **39,** 293–326.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. John Wiley, New York.

Raghavarao, D. and Padgett, Lakshmi V. (2005). Block designs. Analysis, combinatorics and applications. *Series on Applied Mathematics*, **17**, World Scientific, Singapore.

Saurabh, S. and Sinha, K. (2020). Some new resolvable group divisible designs. *Communications in Statistics - Theory and Methods*, doi. org/10.1080/ 03610926.2020.1817487.

Saurabh, S., Sinha, K. and Singh, M. K. (2021). Unifying constructions of group divisible Designs. *Statistics and Applications*, **19(1)** To appear.

Sinha, K. and Dey, A. (1982). On resolvable PBIB designs. *Journal of Statistical Planning and Inference*, **6**, 179 –181.

Xu, H., Yu, Z., Feng, D. and Zhu, H. (2019). New construction of partial geometries based on group divisible designs and their associated LDPC codes. *Physical Communication*, https://doi.org/10.1016/j.phycom.2019.100970.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series) pp 149-164

An Economical Study of Two-State Queueing Model with Initial Customers and Bernoulli Schedules

Rimmy Sharma and Indra Rani
Department of Statistics and O.R.,
Kurukshetra University, Kurukshetra (Haryana), INDIA

Received: 20 January 2021; Revised: 08 March 2021; Accepted: 11 March 2021

Abstract

Present paper studies the behaviour of few clientele available in the system, when the system starts and derives the time dependent probabilities of a single server queueing model with impatient customers ("balking and reneging"), Bernoulli schedule and multiple vacations. "Server accepts a customer with fix probability ρ or commences a vacation of random duration with probability $(1-\rho)$". An arriving customer may balk (do not enter) or renege pursuant to the negative exponential distribution. Time dependent probabilities are computed with the help of recurrence relations and provide us a better understanding of the behaviour of the system. Finally, measureable outcomes are calculated with the help of Maple software.

Key words: M/M/1; Bernoulli schedule; Impatient customers; Multiple vacations; Laplace transform.

1. Introduction

In recent years, various authors have studied queueing models from productive point of view. Many real-life situations occur where clienteles are dejected by longer queue and as a result clienteles have to wait long to get into service upon arrival. In queues, "balking and reneging are common phenomena, as a consequences the customer either decides to join the queue or depart after joining the queue without getting service due to impatience". Queueing systems with impatient units ("balking and reneging") have engross many authors because of their extensive applications in many practical situations such as perishable goods in supermarkets, emergency room in hospitals *etc.* Haight (1957), Haight (1959) obtained probabilities for impatient customers ("balking and reneging") respectively. Anker and Gafarian (1963), Anker and Gafarian (1963) calculated steady state probabilities with impatient clienteles for a finite and an infinite queueing model respectively. Abou-El- Ata (1991) derived steady state probabilities for single-server Markovian queue with state dependent arrivals and impatient clienteles. Seddy *et al.* (2009) obtained time dependent probabilities by using generating function technique for c-servers Markovian queueing model. Bouchentouf and Messabihi (2018) obtained time independent probabilities for a heterogeneous server queueing system with feedback. Sharma and Indra (2020) obtained time dependent probabilities for a two-dimensional state Markovian queueing model with reneging.

From the past few decades, Vacations Queueing system has attracted much attention from numerous researchers. "Vacation: when the server finishes serving a unit and finds the system empty, however, it goes away for a length of time". Cooper (1970) was the first who

Corresponding Author: Indra Rani
E-mail: indra@kuk.ac.in

talked about vacation model and obtained waiting time distribution for M/G/1 model by using Laplace Stieltjes transform. There are different kinds of vacation policies available in literature *i.e.*, single vacation, multiple vacations, Bernoulli schedule, t- policy and *so on*. "In Multiple vacations policy, server keeps on taking vacations until it finds at least one customer waiting in the system at the instant of vacation completion". Bacot and Dshalalow (2001) obtained time independent probabilities for single server bulk queueing system with multiple vacations. Altman and Yechailli (2006) analyzed both single and multiple vacations cases and calculated time independent probabilities by using PGF technique for different markovian model. Banik (2009) obtained time independent probabilities and queue length distribution at various epochs for an infinite-buffer single server queueing model. Another feature that is widely used in queueing models is the Bernoulli schedule. "In Bernoulli schedule the server serves the new customer with probability ρ or takes a vacation with probability $(1-\rho)$". Keilson and Servi (1987) introduced the concept of Bernoulli schedule and obtained steady state probabilities for M/G/1 queueing model. Khedhairi and Tadj (2007) studied bulk service queueing system for both discrete and continuous time by using semi regenerative technique. The combined effects of impatience customers and multiple vacations were studied by numerous researchers such as: Ramaswamy and Servi (1988) calculated joint distribution for busy period of M/G/1 model. Madan *et al.* (2003) calculated steady state probabilities by using generating function technique for a queueing system with two parallel servers. Yue *et al.* (2003) derived the closed-form expressions for the system sizes for a queueing model with variant of multiple vacations. Choudhary *et al.* (2007) calculated steady state behaviour in terms of recursive solutions of batch arrival queue with two phases of heterogeneous service. Ammar (2015) obtained transient probabilities in terms of modified Bessel function by employing PGF technique.

All aforesaid authors have worked on the concept that, there is no clienteles available in the system when the system starts. Thus, the main aim of the paper is to make a model that is more applicable in day-to-day life activities such as railway booking counters, banks, doctor clinics, *etc*. In call centre: Calls arriving to a call centre are managed by agent to answer the calls. Primary calls are automatically answered by machines (*i.e.* initially a few clienteles are always present). The behaviour of the call may depend on several circumstances including waiting time and others. Each individual call may decide to balk or wait for some time and it may happen that clienteles abandon their call when their patience time expires. Server after completing all the clienteles (calls) in the system can go for vacation and after coming back from vacation if there are no clienteles available (calls) in the system server can go on vacation again.

To obtain the time dependent solution by taking together all the above mentioned parameters is very interesting. As transient probability obtained by recursive technique does not involve heavy algebraic manipulations. "The two dimensional concept helps us to understand the probability of exactly *a*- arrivals and *b*- services occurs over a time interval of length *t*". Validation of the model in form of tables is also done with the previous existing results. Graphical analysis shows the impact of parameters on measuring outcomes. Finally an expected cost model is discussed.

2. Assumptions and Notations

i. Inter-arrival time, Service times, vacation times and reneging times are exponentially distributed with parameter λ, μ, w, and ξ respectively.

ii. On arrival a customer either decides to join the queue with probability β or balk with probability $(1-\beta)$.
iii. Initially there are 'n' customers present at time $t=0$ *i.e.* $P_{n,0}(n,0)=1$.
iv. When a customer has just been served and other customers are present, the server accepts a customer with fix probability ρ or commences a vacation of random duration with probability $(1-\rho)$
v. The system state is given by (a, b), where a is the number of arrivals and b is the number of departures up to time t, *i.e.*

$$P(n,0)=\sum_{b=0}^{\infty}P_{n+b,b,V}(n,0)=1 \tag{1}$$

3. Model

Define
$P_{a,b,B}(c,t)$ = The probability of exactly a arrivals, b departures and c- customers remain in the system by time t and the server is busy corresponding to the queue; $b< a$
$P_{a,b,V}(c,t)$ = The probability of exactly a arrivals, b departures and c- customers remain in the system by time t and the server is on vacation; $b \leq a$
$P_{a,b}(c,t)$ = The probability that there are exactly a arrivals and b departures and c- customers remain in the system by time t; $b \leq a$

3.1. Equations of the system

$$\frac{d}{dt}P_{a,b,V}(c,t)=-(\lambda\beta+w)P_{a,b,V}(c,t)+(\lambda\beta)P_{a-1,b,V}(c-1,t)+\mu(1-\rho)P_{a,b-1,B}(c+1,t)\left(1-\delta_{b,0}\right)$$

$$0\leq b<a\,,c\geq 1 \tag{2}$$

$$\frac{d}{dt}P_{a,a,V}(0,t)=-(\lambda\beta)P_{a,a,V}(0,t)+\mu P_{a,a-1,B}(1,t)\left(1-\delta_{a,0}\right),\qquad a\geq 0 \tag{3}$$

$$\frac{d}{dt}P_{a,b,B}(c,t)=-(\lambda\beta+\mu+(c-1)\xi)P_{a,b,B}(c,t)+\lambda\beta P_{a-1,b,B}(c-1,t)\left(1-\delta_{b,a-1}\right)+wP_{a,b,V}(c,t)+(\mu\rho+c\xi)P_{a,b-1,B}(c+1,t)$$

$$0\leq b<a,c\geq 1 \tag{4}$$

Clearly,

$$P_{a,b}(c,t)=P_{a,b,V}(c,t)+P_{a,b,B}(c,t)(1-\delta_{(a,b)})\;a\geq b\geq 0 \tag{5}$$

3.2. Findings of equations

Solving above equation recursively with the help of Laplace transform:

$$\bar{P}_{a,b,V}(0,s)=\frac{1}{(s+\lambda\beta)}\delta_{(n,0)}P_{0,0,V}(0,0),\ a=0=b \tag{6}$$

$$\bar{P}_{a,0,V}(a,s)=(\lambda\beta)^{a}\bar{H}_{a,1,0}^{\lambda\beta+w,\lambda\beta,0}(s)\delta_{(n,0)}\,P_{0,0,V}(0,0)+\sum_{d=1}^{\infty}(\lambda\beta)^{a-d}\bar{H}_{a-d+1,0,0}^{\lambda\beta+w,0,0}(s)\delta_{(n,d)}P_{d,0,V}(d,0),a>0 \tag{7}$$

$$\bar{P}_{a,0,B}(a,s)=\ w.\sum_{f=1}^{\infty}\frac{(\lambda\beta)^{a-f}}{\prod_{d=0}^{a-f}\{s+\lambda\beta+\mu+(a-d-1).\xi\}}\ \bar{P}_{f,0,V}(f,s),\quad a\ >\ 0 \tag{8}$$

$$\bar{P}_{a,b,V}(c,s)=\sum_{e=0}^{a-b}(\lambda\beta)^{a-e-b}\,\mu(1-\rho)^{(1-\delta_{(e,0)})}\,\bar{H}_{a-b+1-e-\delta_{(e,0)},\delta_{(e,0)},0}^{\lambda\beta+w,\lambda\beta,0}(s)\bar{P}_{b+e,b-1,B}(e+1,s)+(\lambda\beta)^{a-b}\bar{H}_{a-b,1,0}^{\lambda\beta+w,\lambda\beta,0}(s)\ \delta_{(n,0)}P_{b,b,V}(0,0)+\sum_{d=b+1}^{\infty}(\lambda\beta)^{a-d}\ \bar{H}_{a-d+1,0,0}^{\lambda\beta+w,0,0}(s)\ \delta_{(n,d-b)}P_{d,b,V}(d-b,0),\qquad a>b>\ 0 \tag{9}$$

$$\bar{P}_{a,b,B}(c,s)=\sum_{f=b+1}^{\infty}\frac{(\lambda\beta)^{a-f}.\{\mu\rho+(f-b).\xi\}}{\prod_{d=0}^{a-f}\{s+\lambda\beta+\mu+(a-d-b-1).\xi\}}\ \bar{P}_{f,b-1,B}(f-b+1,s)+w.\sum_{f=b+1}^{\infty}\frac{(\lambda\beta)^{a-f}.}{\prod_{d=0}^{a-f}\{s+\lambda\beta+\mu+(a-d-b-1).\xi\}}\ \bar{P}_{f,b,V}(f-b,s),\ a\ >b>\ 0 \tag{10}$$

$$\bar{P}_{a,a,V}(0,s)\ =\frac{\mu}{(s+\lambda\beta)}\bar{P}_{a,a-1,B}(1,s)(1-\delta_{(a,0)})+\frac{1}{(s+\lambda\beta)}\delta_{(n,0)}P_{a,a,V}(0,0),\quad a\ >0 \tag{11}$$

3. Substantiations

The Laplace Transform $\bar{P}_{a.}(c,s)$of the probability $P_{a.}(c,t)$ that exactly *a* unit arrives by the time *t*":

a) $\bar{P}_{a.}(s)=\sum_{b=0}^{a}\left[(\bar{P}_{a,b,V}(c,s)+\bar{P}_{a,b,B}(c,s)(1-\delta_{a,b})\right]$

$=\sum_{b=0}^{a}\bar{P}_{a,b}(c,s)=\ \frac{(\lambda\beta)^{a}}{(s+\lambda\beta)^{a+1}}$

And its Inverse Laplace transform is $\quad P_{a.}(c,t)=\ \frac{e^{-\lambda\beta t}(\lambda\beta t)^{a}}{a!}$

b) $\sum_{a=0}^{\infty}\sum_{b=0}^{a}\{\bar{P}_{a,b,V}(c,s)+\bar{P}_{(a,b,B}(c,s)(1-\delta_{(a,b)})\}=\frac{1}{s}$

$\sum_{a=0}^{\infty}\sum_{b=0}^{a}\{P_{a,b,V}(c,t)+P_{a,b,B}(c,t)(1-\delta_{(a,b)})\}=\ 1$

5. Analytical Results

i. Pr $\{a$ arrivals in $(0, t)\} = \frac{e^{-\lambda t}(\lambda t)^a}{a!} = \sum_{b=0}^{a} P_{a,b}(c,t) = P_{a,.}(c,t)$

Table 1: For Exactly *a* customers served by time *t*

λ	μ	w	β	ξ	T	a	$\frac{e^{-\lambda t} * (\lambda t)^a}{a!}$	$\sum_{b=0}^{a} P_{a,b}(t)$
1	2	1	1	1	3	1	0.149361	0.1493612051
1	2	1	1	1	3	3	0.224042	0.2240418076
1	2	1	1	1	3	5	0.100819	0.1008188135
2	2	1	1	1	3	1	0.014873	0.01487251306
2	2	1	1	1	3	3	0.089235	0.08923507838
2	2	1	1	1	3	5	0.160623	0.1606231410
1	2	1	1	1	4	1	0.073263	0.07326255556
1	2	1	1	1	4	3	0.195367	0.1953668148
1	2	1	1	1	4	5	0.156293	0.1562934518
2	2	1	1	1	4	1	0.002684	0.002683701023
2	2	1	1	1	4	3	0.028626	0.02862614425
2	2	1	1	1	4	5	0.091604	0.09160366157
2	4	1	1	1	4	5	0.091604	0.09160366160
1	2	1	1	1	4	4	0.195367	0.1953668148
1	2	1	1	1	3	6	0.050409	0.05040940672
3	2	1	1	1	3	1	0.0011106	0.001110688237
3	2	1	1	1	3	3	0.0149942	0.01499429120
3	2	1	1	1	3	5	0.0607268	0.06072687936

The last Column of Table-1 completely matches with Table-1 of Pegden and Rosenshine (1982).

ii. The probability that exactly *b* number of customers have been served.

Server is on vacation *i.e.* $\sum_{a=b}^{\infty} P_{a,b,V}(c,t)$

Server is busy *i.e.* $\sum_{a=b}^{\infty} P_{a,b,B}(c,t)$ are based on the following relationship

$P_{.,b}(c,t) = \sum_{a=b}^{\infty} P_{a,b}(c,t)$ where $P_{a,b}(c,t)$ is defined in equation (5)".

Table 2: For exactly *b* customers served by time *t*

$\lambda = 1,\ \mu = 4,\ w = 1,\ n = 0,\ \beta = 1,\ \xi = 1,\ \rho = 1,\ b = 0$ to 6 $P_{.b}(c,t) = P_{.b,B}(c,t) + P_{.b,V}(c,t)$				
$t = 1$	$t = 3$	$t = 5$	$t = 7$	$t = 10$
.483485	.392200	.338348	.222880	.0739809
.333343	.293382	.232925	.136406	.0375456
.13382	.156248	.112178	.0580149	.0132014
.03866	.0738856	.048044	.0216532	.0040095
.00875	.032712	.019831	.0075688	.0010965
.00161	.013342	.007936	.0024681	.0002598
.000252	.0047242	.002920	.0007192	.0000477
.99992	**.966494**	**.762184**	**.44971**	**.130142**

Table-2 Coincides with table I of Hubbard *et al.* (1986)

iii. $P_N(t) = P(Exactly\ N\ customers\ in\ the\ system\ by\ time\ t)$. $P_N(t)$ can be expressed in terms of $P_{a,b}(c,t)$ and is based on the relationship, we have

$$P_N(t) = \sum_{b=0}^{\infty} P_{b+N,b}(N,t) \quad \& \quad P_N(t) = P_B(N,t) + P_V(N,t)$$

where, $P_B(N,t) = \sum_{b=0}^{\infty} P_{b+N,b,B}(N,t)$, $P_V(N,t) = \sum_{b=0}^{\infty} P_{b+N,b,V}(N,t)$

Table 3: Exactly *N* customers in the system

$\lambda = 1, \mu = 2$, $w = 1$, $n = 1$, $\rho = 0.2$, $\beta = 0.6$, $\xi = 0.7$, $N = 0$ to 6		
$t = 1$	$t = 3$	$t = 5$
.245353	.367102	.362711
.462100	.323646	.303897
.221013	.181321	.158741
.058932	.079292	.063808
.010870	.028463	.020800
.001528	.008188	.005287
.000166	.001609	.000837
.999963	**.989623**	**.916082**

iv. The server's utilization time, server's vacation time *i.e.* the fraction of time the server is busy and the fraction of time server is on vacation until time *t* can also be expressed in terms of $P_{a,b}(c,t)$

Server's utilization time: $U(t) = \sum_{a=0}^{\infty} \sum_{b=0}^{a} P_{a,b,B}(c,t)$

Server's vacation time:$V(t) = \sum_{a=0}^{\infty} \sum_{b=0}^{a} P_{a,b,V}(c,t)$

Table 4: Server's utilization time and Server's vacation time

λ =1.7, μ = 2.5, w = 1.5, n = 1, ρ = 0.6, β = 0.7, ξ = 1.1			
	$\sum_{a=0}^{\infty} \sum_{b=0}^{a} P_{a,b,V}(c,t)$	$\sum_{a=0}^{\infty} \sum_{b=0}^{a} P_{a,b,B}(c,t)$	Total
t = 1	0.650353	0.342156	**0.992509**
t = 2	0.621177	0.285446	**0.906623**
t = 3	0.513311	0.198857	**0.712168**
t = 4	0.364730	0.118833	**0.483563**
t = 5	0.228494	0.0633109	**0.291805**

6. Performance Indices

(a) The expected number of customers in the system *E(L)* is given by

$$E(L) = \sum_{N=1}^{\infty} N \ [P_B(N,t) + P_V(N,t)]$$

(b) The expected number of customers in the queue is given by

$$E(L_q) = \sum_{N=1}^{\infty} (N-1)\ [P_B(N,t) + P_V(N,t)]$$

(c) The throughput is

$$T(P) = \sum_{N=1}^{\infty} \mu[\, P_B(N,t) + P_V(N,t)]$$

(d) Mean balking rate is given by

$$B.R. = \sum_{N=1}^{\infty} \lambda(1-\beta)\,\{P_B(N,t) + P_V(N,t)\}$$

(e) Mean reneging rate is given by

$$R.R. = \sum_{N=1}^{\infty} \xi(N-1)\ [P_B(N,t) + P_V(N,t)]$$

(f) Average rate of customer loss *(L.R.)* is given by

$$L.R. = B.R. + R.R.$$

7. Cost Model

We make a expected cost function for the given system, considering cost per some unit of time of server for all the parameters considered above.

Let

$C1$ = Vacation Cost

$C2$ = Busy cost

$C3$ = Idle Cost.

$C4$ = unit is waiting for service.

$C5$ = unit joins the system and is served.

$C6$ = customer renege or balks.

Mean cost function per some unit time:

$$C = C1*P_{VAC}+C2*P_{BUSY}+C3*P_{IDLE}+C4*E(Lq)+C5*[E(L)-E(Lq)]+C6*L.R.$$

We fix cost elements $C1 = 100$, $C2 = 110$, $C3 = 120$, $C4 = 150$, $C5 = 130$, $C6 = 140$.

8. Graphical Presentations

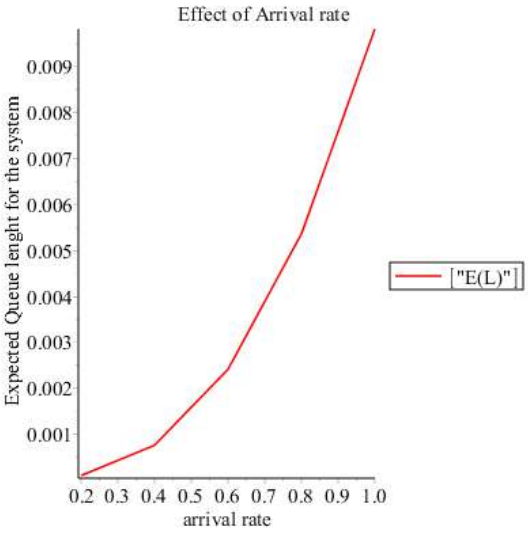


Figure 1: Arrival rate on *E(L)*

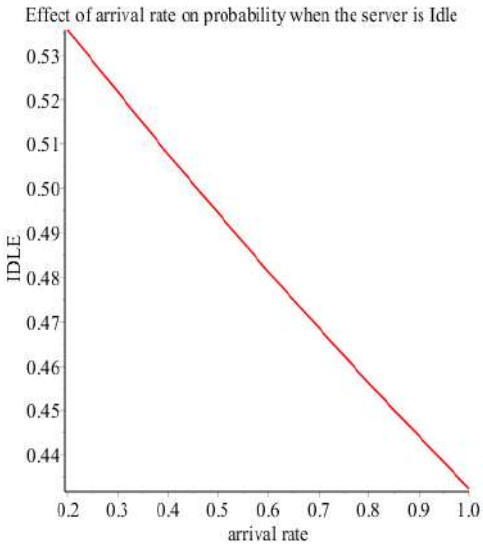


Figure 2: Arrival rate on probability of server remains idle

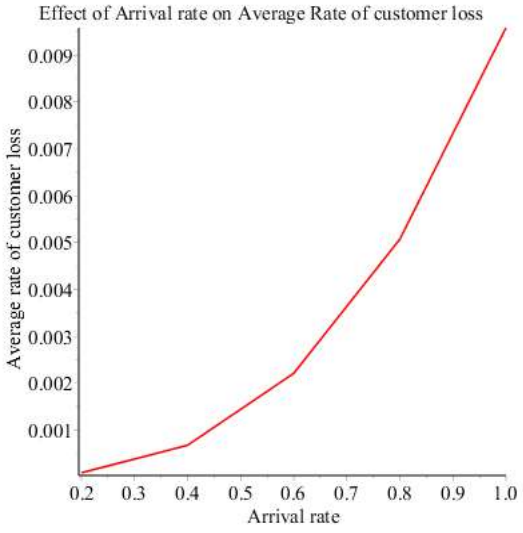


Figure 3: Arrival rate on *L.R.*

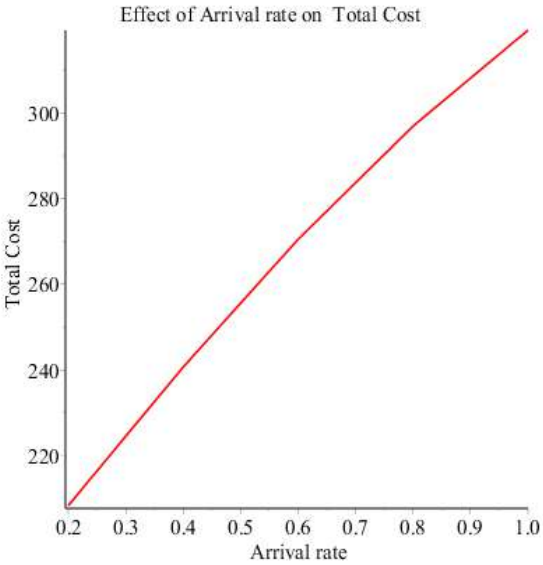


Figure 4: Arrival rate on Cost function

In Figures 1 to 4 we fix $w = 1.5$, $n = 1$, $\mu = 2.5$, $\beta = 0.4$, $\rho = 0.6$, $\xi = 1.1$, $t = 1$ and vary the values of λ. These graphs show that the expected number of customers in the system, expected cost and average rate of customer loss increase as arrival rate increases but probability of server remains idle decreases as λ increases

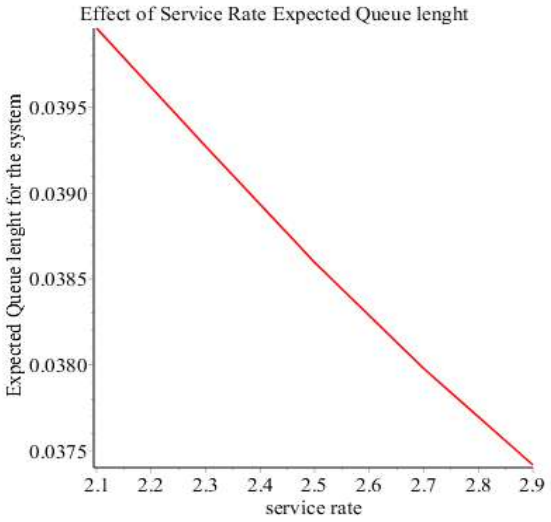


Figure 5: Service rate on *E(L)*

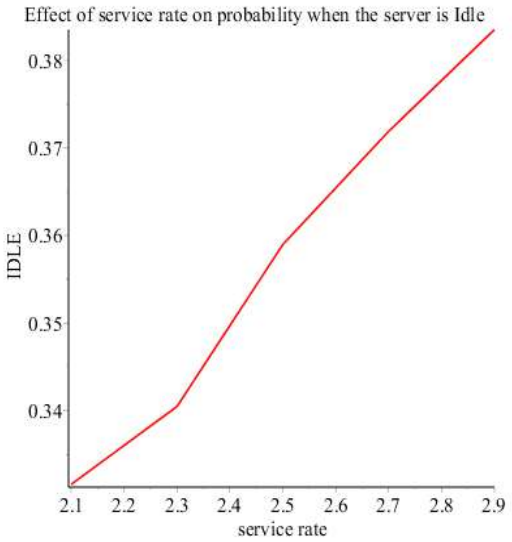


Figure 6: Service rate on throughput

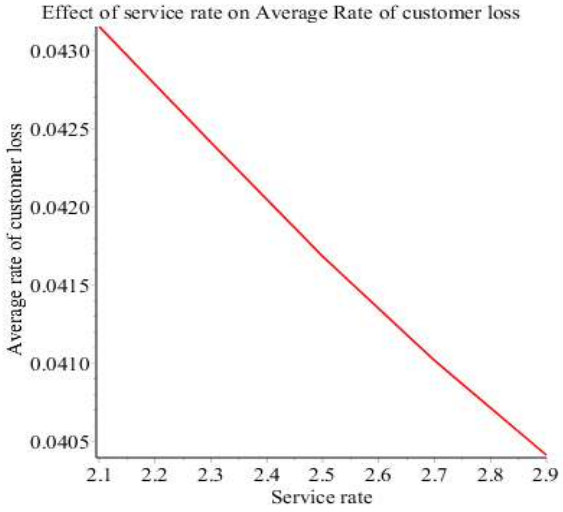


Figure 7: Service rate on *L.R*

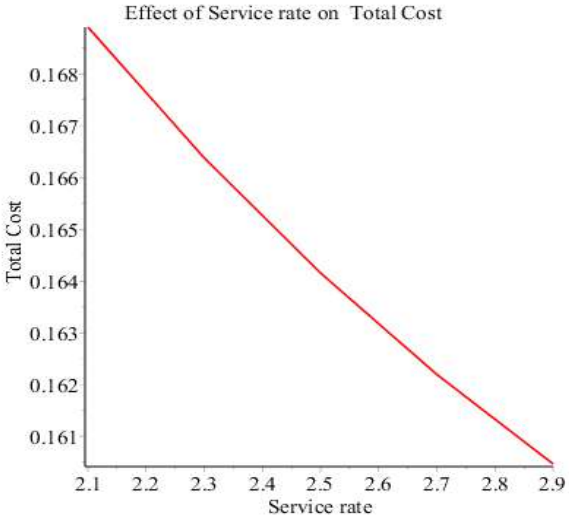


Figure 8: Service rate on cost function

In Figures 5 to 8 we fix $\lambda = 1.7$, $w = 1.5$, $n = 1$, $\beta = 0.4$, $\xi = 1.1$, $\rho = 0.6$ and vary the values of μ. These graphs show that the expected number of customers in the system, expected cost and average rate of customer loss decrease as μ increases but probability of server remains idle increases as μ increases".

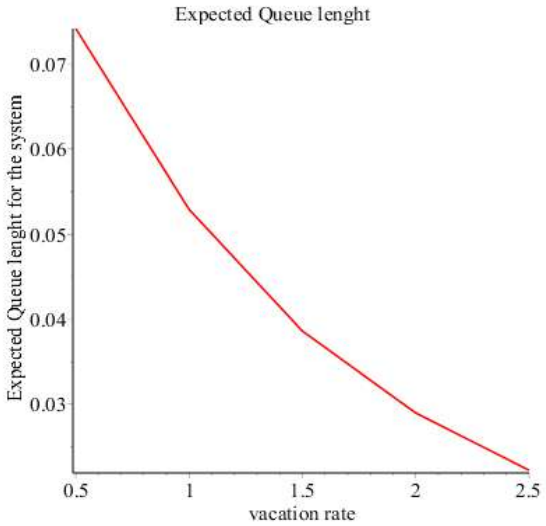


Figure 9: Vacation Rate (w) on *E(L)*

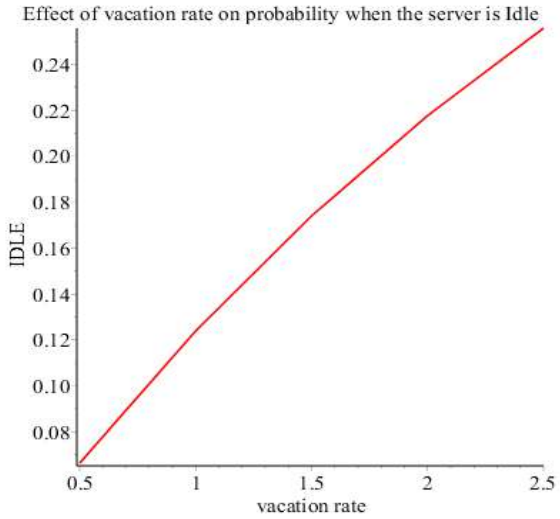


Figure 10: Vacation Rate (w) on Probability of server remains idle

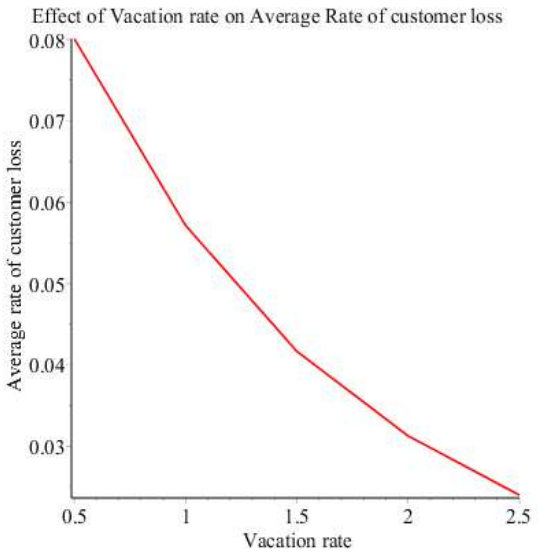


Figure 11: Vacation Rate (w) on *L.R.*

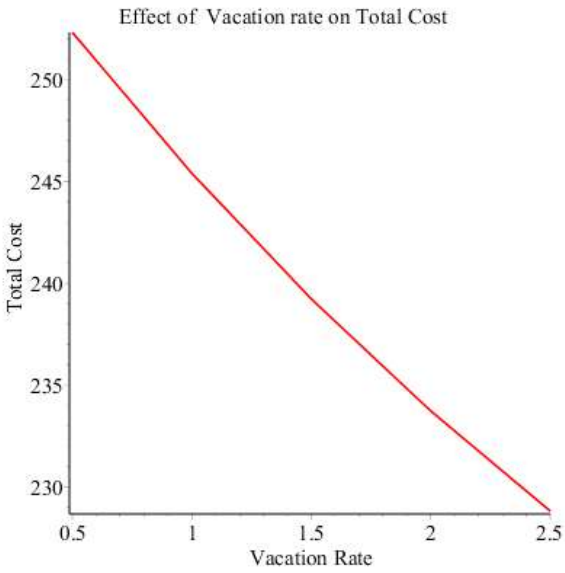


Figure 12: Vacation Rate (w) on Cost Function

In Figures 9 to 12 we fix λ=1.7, μ= 2.5, n=1, β = 0.4, ρ = 0.6, ξ=1.1 and vary the values of w. "These graphs show that the expected number of customers in the system, expected cost, expected queue length and average rate of customer loss decrease as w increases but probability of server remains idle increases as vacation rate increases".

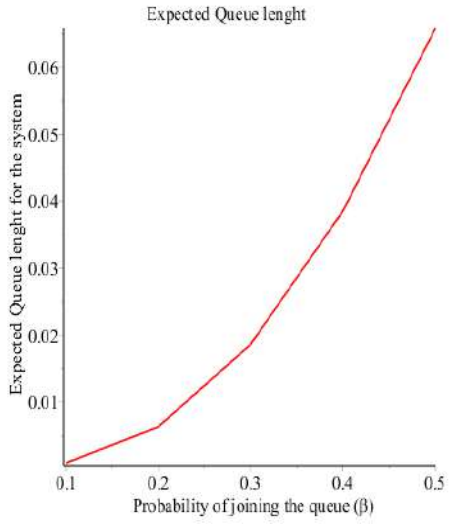


Figure 13: β on $E(L)$

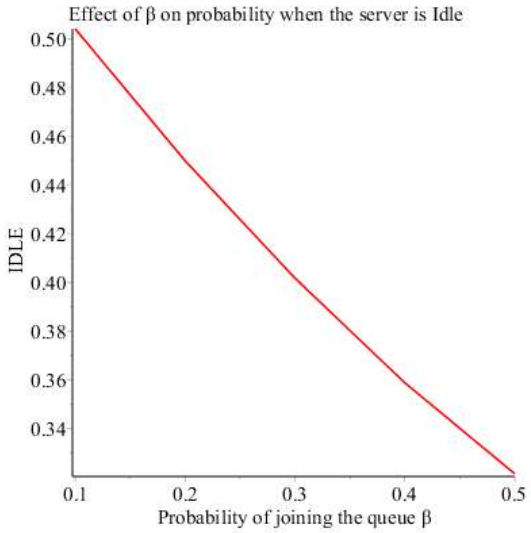


Figure 14: β on Probability of server remains idle

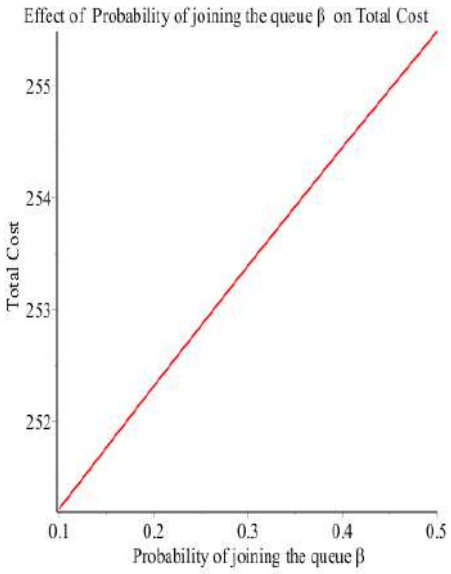


Figure 15: Impact of β on Total Cost

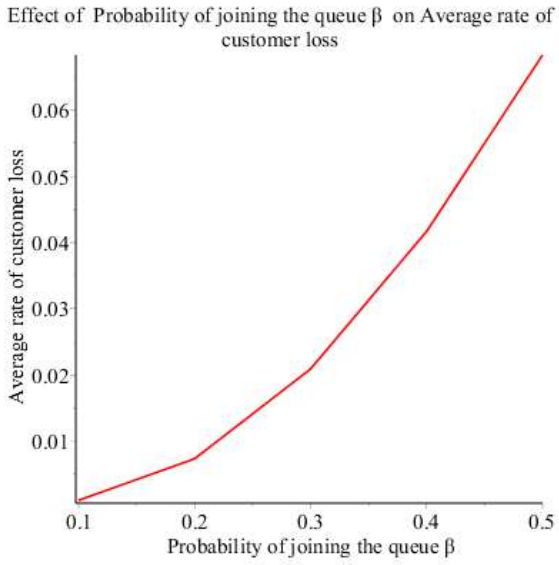


Figure 16: Impact of β on *L.R.*

In Figures 13 to 16 we fix λ=1.7, μ=2.5, n=1,w = 1.5, ξ=1.1, ρ = 0.6 and vary the values of β. "These graphs show that the expected number of customers in the system, average rate of customer loss, expected cost increase as β increases but probability of server remains idle decreases as β increases".

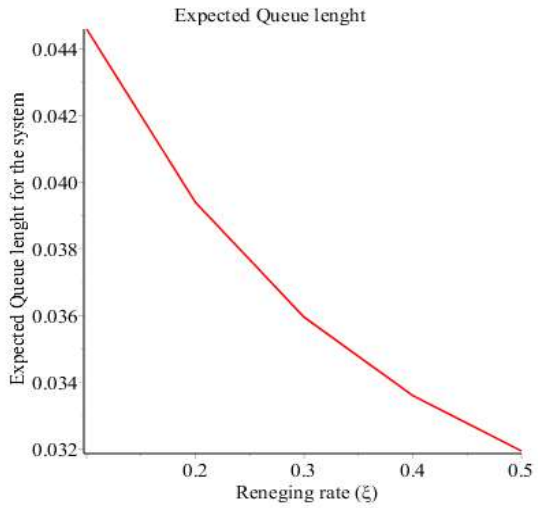


Figure 17: Reneging rate on *E(L)*

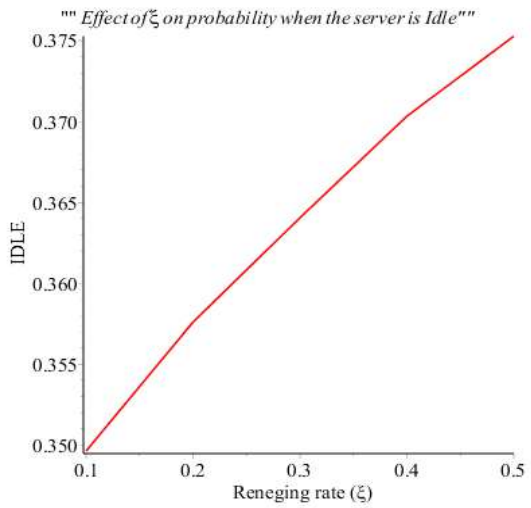


Figure 18: Impact of ξ on probability that server remains Idle

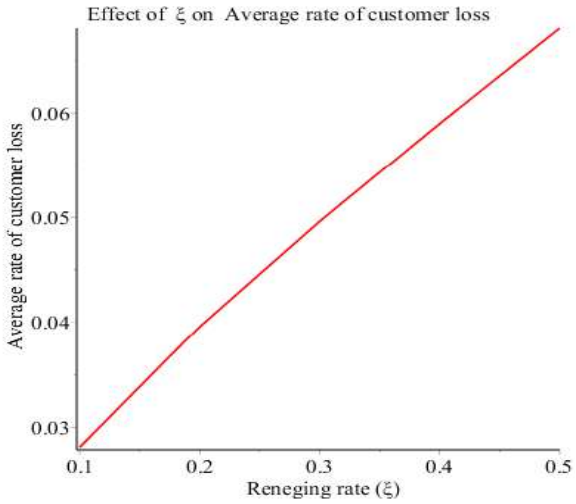


Figure 19: Impact of ξ on *L.R.*

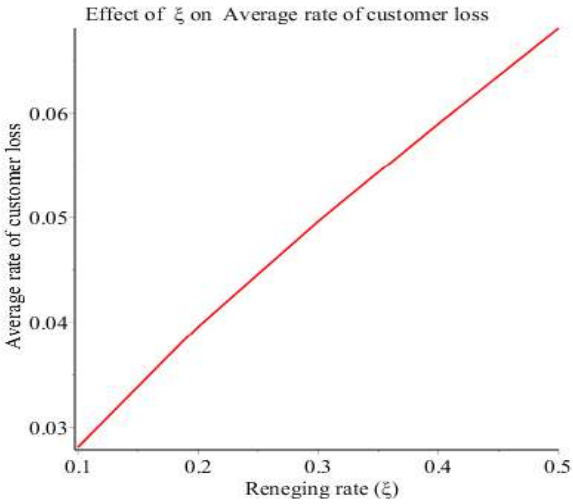


Figure 20: Impact of ξ on cost function

In Figures 17 to 20 we fix $\lambda = 1.7$, $w = 1.5$, $n = 1$, $\mu = 2.5$, $\beta = 0.4$, $\rho = 0.6$, $t = 1$ and vary the values of ξ. These graphs show that expected cost, probability of server remains idle and average rate of customer loss increase as ξ increases but expected queue length decreases as ξ increases.

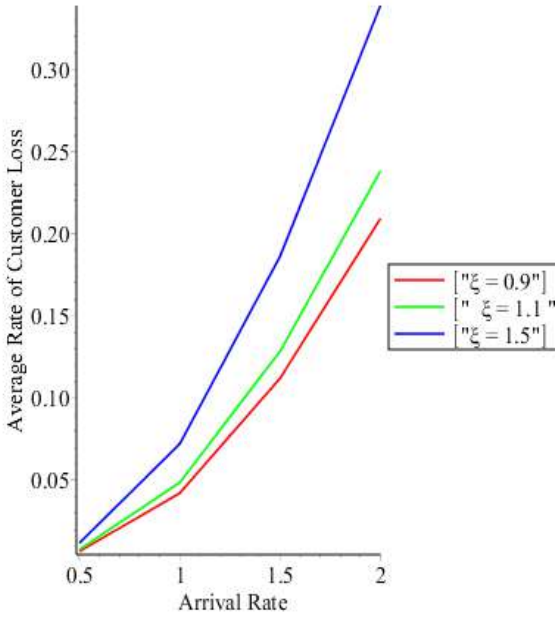


Figure 21: Impact of λ on *E(L)*

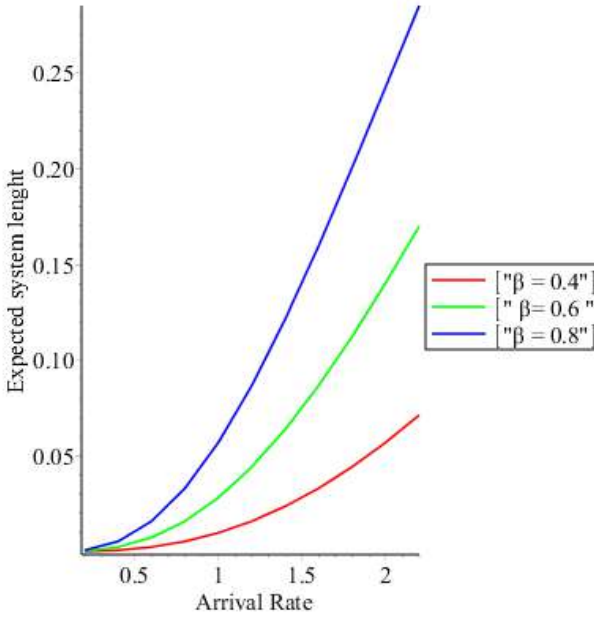


Figure 22: Impact of λ on *L.R.*

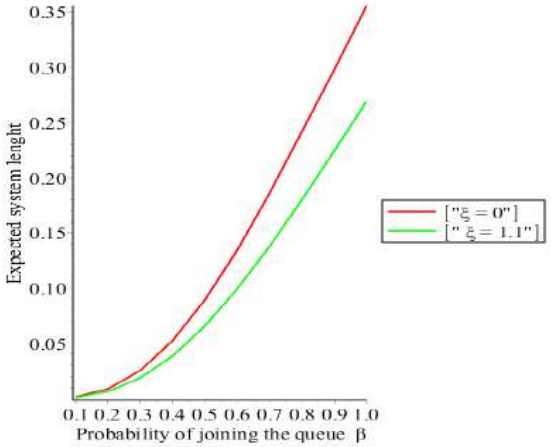


Figure 23: Impact of β on *E(L)*

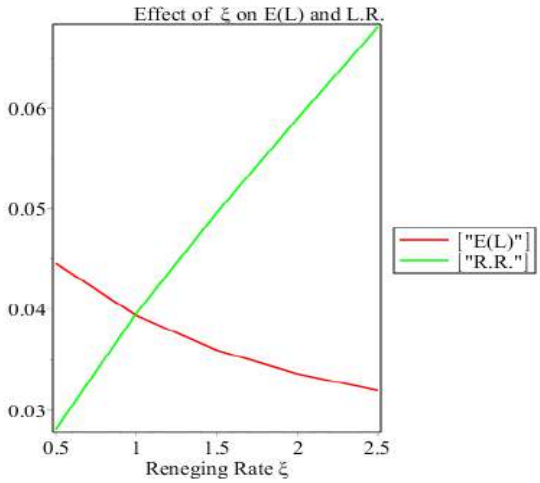


Figure 24: Impact of λ on *E(L)* and *L.R.*

Figures 23 and 24 give the effect of β and λ on mean system length and average loss of customer. As probability of joining the queue increases *E [L]* increases and as reneging rate increases *E[L]* decreases and average rate of customer loss increases

9. Conclusions

This paper considers two-dimensional state Markovian queueing model with Bernoulli Schedule, multiple vacations and impatience customers in which the state of the system is given by (*a, b*). The concept of few clienteles (say "*n*") available in the system makes this model different from the previous models available in literature. The governing systems of equations are solved by using the Laplace transform and different measures of effectiveness (Expected

system length, throughput of system, mean balking rate, mean reneging rate, *etc.*) are calculated that provide better perception of a queueing system. Finally, an expected cost function is discussed, and it shows that if we increase service rate then the probability that customers may balk or renege from the system is reduced which minimize expected cost for the system. Different firms can utilize this model to model their system accordingly and can have an idea about the minimum cost that system will generate.

References

Abou-El-Ata, M. O. (1991). The state-dependent queue: M/M/1/N with reneging and general balk functions. *Microelectronics Reliability*, **31**(**5**), 1001-1007.

Al-Seedy, R. O., El-Sherbiny, A. A., El-Shehawy, S. A. and Ammar, S. I. (2009). Transient solution of the M/M/c queue with balking and reneging. *Computers and Mathematics with Applications*, **57**(**8)**, 1280-1285.

Altman, E. and Yechiali, U. (2006). Analysis of customers' impatience in queues with server vacations. *Queueing Systems*, **52**(**4**), 261-279.

Al-Khedhairi, A. and Tadj, L. (2007). A bulk service queue with a choice of service and re-service under Bernoulli schedule. *International Journal of Contemporary Mathematical Sciences*, **2**(**23**), 1107-1120.

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.

Ancker, C. J. and Gafarian, A. V. (1963). Queuing problems with Balking and Reneging-I. *Operations Research*, **11**, 88-100.

Ancker, C. J. and Gafarian, A. V. (1963). Queuing problems with Balking and Reneging-II. *Operations Research*, **11**, 928-937.

Bacot, J. B. and Dshalalow, J. H. (2001). A bulk input queueing system with batch gated service and multiple vacation policy. *Mathematical and Computer Modelling*, **34(7-8)**, 873-886.

Banik, A. D. (2009). The infinite-buffer single server queue with a variant of multiple vacation policy and batch Markovian arrival process. *Applied Mathematical Modelling*, **33**(**7)**, 3025-3039.

Bouchentouf, A. A. and Messabihi, A. (2018). Heterogeneous two-server queueing system with reverse balking and reneging. *Opsearch*, **55**(**2**), 251-267.

Cooper, R. B. (1970). Queues served in cyclic order waiting times. *Bell Systems*, **49**, 339 – 413.

Choudhury, G., Tadj, L. and Paul, M. (2007). Steady state analysis of an $M^x/G/1$ queue with two phase service and Bernoulli vacation schedule under multiple vacation policy. *Applied Mathematical Modelling*, **31**(**6**), 1079-1091.

Haight, F. A. (1957). Queueing with balking. *Biometrika*, **44**(3/4), 360–369.

Haight, F. A. (1959). Queueing with reneging. *Metrika*, **2**(**1**), 186–197.

Hubbard, J. R., Pegden, C. D. and Rosenshine, M. (1986). The departure process for the M/M/1 queue. *Journal of Applied Probability*, **23**(**1**), 249-255.

Keilson, J. and Servi, L. D. (1987). Dynamics of the M/G/1 vacation model. *Operations Research*, **35**(**4**), 575-582.

Madan, K. C., Abu-Dayyeh, W. and Taiyyan, F. (2003). A two-server queue with Bernoulli schedules and a single vacation policy. *Applied Mathematics and Computation*, **145**(**1**), 59-71.

Pegden, C. D. and Rosenshine, M. (1982). Some new results for the M/M/1 queue. *Management Science*, **28**(**7**), 821-828.

Ramaswamy, R. and Servi, L. D. (1988). The busy period of the migi1 vacation model with a bernoulli schedule. Communications in Statistics. *Stochastic Models*, **4**(**3**), 507-521.

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*, **1531**(**1**), 012060.

Yue, D., Yue, W., Saffer, Z. and Chen, X. (2014). Analysis of an M/M/1 queueing system with impatient customers and a variant of multiple vacation policy. *Journal of Industrial & Management Optimization*, **10**(**1**), 89-112.

Statistics and Application {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 165-179

A General Class of Product-cum-Ratio-Type Exponential Estimators in Double Sampling for Stratification of Finite Population Mean

Housila P. Singh and Pragati Nigam
School of Studies in Statistics, Vikram University, Ujjain, (*M.P.*), *India.*

Received: 16 April 2020; Revised: 28 February 2021; Accepted: 13 March 2021

Abstract

This paper addresses the problem of estimating the population mean of the study variable *y* using information on the auxiliary variable *x* in double sampling for stratification. A general class of product-cum-ratio-type estimators have been defined in this paper. The properties of the suggested class of estimators are studied up to terms of order $O\left(n^{-1}\right)$. Asymptotic optimum estimator (*AOE*) in the class is also identified. In particular, to illustrate the general results, we have suggested a subclass of estimators $\hat{\bar{Y}}_{(1)}$ along with its properties. Preference regions are obtained in which the proposed estimator $\hat{\bar{Y}}_{(1)}$ is better than the existing estimators. In support of the present study, an empirical study is also carried out.

Key words: Auxiliary variable; Bias, Mean squared error; Double sampling for stratification.

Mathematics Subject Classification Code: 62D05.

1. Introduction

It is a well-established fact that the use of auxiliary information in the estimation of population mean provides efficient estimators. Out of many, ratio, product and regression methods of estimation are good examples in this context. A large amount of work has been carried out in estimating the population mean using simple random sampling (*SRS*) with or without replacement (*WOR*) scheme, for instance, see Singh (1986), Singh (2003) among others. Usually, heterogeneous populations are encountered in practice. In such a situation, stratification (or stratified sampling) is extensively used procedure in sample surveys to provide samples that are representatives of major sub-groups of a population. When the sampling frame within strata is known, stratified sampling is used, but there are many situations of practical importance where strata weights are known but a frame within the strata is not available; post-stratification may then be employed to cope with this problem. For example, in household survey in a city, number of households in different colonies may be available, but list of households may not be available. In such a situation post-stratification is used. However, in other situations with the passage of time, the stratum weights may not be known exactly as they become out-of-date. Further, the information on the stratification variable may not be readily available but could be made available by diverting a part of the survey budget to its collection. This type of situation occurs during the household surveys, when the investigator does not have information about newly added households in different colonies, see Tailor *et al*.

Corresponding Author: Pragati Nigam
Email: pragatinigam1@gmail.com

(2014). In such a situation we employ the procedure of double sampling for stratification (*DSS*) introduced by Neyman (1938). Double sampling for stratification is a sampling design that is extensively employed in forest and other resource inventories in forest ecosystems. Double sampling is a powerful and cost-effective procedure. For more studies on this topic the reader is referred to the papers by Rao (1973), Ige and Tripathi (1987, 1991), Singh and Vishwakarma (2007), Vishwakarma and Singh (2012), Tailor and Lone (2014), Vishwakarma and Zeeshan (2018) and Singh and Nigam (2020 a, b).

2. Procedure of Selecting a Sample and Terminologies

Consider a finite population $U=\{U_1,U_2,...U_N\}$ of N units. Let y and x be study variate and auxiliary variate, respectively. Let $\bar{Y}$ be the population mean of the study variate y. Suppose we want to estimate the population mean $\bar{Y}$ of y and consider it desirable to stratify the population on the basis of the values of ran auxiliary variate x but the frequency distribution of x is unknown. Let the population of size N be stratified into L strata of size N_h with strata weights $W_h=\frac{N_h}{N}$, (h = 1,2,…,L). The sampling frame for different strata and the strata weights $W_h=\frac{N_h}{N}$, h = 1,2,…,L are not known although the strata may be fixed in advance. Under these circumstances we employ the procedure of double sampling for stratification (*DSS*). It consists of the following steps [see Rao (1973) and Ige and Tripathi (1987)]:

(i) We select first phase sample S' of size n' using simple random sampling without replacement (*SRSWOR*) and measure only auxiliary variate x.

(ii) The first phase sample S' is stratified into L strata based on measured x-values. Let n'_h be the number of units in S' falling into stratum h (h=1,2,…,L; $n'=\sum_{h=1}^{L}n'_h$) and $n'=\{n'_1,n'_2,...,n'_L\}$ denote the resulting configuration of S'.

(iii) Sub-sample of sizes $n_h=\upsilon_h n'_h, 0<\upsilon_h<1, h=1,2,...,L$, υ_h being predetermined for all h, are drawn from strata, independently from each other, using *SRSWOR*. Thus, it constitutes a second phase sample S of size $n=\sum_{h=1}^{L}n_h$. The study variable y is measured on all n_h sampled unites, for all h.

We use the following notations:

$w_h=\frac{n'_h}{n'}$:is an unbiased estimator of strata weights $W_h=\frac{N_h}{N}$(or proportion of first sample falling in stratum h), see Cochran (1977, p.328),

$\bar{Y}=\sum_{h=1}^{L}W_h\bar{Y}_h$ is the population mean of the study variable y,

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

$S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^2$ is the population mean square of y,

$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$ is the population mean square of x,

$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)$ is the population covariance between y and x,

$S_{yh}^2 = \frac{1}{N_h-1}\sum_{j=1}^{N_h}\left(y_{hj} - \bar{Y}_h\right)^2$ is the population mean square of y of h^{th} stratum,

$S_{xh}^2 = \frac{1}{N_h-1}\sum_{j=1}^{N_h}\left(x_{hj} - \bar{X}_h\right)^2$ is the population mean square of x of h^{th} stratum,

$\rho_{yxh} = \frac{S_{yxh}}{S_{yh}S_{xh}}$ is the population correlation coefficient between y and x in the h^{th} stratum,

$\bar{y}_{ds} = \sum_{h=1}^{L} w_h \bar{y}_h$ is an unbiased estimator of the population mean $\bar{Y}$,

$\bar{x}_{ds} = \sum_{h=1}^{L} w_h \bar{x}_h$ is an unbiased estimator of the population mean $\bar{X}$,

$\bar{x}' = \sum_{h=1}^{L} w_h \bar{x}'_h$ is an unbiased estimator of the population mean $\bar{X}$,

$\bar{y}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj}$ is the mean of the second phase taken from h^{th} stratum for y,

$\bar{x}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} x_{hj}$ is the mean of the second phase taken from h^{th} stratum for x,

$\bar{x}'_h = \frac{1}{n'_h}\sum_{j=1}^{n'_h} x_{hj}$ is the mean of the first phase sample of h^{th} stratum for x,

$f = \frac{n'}{N}$ is the first phase sampling fraction,

$\beta_h = \frac{S_{yxh}}{S_{xh}^2}, a_h = \left(\frac{1}{\upsilon_h} - 1\right)W_h S_{xh}^2, \beta = \frac{\sum_{h=1}^{L} a_h \beta_h}{\sum_{h=1}^{L} a_h} = \frac{A_{yx}}{A_x}$ is the weighted average of the strata population regression coefficient, $R = \frac{\bar{Y}}{\bar{X}}$,

$$A_y = \sum_{h=1}^{L} W_h S_{yh}^2 \left(\frac{1}{\upsilon_h} - 1\right),$$

$$A_x = \sum_{h=1}^{L} W_h S_{xh}^2 \left(\frac{1}{\upsilon_h} - 1\right),$$

$$A_{yx} = \sum_{h=1}^{L} W_h S_{yxh} \left(\frac{1}{\upsilon_h} - 1\right) \text{ and } \rho = \frac{A_{yx}}{\sqrt{A_y A_x}}.$$

To obtain the bias and mean squared error (*MSE*) of the suggested estimator, we write

$$\bar{y}_{ds} = \bar{Y}(1+e_0), \bar{x}_{ds} = \bar{X}(1+e_1) \text{and } \bar{x}' = \bar{X}(1+e'_1)$$

such that

$$E(e_0) = E(e_1) = E(e'_1) = 0$$

and

$$E(e_0^2) = \frac{1}{\bar{Y}^2}\left[S_y^2\left(\frac{1-f}{n'}\right) + \frac{1}{n'}A_y\right],$$

$$E(e_1^2) = \frac{1}{\bar{X}^2}\left[S_x^2\left(\frac{1-f}{n'}\right) + \frac{1}{n'}A_x\right],$$

$$E(e_1'^2) = \left(\frac{1-f}{n'}\right)\frac{1}{\bar{X}^2}S_x^2,$$

$$E(e_0 e_1) = \frac{1}{\bar{Y}\,\bar{X}}\left[S_{yx}\left(\frac{1-f}{n'}\right) + \frac{1}{n'}A_{yx}\right],$$

$$E(e_0 e'_1) = \left(\frac{1-f}{n'}\right)\frac{1}{\bar{Y}\,\bar{X}}S_{yx}, \text{ and}$$

$$E(e_1 e'_1) = \left(\frac{1-f}{n'}\right)\frac{1}{\bar{X}^2}S_x^2.$$

3. Review of Some Existing Estimators

The usual unbiased estimator for population mean $\bar{Y}$ is defined by

$$\bar{y}_{ds} = \sum_{h=1}^{L} w_h \bar{y}_h . \tag{1}$$

The variance/*MSE* of $\bar{y}_{ds}$ is given by

$$V(\bar{y}_{ds}) = MSE(\bar{y}_{ds}) = \left(\frac{1-f}{n'}\right)S_y^2 + \frac{1}{n'}A_y . \tag{2}$$

In double sampling for stratification, ratio and product estimators due to Ige and Tripathi (1987) are respectively given by

$$\hat{\bar{Y}}_{R(ds)} = \bar{y}_{ds}\left(\frac{\bar{x}'}{\bar{x}_{ds}}\right), \tag{3}$$

$$\hat{\bar{Y}}_{P(ds)} = \bar{y}_{ds}\left(\frac{\bar{x}_{ds}}{\bar{x}'}\right). \tag{4}$$

The *MSE*s of $\hat{\bar{Y}}_{R(ds)}$ and $\hat{\bar{Y}}_{P(ds)}$ up to terms of order $O(n^{-1})$, are given respectively by

$$MSE\left(\hat{\bar{Y}}_{R(ds)}\right) = \left(\frac{1-f}{n'}\right)S_y^2 + \frac{1}{n'}\left\{A_y + R^2 A_x\left(1 - \frac{2\beta}{R}\right)\right\}, \tag{5}$$

$$MSE\left(\hat{\bar{Y}}_{P(ds)}\right)=\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+R^2A_x\left(1+\frac{2\beta}{R}\right)\right\}, \tag{6}$$

where $\upsilon_h=\frac{n_h}{n'_h}$.

Motivated by Bahl and Tuteja (1991), Tailor *et al*. (2014) suggested the ratio-type and product-type exponential estimators in double sampling for stratification for population mean $\bar{Y}$ respectively as

$$\hat{\bar{Y}}_{\mathrm{Re}(ds)}=\bar{y}_{ds}\exp\left\{\frac{\bar{x}'-\bar{x}_{ds}}{\bar{x}'+\bar{x}_{ds}}\right\}, \tag{7}$$

$$\hat{\bar{Y}}_{Pe(ds)}=\bar{y}_{ds}\exp\left\{\frac{\bar{x}_{ds}-\bar{x}'}{\bar{x}'+\bar{x}_{ds}}\right\}. \tag{8}$$

The *MSE*s of $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ up to terms of order $O\left(n^{-1}\right)$, are respectively given by

$$MSE\left(\hat{\bar{Y}}_{\mathrm{Re}(ds)}\right)=\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+\frac{R^2A_x}{4}\left(1-\frac{4\beta}{R}\right)\right\}, \tag{9}$$

$$MSE\left(\hat{\bar{Y}}_{Pe(ds)}\right)=\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+\frac{R^2A_x}{4}\left(1+\frac{4\beta}{R}\right)\right\}. \tag{10}$$

From (2), (5), (6), (9) and (10) it can be observed that the:

(i) ratio estimator $\hat{\bar{Y}}_{R(ds)}$ is more efficient than the unbiased estimator $\bar{y}_{(ds)}$ if

$$\frac{\beta}{R}>\frac{1}{2} \tag{11}$$

(ii) product estimator $\hat{\bar{Y}}_{P(ds)}$ is better than the unbiased estimator $\bar{y}_{(ds)}$ if

$$\frac{\beta}{R}<-\frac{1}{2} \tag{12}$$

(iii) ratio-type exponential estimator $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ is superior to the unbiased estimator $\bar{y}_{(ds)}$ if

$$\frac{\beta}{R}>\frac{1}{4} \tag{13}$$

(iv) product-type exponential estimator $\hat{\bar{Y}}_{Pe(ds)}$ is more precise than the unbiased estimator $\bar{y}_{(ds)}$ if

$$\frac{\beta}{R} < -\frac{1}{4} \tag{14}$$

In this paper we have suggested a general class of product-cum-ratio-type exponential estimators for population mean $\bar{Y}$ of y using double sampling for stratification based on auxiliary information. Expressions of bias and *MSE* of the proposed estimator up to $O\left(n^{-1}\right)$ are derived. Asymptotic optimum estimator (*AOE*) in the class is identified with its approximate *MSE* formula. To illustrate the general results we have considered a subclass of estimators $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ along with its properties up to terms of order $O\left(n^{-1}\right)$. An empirical study is carried out in support of the present study.

4. The Suggested Class of Product-Cum-Ratio-Type Exponential Estimators

We define a general class of product-cum-ratio-type exponential estimators for population mean $\bar{Y}$ in double sampling for stratification based on auxiliary information, as

$$\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} = \bar{y}_{ds}\left[\delta\left(\frac{a\bar{x}_{ds}+b}{a\bar{x}'+b}\right)^{\eta} + (1-\delta)\exp\left\{\frac{\phi a\left(\bar{x}'-\bar{x}_{ds}\right)}{a\left(\bar{x}'+\bar{x}_{ds}\right)+2b}\right\}\right], \tag{15}$$

where δ is a suitable chosen constant, (η,ϕ) are scalars taking values (-1,0,1) for generating ratio and product-type estimators, $a\,(\neq 0)$ and b are either real numbers or functions of known parameters of the auxiliary variable x like coefficient of variation C_x, standard deviation S_x, coefficient of skewness $\beta_1(x)$, coefficient of kurtosis $\beta_2(x)$, correlation coefficient ρ_{yx} between y and x; and $\Delta(x)=\left(\beta_2(x)-\beta_1(x)-1\right)$. One may also take the values of a and b as $\varphi_1=\sum_{h=1}^{L}W_hS_{xh}, \varphi_2=\sum_{h=1}^{L}W_hC_{xh}, \varphi_3=\sum_{h=1}^{L}W_h\beta_{1h}(x), \varphi_4=\sum_{h=1}^{L}W_h\beta_{2h}(x), \varphi_5=\sum_{h=1}^{L}W_h\rho_{yxh}$ and $\varphi_6=\sum_{h=1}^{L}W_h\Delta_h(x)$ with $\Delta_h(x)=\left(\beta_{2h}(x)-\beta_{1h}(x)-1\right)$, for instance, see Koyuncu and Kadilar (2009, p.2553).

A large number of estimators can be generated from the proposed estimator $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ for suitable values of (δ,η,a,b). For example:

(i) $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} \to \bar{y}_{ds}$ for $(a,b,\delta,\eta,\phi)=(a,b,\delta,0,0)$,

(ii) $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} \to \hat{\bar{Y}}_{R(ds)}$ for $(a,b,\delta,\eta,\phi)=(a,0,1,-1,\phi)$,

(iii) $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} \to \hat{\bar{Y}}_{P(ds)}$ for $(a,b,\delta,\eta,\phi)=(a,0,1,1,\phi)$,

(iv) $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} \to \hat{\bar{Y}}_{\mathrm{Re}(ds)}$ for $(a,b,\delta,\eta,\phi)=(a,0,0,\eta,1)$,

(v) $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} \to \hat{\bar{Y}}_{Pe(ds)}$ for $(a,b,\delta,\eta,\phi)=(a,0,0,\eta,-1)$,

etc.

4.1. The bias and *MSE* of the proposed class of estimators $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$

Expressing $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ at (15) in terms of e's we have

$$\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} = \bar{Y}(1+e_0)\left[\delta\left(\frac{a\bar{X}(1+e_1)+b}{a\bar{X}(1+e'_1)+b}\right)^{\eta} + (1-\delta)\exp\left\{\frac{\phi a\bar{X}(e'_1-e_1)}{a\bar{X}(2+e'_1+e_1)+2b}\right\}\right]$$

$$= \bar{Y}(1+e_0)\left[\delta(1+\xi e_1)^{\eta}(1+\xi e'_1)^{-\eta} + (1-\delta)\exp\left\{\frac{\xi\phi}{2}(e'_1-e_1)\left(1+\frac{\xi}{2}(e'_1+e_1)\right)^{-1}\right\}\right], \quad (16)$$

where the $\xi = \dfrac{a\bar{X}}{(a\bar{X}+b)}$.

Expanding right hand side of (16), multiplying out, subtracting $\bar{Y}$ from both sides and neglecting terms of e's having power greater than two we have

$$\left(\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)} - \bar{Y}\right) \cong \bar{Y}\left[\begin{array}{l} e_0 + \dfrac{1}{2}\{2\eta\delta - (1-\delta)\phi\}\xi(e_1 - e'_1 + e_0e_1 - e_0e'_1) + \dfrac{\xi^2}{8}\{4\delta\eta(\eta-1) + (1-\delta)\phi(\phi+2)\}e_1^2 \\ + \dfrac{\xi^2}{8}\{4\delta\eta(\eta+1) + (1-\delta)\phi(\phi-2)\}e'^2_1 - \dfrac{\xi^2}{4}\{4\delta\eta^2 + (1-\delta)\phi^2\}e_1e'_1 \end{array}\right] \quad (17)$$

Taking expectation of both sides of (17) we get the bias of $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ up to terms of order $O(n^{-1})$ as

$$B\left(\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}\right) = \frac{\xi}{8n'\bar{X}}\left[4\{2\delta\eta - (1-\delta)\phi\}A_{yx} + \xi R\{4\delta\eta(\eta-1) + (1-\delta)\phi(\phi+2)\}A_x\right]. \quad (18)$$

Squaring both sides of (17), neglecting terms of e's having power greater than two and then taking expectation of both sides we get the *MSE* of $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ up to terms of order $O(n^{-1})$ as

$$MSE\left(\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}\right) = \left[\frac{(1-f)}{n'}S_y^2 + \frac{1}{n'}\left\{A_y + \frac{\theta R^2\xi^2 A_x}{4}\left(\theta - \frac{4\beta}{R\xi}\right)\right\}\right], \quad (19)$$

which is minimized for

$$\theta = \frac{2\beta}{R\xi} = \theta_0(\text{say}), \quad (20)$$

where $\theta = \{(1-\delta)\phi - 2\delta\eta\}$.

Thus the resulting minimum *MSE* of $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ up to terms of order $O(n^{-1})$ is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}\right)=\left[\frac{(1-f)}{n'}S_y^2+\frac{A_y}{n'}\left(1-\rho^2\right)\right]. \tag{21}$$

Now we state the following theorem.

Theorem 1: Up to terms of order $O\left(n^{-1}\right)$,

$$MSE\left(\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}\right)\geq\left[\frac{(1-f)}{n'}S_y^2+\frac{A_y}{n'}\left(1-\rho^2\right)\right]$$

with equality holding if

$$\theta=\frac{2\beta}{\xi R}.$$

To illustrate the general results of the class of estimators $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$, we consider the following sub class of estimators for $\bar{Y}$ as

$$\hat{\bar{Y}}_{(1)}=\bar{y}_{ds}\left[\delta\left(\frac{a\bar{x}_{ds}+b}{a\bar{x}'+b}\right)+(1-\delta)\exp\left\{\frac{a\left(\bar{x}'-\bar{x}_{ds}\right)}{a\left(\bar{x}'+\bar{x}_{ds}\right)+2b}\right\}\right] \tag{22}$$

which is obtained on putting $(\eta,\phi)=(1,1)$ in (15). We designate the estimator $\hat{\bar{Y}}_{(1)}$ as 'product-cum-ratio-type exponential' estimator.

Inserting $(\eta,\phi)=(1,1)\Rightarrow\theta=(1-3\delta)$ in (18) and (19) we get the bias and *MSE* of $\hat{\bar{Y}}_{(1)}$ up to terms of order $O\left(n^{-1}\right)$ respectively as

$$B\left(\hat{\bar{Y}}_{(1)}\right)=\frac{\xi^2RA_x}{2n'\bar{X}}\left[\frac{(3\delta-1)\beta}{\xi R}+\frac{3}{4}(1-\delta)\right], \tag{23}$$

$$MSE\left(\hat{\bar{Y}}_{(1)}\right)=\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]. \tag{24}$$

The $MSE\left(\hat{\bar{Y}}_{(1)}\right)$ at (24) is minimum when

$$\delta=\frac{1}{3}\left(1-\frac{2\beta}{R\xi}\right)=\delta_{(o)}\,(\text{say}). \tag{25}$$

Substitution of (25) in (24) yields the minimum *MSE* of $\hat{\bar{Y}}_{(1)}$ as

$$MSE_{\min}\left(\hat{\bar{Y}}_{(1)}\right)=\left[\frac{(1-f)}{n'}S_y^2+\frac{A_y}{n'}\left(1-\rho^2\right)\right]. \tag{26}$$

Now we give the conditions under which the proposed estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$.

From (2), (5), (6), (9), (10) and (24) it can be easily shown that the proposed product-cum-ratio-type exponential estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than:

(i) the usual unbiased estimator $\bar{y}_{ds}$ if

$$\min\left\{\frac{1}{3},-\frac{1}{3}\left(\frac{4\beta}{R\xi}-1\right)\right\}<\delta<\max\left\{\frac{1}{3},-\frac{1}{3}\left(\frac{4\beta}{R\xi}-1\right)\right\}. \tag{27}$$

(ii) the ratio estimator $\hat{\bar{Y}}_{R(ds)}$ if

$$\min\left\{\frac{1}{3}\left(1-\frac{2}{\xi}\left(\frac{2\beta}{R}-1\right)\right),\frac{1}{3}\left(1-\frac{2}{\xi}\right)\right\}<\delta<\max\left\{\frac{1}{3}\left(1-\frac{2}{\xi}\left(\frac{2\beta}{R}-1\right)\right),\frac{1}{3}\left(1-\frac{2}{\xi}\right)\right\}. \tag{28}$$

(iii) the product estimator $\hat{\bar{Y}}_{P(ds)}$ if

$$\min\left\{\frac{1}{3}\left(1-\frac{2}{\xi}\left(\frac{2\beta}{R}+1\right)\right),\frac{1}{3}\left(1+\frac{2}{\xi}\right)\right\}<\delta<\max\left\{\frac{1}{3}\left(1-\frac{2}{\xi}\left(\frac{2\beta}{R}+1\right)\right),\frac{1}{3}\left(1+\frac{2}{\xi}\right)\right\}. \tag{29}$$

(iv) the ratio-type exponential estimator $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ if

$$\min\left\{\frac{1}{3}\left(1-\frac{1}{\xi}\right),\frac{1}{3}\left(\frac{1}{\xi}\left(1-\frac{4\beta}{R}\right)+1\right)\right\}<\delta<\max\left\{\frac{1}{3}\left(1-\frac{1}{\xi}\right),\frac{1}{3}\left(\frac{1}{\xi}\left(1-\frac{4\beta}{R}\right)+1\right)\right\}. \tag{30}$$

(v) the product-type exponential estimator $\hat{\bar{Y}}_{Pe(ds)}$ if

$$\min\left\{\frac{1}{3}\left(1-\frac{1}{\xi}\left(\frac{4\beta}{R}+1\right)\right),\frac{1}{3}\left(\frac{1}{\xi}+1\right)\right\}<\delta<\max\left\{\frac{1}{3}\left(1-\frac{1}{\xi}\left(\frac{4\beta}{R}+1\right)\right),\frac{1}{3}\left(\frac{1}{\xi}+1\right)\right\}. \tag{31}$$

5. Numerical Illustration

To illustrate the performance of the suggested estimator $\hat{\bar{Y}}_{(1)}$ over other existing estimators, we have considered three data sets whose descriptions are given below.

Data 1 [Source: Tailor *et al*. (2014)]
x: Area in '000 Hectare, *y*: Productivity (MT/Hectare)
N=20, n=8, $n_1=4, n_2=4, n'_1=7, n'_2=7, N_1=10, N_2=10, \bar{Y}_1=142.80, \bar{Y}_2=102.60,$
$\bar{X}_1=1632.00, \bar{X}_2=2036.00, S_{x1}=102.17, S_{x2}=103.46, S_{y1}=6.09, S_{y2}=12.60,$
$S_{yx1}=-239.30, S_{yx2}=-655.30, S_y^2=528.43.$

Data 2 [Source: Chouhan (2012)]
x: Area in '000 Hectare, *y*: Productivity (MT/Hectare)
N=20, n=8, $n_1=4, n_2=4, n'_1=7, n'_2=7, N_1=10, N_2=10, \bar{Y}_1=1.70, \bar{Y}_2=3.67, \bar{X}_1=6.32,$
$\bar{X}_2=8.67, S_{x1}=1.19, S_{x2}=10.82, S_{y1}=0.50, S_{y2}=1.41, S_{yx1}=-0.05, S_{yx2}=-7.04, S_y^2=2.20.$

Data 3 [Source: Murthy (1967), p228]
x: Fixed capital, *y*: Output
N=10, n=4, $n_1=2, n_2=2, n'_1=4, n'_2=4, N_1=5, N_2=5, \bar{Y}_1=1925.8, \bar{Y}_2=3115.6,$
$\bar{X}_1=214.4, \bar{X}_2=333.8, S_{x1}=74.87, S_{x2}=66.35, S_{y1}=615.92, S_{y2}=340.38,$
$S_{yx1}=39360.68, S_{yx2}=22356.50, S_y^2=668351.00.$

We have computed the percent relative efficiencies (*PRE*s) of the suggested estimator $\hat{\bar{Y}}_{(1)}$ with respect to $\bar{y}_{(ds)}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ by using the following formulae:

$$PRE\left(\hat{\bar{Y}}_{(1)}, \bar{y}_{ds}\right)=\frac{\left[\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}A_y\right]}{\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]}*100, \tag{32}$$

$$PRE\left(\hat{\bar{Y}}_{(1)}, \hat{\bar{Y}}_{R(ds)}\right)=\frac{\left[\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+R^2A_x\left(1-\frac{2\beta}{R}\right)\right\}\right]}{\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]}*100, \tag{33}$$

$$PRE\left(\hat{\bar{Y}}_{(1)}, \hat{\bar{Y}}_{P(ds)}\right)=\frac{\left[\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+R^2A_x\left(1+\frac{2\beta}{R}\right)\right\}\right]}{\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]}*100, \tag{34}$$

$$PRE\left(\hat{\bar{Y}}_{(1)}, \hat{\bar{Y}}_{\mathrm{Re}(ds)}\right)=\frac{\left[\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+\frac{R^2A_x}{4}\left(1-\frac{4\beta}{R}\right)\right\}\right]}{\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]}*100, \tag{35}$$

$$PRE\left(\hat{\bar{Y}}_{(1)},\hat{\bar{Y}}_{Pe(ds)}\right)=\frac{\left[\left(\frac{1-f}{n'}\right)S_y^2+\frac{1}{n'}\left\{A_y+\frac{R^2A_x}{4}\left(1+\frac{4\beta}{R}\right)\right\}\right]}{\left[\frac{(1-f)}{n'}S_y^2+\frac{1}{n'}\left\{A_y+\frac{(3\delta-1)R^2\xi^2A_x}{4}\left((3\delta-1)+\frac{4\beta}{R\xi}\right)\right\}\right]}*100. \quad (36)$$

Here we note that for the sake of simplicity we have taken (a, b) = (1, 0) for computing the *PRE* in the suggested estimator $\hat{\bar{Y}}_{(1)}$. Also for computing the range of δ we have taken (a, b) = (1, 0) in equation (27)-(31).

Table 1: Different ranges of δ in which the proposed estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than $\bar{y}_{(ds)}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$

Estimators	Data 1		Data 2		Data 3	
	δ (min)	δ (max)	δ (min)	δ (max)	δ (min)	δ (max)
$\bar{y}_{ds}$	0.33333	1.17659	0.33333	0.55604	0.33333	–0.5608
$\hat{\bar{Y}}_{R(ds)}$	–0.33333	1.84326	–0.33333	1.22271	–0.33333	0.10589
$\hat{\bar{Y}}_{P(ds)}$	0.50993	1	–0.11063	1	–1.22745	1
$\hat{\bar{Y}}_{RExp(ds)}$	0	1.50993	0	0.88937	–0.2274	0
$\hat{\bar{Y}}_{PExp(ds)}$	0.84326	0.66667	0.22271	0.66667	–0.89411	0.66667
$\delta_{(opt)}$	0.75946		0.44469		–0.1137	

Table 2: *PRE* of $\hat{\bar{Y}}_{(1)}$ with respect to $\bar{y}_{ds}$ at different values of δ

Data 1		Data 2		Data 3	
δ	***PRE***	δ	***PRE***	δ	***PRE***
0.33333	100	0.333333	100	–0.56	100.13
0.4	101.81	0.4	109.77	–0.5	110.33
0.5	104.04	0.444	111.87	–0.25	150.42
0.6	105.59	0.55604	100	–0.11372	158.62
0.7	106.4	0.60	90.88	0	152.82
0.75	106.52	-	-	0.25	114.28
0.8	106.44	-	-	0.3	105.6
0.9	105.7	-	-	0.33333	100
1	104.22	-	-	0.40	89.41
1.176595	100	-	-	-	-

Table 3: *PRE* of $\hat{\bar{Y}}_{(1)}$ with respect to $\hat{\bar{Y}}_{R(ds)}$ at different values of δ

Data 1		Data 2		Data 3	
δ	***PRE***	δ	***PRE***	δ	***PRE***
–0.3333	100	–0.33333	100	–0.33333	100
-0.25	104.67	-0.25	120.91	-0.3	103.6
0	118.63	0	234.87	–0.25	108.25
0.25	131.16	0.25	498.58	–0.2	111.71
0.5	140.09	0.5	660.16	–0.1	114.08
0.75	143.42	0.75	359.07	0	109.97
1	140.34	1	171.93	0.10589	100
1.25	131.6	1.22271	100	0.15	94.81
1.5	119.17	1.25	94.26	-	-
1.75	105.23	-	-	-	-
1.8432	100	-	-	-	-

Table 4: *PRE* of $\hat{\bar{Y}}_{(1)}$ with respect to $\hat{\bar{Y}}_{P(ds)}$ at different values of δ

Data 1		Data 2		Data 3	
δ	***PRE***	δ	***PRE***	δ	***PRE***
0.50993	100	–0.11062	100	–1.22744	100
0.6	101.31	0	136.61	–1	140.38
0.7	102.09	0.25	289.99	–0.75	212.03
0.8	102.13	0.5	383.97	–0.5	322.61
0.9	101.42	0.75	208.85	–0.25	439.84
1	100	1	100	0	446.85
1.25	93.77	1.25	54.83	0.25	334.14
-	-	-	-	0.5	220.36
-	-	-	-	0.75	145.48
-	-	-	-	1	100

Table 5: *PRE* of $\hat{\bar{Y}}_{(1)}$ with respect to $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ at different values of δ

Data 1		Data 2		Data 3	
δ	***PRE***	δ	***PRE***	δ	***PRE***
0	100	0	100	–0.2274	100
0.25	110.56	0.1	135.36	–0.2	101.58
0.5	118.08	0.2	183.9	–0.1	103.74
0.75	120.9	0.25	212.28	0	100
1	118.29	0.3	241.01	0.1	91.53
1.25	110.93	0.5	281.08	-	-
1.5099	100	0.6	235.03	-	-
1.65	93.45	0.7	178.14	-	-
-	-	0.88937	100	-	-

Table 6: *PRE* of $\hat{\bar{Y}}_{(1)}$ with respect to $\hat{\bar{Y}}_{Pe(ds)}$ at different values of δ

Data 1		Data 2		Data 3	
δ	***PRE***	δ	***PRE***	δ	***PRE***
0.66667	100	0.22271	100	–0.89411	100
0.7	100.17	0.25	107.99	–0.75	127.37
0.75	100.28	0.3	122.6	–0.5	193.81
0.8	100.21	0.4	144.41	–0.25	264.23
0.84326	100	0.5	142.98	0	268.44
0.90	99.52	0.6	119.56	0.25	200.73
-	-	0.66666	100	0.5	132.38
-	-	0.70	90.62	0.66666	100

The optimum value of δ and the range of δ in which the proposed estimator $\hat{\bar{Y}}_{(1)}$ dominates over the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ are displayed in Table 1.

The percent relative efficiencies (*PRE*s) of the suggested estimator $\hat{\bar{Y}}_{(1)}$ with respect to $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ for varying values of δ are presented in Tables 2 to 6, respectively.

It is observed from Tables 2 to 6 that the proposed estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ in certain range of δ as given in Table 1. Further, we observed that there is largest gain in efficiency by using the estimator $\hat{\bar{Y}}_{(1)}$ over the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ for all the data sets 1, 2 and 3 at the optimum value of δ.

Table 2 exhibits that the proposed estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than the usual unbiased estimator $\bar{y}_{ds}$ for a wider range of δ for data sets 1 and 3 while it is better than $\bar{y}_{ds}$ for a shorter range of δ. For data set 3, the gain in efficiency is also substantial by using the estimator $\hat{\bar{Y}}_{(1)}$ over $\bar{y}_{ds}$.

It is observed from Table 3 that the estimator $\hat{\bar{Y}}_{(1)}$ gives the largest amount of gain in efficiency over ratio estimator $\hat{\bar{Y}}_{R(ds)}$ for a broad range of δ for the data sets 1 and 2. While it is marginal for a smaller range of δ for the data set 3.

Table 4 presents that the estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than the product estimator $\hat{\bar{Y}}_{P(ds)}$ for a broader range of δ with substantial gain in efficiency for data set 3. It also presents

considerable gain in efficiency by using $\hat{\bar{Y}}_{(1)}$ over the product estimator $\hat{\bar{Y}}_{P(ds)}$ but for a smaller range of δ for data set 2. The gain in efficiency is marginal for data set 1.

Table 5 demonstrate that the suggested estimator $\hat{\bar{Y}}_{(1)}$ is more efficient than ratio-type exponential estimator $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ with substantial gain in efficiency for a wide range of δ for data sets 1 and 2, while it is marginal for data set 3.

It is observed from Table 6 that there is sizable gain in efficiency by using the estimator $\hat{\bar{Y}}_{(1)}$ over the product-type exponential estimator $\hat{\bar{Y}}_{Pe(ds)}$ for a wide range of δ for data sets 2 and 3 while it is marginal in case of population data set 1.

Finally, we conclude from Tables 1 to 6 that even if the scalar δ deviates from its "exact optimum value" the gain in efficiency by using $\hat{\bar{Y}}_{(1)}$ over $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(ds)}$, $\hat{\bar{Y}}_{P(ds)}$, $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and $\hat{\bar{Y}}_{Pe(ds)}$ is considerable. Also, there is enough scope of selecting the values of δ for obtaining better estimators than existing estimators. Thus, we recommend the use of the proposed estimator $\hat{\bar{Y}}_{(1)}$ in practice.

6. Discussion

In this article, we have discussed the problem of estimating the population mean using auxiliary information in double sampling for stratification. A class of product-cum-ratio-type estimators $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ have been developed. Expressions of bias and mean squared error of the developed estimator $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ have been derived up to the first order of approximation. Optimum condition is derived at which the mean squared error of the proposed estimator $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ is minimized. In particular, to demonstrate the utility of the general results, we consider a subclass of the developed estimator $\hat{\bar{Y}}^{(c)}_{P\mathrm{Re}(ds)}$ named as 'product-cum-ratio-type exponential' estimator $\hat{\bar{Y}}_{(1)}$. Properties of the subclass of estimators $\hat{\bar{Y}}_{(1)}$ have been studied. Regions of preferences have been derived in which the suggested subclass of estimators $\hat{\bar{Y}}_{(1)}$ is more efficient than the usual unbiased estimator $\bar{y}_{ds}$, Ige and Tripathi (1987) ratio estimator $\hat{\bar{Y}}_{R(ds)}$ and product estimator $\hat{\bar{Y}}_{P(ds)}$, Tailor *et al.* (2014) ratio-type exponential estimator $\hat{\bar{Y}}_{\mathrm{Re}(ds)}$ and product-type exponential estimator $\hat{\bar{Y}}_{Pe(ds)}$. We have also carried out an empirical study to demonstrate the performance of the proposed estimator $\hat{\bar{Y}}_{(1)}$ over other existing estimators.

Acknowledgement

The authors would like to express their gratitude to the reviewer for his valuable suggestions that led to considerable improvement in the presentation of the contents. We are

also thankful to Professor Vinod Kumar Gupta, Chair Editor for his careful reading the paper and forwarding valuable suggestions regarding the improvement of the paper.

References

Bahl, S. and Tuteja, R. K. (1991). Ratio and product-type exponential estimator. *Journal of Information and Optimization Sciences*, **12**(**1**), 159-164.

Chouhan, S. (2012). *Improved Estimation of Parameters Using Auxiliary Information in Sample Surveys.* Unpublished Ph.D. Thesis, Vikram University, Ujjain, M.P. India.

Cochran, W. G. (1977). Sampling techniques, third edition, Wiley Eastern Limited, New Delhi, India.

Ige, A. F. and Tripathi, T. P. (1987). On doubling for stratification and use of auxiliary information. *Journal of the Indian Society of Agricultural Statistics*, **39**, 191-201.

Koyuncu, N. and Kadilar, C. (2009). Ratio and product estimators in stratified random sampling. *Journal of Statistical Planning and Inference*, **139**(**8**), 2552-2558.

Murthy, M. N. (1967). *Sampling Theory and Methods*. *Statistical Publishing Society, Calcutta, India*, 228.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **37**, 101-116.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **6**(**1**), 125-133.

Singh, H. P. (1986). A generalized class of estimators of ratio, product and mean using supplementary information on an auxiliary character in *PPSWR* sampling scheme. *Gujarat Statistical Review*, **13(2**), 1-30.

Singh, H. P. and Nigam, P. (2020a). Product-product-type (*PPT*) exponential estimator in double sampling for stratification (*DSS*) of finite population mean. *International Journal of Mathematics and Statistics*, **21**(**3**), 50-59.

Singh, H. P. and Nigam, P. (2020b). Ratio-ratio-type exponential estimator of finite population mean in double sampling for stratification. *International Journal of Agricultural and Statistical Sciences*, **16(1**), 251-257.

Singh, H. P. and Vishwakarma, G. K. (2007). A general procedure for estimating the mean using double sampling for stratification. *Model Assisted Statistics and Applications*, **2**(**4**), 225-237.

Singh, S. (2003). *Advance Sampling Theory and Applications-How Michael 'Selected' Amy*. Vol. I, Springer Sciences and Business Media.

Tailor, R. and Lone, H. A. (2014). Dual to separate ratio type exponential estimator in post-stratification. *Journal of Statistics Application and Probability*, **3**(**3**), 425-432.

Tailor, R., Chouhan, S. and Kim, J. M. (2014). Ratio and product type exponential estimators of population mean in double sampling for stratification. *Communications for Statistical Applications and Methods*, **21**(**1**), 1-9.

Vishwakarma, G. K. and Singh, H. P. (2012). A general procedure for estimating the mean using double sampling for stratification and multi-auxiliary information. *Journal of Statistical Planning and Inference*, **142**, 1252-1261.

Vishwakarma, G. K. and Zeeshan, S. M. (2018). A ratio-cum-product type exponential estimator using double sampling for stratification. *Romanian Statistical Review*, **2**, 3-13.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 181-192

Methods of Stratification for Allocation Proportional to Stratum Total Under a Superpopulation Model

Bhuwaneshwar Kumar Gupt[1] **and Md. Irphan Ahamed**[2]
[1]*Department of Statistics, North-Eastern Hill University, Shillong, India*
[2]*Department of Mathematics, Umshyrpi College, Shillong, India*

Received: 08 September 2020; Revised: 23 March 2021; Accepted: 28 March 2021

Abstract

In this paper, the problem of optimum stratification in stratified sampling is considered on a concomitant variable *x* which is highly correlated with the estimation variable *y*, in the light of a-priori distributions, for the allocation proportional to stratum total of the auxiliary variable *x*. Unlike earlier techniques and approaches, available in literature, used in obtaining methods of stratification for the said allocation, the problem is dealt with in a different way and hence a distinct set of equations giving optimum points of stratification and a few distinct methods for finding approximately optimum points of stratification have been obtained. All these proposed methods of stratification are found efficient as well as easy to use when examined empirically by illustrating them in several generated data.

Keywords: Allocation; Auxiliary variable; Optimum points of stratification; Probability density functions; Simple random sampling with and without replacement.

AMS Subject Classification: 62D05, 62-08

1. Introduction

In the field of construction of strata in stratified sampling, it was Dalenius (1950) who first considered the problem of optimum stratification based on estimation variable for Tschuprow (1923) and Neyman (1934) optimum allocation. Since then, the work in this area has been being progressively extended by several researchers into various ways, dimensions and results. To mention a few pertinent ones, among many, are Dalenius and Gurney (1951), Dalenius and Hodges (1959), Taga (1967), Singh and Sukhatme (1969, 1972), Singh and Prakash (1975), Singh (1971, 1975 a, b, c) *etc*., who considered the problem of construction of strata for various allocations with different methods of sampling.

Rao (1968) obtained allocation proportional to stratum total of auxiliary variable which is highly correlated with estimation variable, auxiliary variable proportional allocation (AVPA), under the following superpopulation model:

$$\begin{aligned} &(i)\ \ \xi\left(y_i|x_i\right) = \alpha + \beta\, x_i \\ &(ii)\ V\left(y_i|x_i\right) = \sigma^2\, x_i^2 \\ &(iii)\ \varsigma\left(y_i, y_j|x_i, x_j\right) = 0 \end{aligned} \qquad (1)$$

Corresponding Author: B.K. Gupt
Email: bhuwaneshwargupt@gmail.com

Yadava and Singh (1984) considered the problem of optimum stratification under simple random sampling scheme for the abovementioned allocation – AVPA - proposed by Rao (1968). They obtained equations giving optimum points of stratification (OPS) and a few methods of finding approximately optimum points of stratification (AOPS). Gupt and Ahamed (2020) too considered problem of optimum stratification for a generalized AVPA obtained by Gupt (2003, 2012), *i.e.*,

$$n_h \alpha Z_h \tag{2}$$

where $Z_h = \sum X_{hj}^{g/2}, j = 1, 2, ..., N_h$, provided $\eta_h(g) = \frac{\sigma_h(x)}{\overline{Z}_h}$ are equal in all strata h=1, 2, …, L under the model (1) when $V(y_i|x_i)$ is proportional to x_i^g and g is the level of heteroscedasticity. A particular case of the methods of stratification obtained by the authors when g =2 gives the methods of stratification for the allocation AVPA which were also obtained by Yadava and Singh (1984).

In this paper too, the problem of optimum stratification for allocation AVPA in the light of the priori distributions (1) is considered, but a different technique and procedure is used. We obtain equations giving OPS and a few methods for finding AOPS which are quite different from the ones obtained by Yadava and Singh (1984), *i.e.*, the particular case of the methods obtained by Gupt and Ahamed (2020). The methods of approximation obtained in this paper are suitable for practical applications. These proposed methods will hold good for stratified simple random sampling without replacement also when finite population correction is ignored in each stratum.

In section 2 of this paper, the equations giving OPS and a few methods giving AOPS are obtained. In section 3, numerical illustrations of all the proposed methods in this paper by using generated data and comparison of the methods with respect to the methods proposed by Gupt and Ahamed (2020) for particular case g=2 are carried out. In section 4, conclusion is given.

2. Equations Giving OPS and Methods of Finding AOPS

The allocation taken is

$$n_h \propto X_h, \tag{3}$$

where, X_h is total of h^{th} stratum. The allocation (3) can be written as

$$\eta_h = n \frac{W_h \overline{X}_h}{\overline{X}}, \tag{4}$$

where W_h is the proportion of population units in the h^{th} stratum, $\overline{X}_h$ is the mean for x in h^{th} stratum and $\overline{X}$ is the population mean. Using (4) in the sampling variance for stratified sampling in simple random sampling with replacement (SRSWR), *i.e.*, $V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_{hy}^2}{n_h}$, we get,

$$V(\overline{y}_{st}) = \frac{\overline{X}}{n} \sum_{h=1}^{L} W_h \frac{\sigma_{hy}^2}{\overline{X}_h} \tag{5}$$

where, L is total number of strata.

Considering the superpopulation model (1), Gupt (2003, 2012) obtained conditional expectation of σ_{hy}^2 given x as

$$E\left(\sigma_{hy}^2 \middle| \underline{\mathbf{X}}_h\right) = \beta^2\sigma_{hx}^2 + \sigma^2\frac{N_h - 1}{N_h}\left(\frac{1}{N_h}\sum_{i=1}^{N_h} X_{hi}^2\right), \text{ where } \underline{\mathbf{X}}_h^{/} = \left(\mathrm{X_{h1}}\ \mathrm{X_{h2}} \ldots\ \mathrm{X_{hN_h}}\right).$$

For strata of large sizes, we can consider $\frac{N_h - 1}{N_h} \approx 1$, and therefore, we obtain the following expression

$$E\left(\sigma_{hy}^2 \middle| \underline{\mathbf{X}}_h\right) = \left(\beta^2 + \sigma^2\right)\sigma_{hx}^2 + \sigma^2\,\overline{X}_h^2$$

$$\Rightarrow E\left(V\left(\bar{y}_{st}\middle|\widehat{\mathbf{X}}\right)\right) = \frac{\overline{X}}{n}\sum_h \frac{W_h}{\overline{X}_h}\left\{\left(\beta^2 + \sigma^2\right)\sigma_{hx}^2 + \sigma^2\,\overline{X}_h^2\right\}, \tag{6}$$

where $\widehat{\mathbf{X}} = (\underline{\mathbf{X}}_1, \underline{\mathbf{X}}_2, \ldots, \underline{\mathbf{X}}_L)$.

The conditional expectation of variance in (6) is partially differentiated with respect to x_h (h=1, 2, …, L-1), and the derivative is equated to zero to obtain equations that minimize the expected conditional variance. Thus, we proceed as follows:

$$\frac{\delta}{\delta x_h}E\left(V(\bar{y}_{st}) \middle| \widehat{\mathbf{X}}\right) = 0$$

$$\Rightarrow \frac{\delta}{\delta x_h}\left[\frac{W_h}{\overline{X}_h}\left\{\left(\beta^2 + \sigma^2\right)\sigma_{hx}^2 + \sigma^2\overline{X}_h^2\right\}\right] + \frac{\delta}{\delta x_h}\left[\frac{W_{h+1}}{\overline{X}_{h+1}}\left\{\left(\beta^2 + \sigma^2\right)\sigma_{(h+1)x}^2 + \sigma^2\overline{X}_{h+1}^2\right\}\right] = 0 \tag{7}$$

Taking the first term, we get

$$\frac{\delta}{\delta x_h}\left[\frac{W_h}{\overline{X}_h}\left\{\left(\beta^2 + \sigma^2\right)\sigma_{hx}^2 + \sigma^2\overline{X}_h^2\right\}\right] = \left(\beta^2 + \sigma^2\right)\frac{\delta}{\delta x_h}\left(\frac{W_h\sigma_{hx}^2}{\overline{X}_h}\right) + \sigma^2\frac{\delta}{\delta x_h}\left(W_h\overline{X}_h\right) \tag{8}$$

and considering the expression

$$\frac{\delta}{\delta x_h}\left(\frac{W_h\sigma_{hx}^2}{\overline{X}_h}\right) = \frac{\overline{X}_h\frac{\delta}{\delta x_h}\left(W_h\sigma_{hx}^2\right) - W_h\sigma_{hx}^2\frac{\delta\overline{X}_h}{\delta x_h}}{\overline{X}_h^2} \tag{9}$$

If we consider $f(x)$ is the probability density function for stratification variable x, we have

$$W_h = \int_{x_{h-1}}^{x_h} f(t)dt$$

$$W_h\frac{\delta\overline{X}_h}{\delta x_h} = \left(x_h - \overline{X}_h\right)f(x_h) \tag{10}$$

Therefore, we can get

$$\frac{\delta}{\delta x_h}\left(W_h\sigma_{hx}^2\right) = \left(x_h - \overline{X}_h\right)^2 f(x_h) \tag{11}$$

Using (10) and (11) in equation (9), we get

$$\frac{\delta}{\delta x_h}\left(\frac{W_h\sigma_{hx}^2}{\overline{X}_h}\right) = \frac{\left(x_h - \overline{X}_h\right)}{\overline{X}_h}\left[\overline{X}_h\left(x_h - \overline{X}_h\right) - \sigma_{hx}^2\right]f(x_h) \tag{12}$$

Again using (10) and (12) in equations (8), we get

$$\frac{\delta}{\delta x_h}\left[\frac{W_h}{\overline{X}_h}\left\{\left(\beta^2+\sigma^2\right)\sigma_{hx}^2+\sigma^2\overline{X}_h^2\right\}\right]$$
$$=\left[\left(\beta^2+\sigma^2\right)\frac{\left(x_h-\overline{X}_h\right)}{\overline{X}_h^2}\left\{\overline{X}_h\left(x_h-\overline{X}_h\right)-\sigma_{hx}^2\right\}+\sigma^2 x_h\right]f\left(x_h\right) \quad (13)$$

Similarly,

$$\frac{\delta}{\delta x_h}\left[\frac{W_{h+1}}{\overline{X}_h}\left\{\left(\beta^2+\sigma^2\right)\sigma_{(h+1)x}^2+\sigma^2\overline{X}_{h+1}^2\right\}\right]$$
$$=-\left[\left(\beta^2+\sigma^2\right)\frac{\left(x_h-\overline{X}_{h+1}\right)}{\overline{X}_{h+1}^2}\left\{\overline{X}_{h+1}\left(x_h-\overline{X}_{h+1}\right)-\sigma_{(h+1)x}^2\right\}+\sigma^2 x_h\right]f\left(x_h\right) \quad (14)$$

Substituting (13) and (14) in (7), we get

$$\left(\beta^2+\sigma^2\right)\frac{\left(x_h-\overline{X}_h\right)}{\overline{X}_h^2}\left\{\overline{X}_h\left(x_h-\overline{X}_h\right)-\sigma_{hx}^2\right\}+\sigma^2 x_h$$
$$=\left(\beta^2+\sigma^2\right)\frac{\left(x_h-\overline{X}_{h+1}\right)}{\overline{X}_{h+1}^2}\left\{\overline{X}_{h+1}\left(x_h-\overline{X}_{h+1}\right)-\sigma_{(h+1)x}^2\right\}+\sigma^2 x_h$$

$$\Rightarrow\left(x_h-\overline{X}_h\right)\left\{\frac{x_h}{\overline{X}_h}-(1+C_{hx}^2)\right\}=\left(x_h-\overline{X}_{h+1}\right)\left\{\frac{x_h}{\overline{X}_{h+1}}-(1+C_{(h+1)x}^2)\right\} \quad (15)$$

where, $C_{hx}=\frac{\sigma_{hx}}{\overline{X}_h}$ is the coefficient of variation of x variable in the h^{th} stratum.

Thus, we have got equations (15) which will give OPS.

For finding methods of approximation, we neglect square of coefficients of variation C_{hx}^2 in (15) as square of coefficients of variation are expected to be very small quantities relatively with unity, then we can get

$$x_h=\sqrt{\overline{X}_h\,\overline{X}_{h+1}} \quad (16)$$

which will give AOPS.

On the other hand, we may consider C_{hx}^2 are not negligible in all strata, but they are approximately equal in two successive strata, $C_{1x}^2\approx C_{2x}^2$, $C_{2x}^2\approx C_{3x}^2$, ..., $C_{L-1\,x}^2\approx C_{L\,x}^2$. If each of C_{hx}^2 and $C_{(h+1)x}^2$ is approximately replaced by their geometric mean in (15), we can get

$$x_h=\sqrt{1+C_{hx}\,C_{(h+1)x}}\;\sqrt{\overline{X}_h\,\overline{X}_{h+1}} \quad (17)$$

which will also give AOPS.

If we replace C_{hx}^2 and $C_{(h+1)x}^2$ by their arithmetic mean such as

$$C_{hx}^2\approx C_{(h+1)x}^2=\frac{C_{hx}^2+C_{(h+1)x}^2}{2}\equiv\overline{C}_{h+\frac{1}{2},x}^{\,2}.$$

Then, we can again obtain,

$$x_h=\sqrt{1+\overline{C}_{h+\frac{1}{2},x}^{\,2}}\;\sqrt{\overline{X}_h\,\overline{X}_{h+1}} \quad (18)$$

which will also give AOPS.

It is clearly seen that in terms of analytical justification, our methods obtained in this paper do have equally strong ground with that of proposed methods by Yadava and Singh (1984) - special case of the Gupt and Ahamed's (2020) methods. However, the different techniques and procedure used in this paper have yielded different methods and results. It is also observed that our methods are much easier to use and provide more options for practical applications, whereas their methods are complicated and provide less option for practical applications, particularly in the practical applications of their approximation methods. Since their approximation methods of stratification are presented in the form of definite integrals involving known probability density function of the auxiliary variable based on which stratification is to be done, if the stratification variable in population does not follow a known probability density function, their methods cannot be used. Our methods are free from such restriction.

3. Numerical Illustrations

In order to examine the efficiencies of proposed methods - equations (15) giving OPS and methods of approximation (16), (17) and (18) giving AOPS - we use all these methods in the populations generated by the following probability density functions.

a) Chi-square distribution: $f(x) = \frac{1}{2} e^{\frac{-x+1}{2}},\ 1 \le x < \infty$

b) Exponential distribution: $f(x) = e^{-x+1},\ 1 \le x < \infty$

c) Normal distribution, $X \sim \text{N}(3, 1)$: $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-3)^2}{2}},\ -\infty < x < \infty$

d) Right triangular distribution: $f(x) = 2(2-x),\ 1 \le x \le 2$

e) Standard Gamma distribution (Erlang distribution): $f(x) = \frac{1}{2} x e^{-x+1},\ 1 \le x < \infty$

f) Uniform distribution: $f(x) = 1,\ 1 \le x \le 2$

In the case of generated populations by Chi-square, Exponential and Standard Gamma distribution functions, we truncate the distributions such that area under the curve to the right of the truncation point is 0.05 whereas normal distribution is truncated at two points such that each area under the curve to left and right of truncation points is 0.05. We use equations (15) and methods of approximation (16), (17) and (18) in stratifying each of the generated populations into numbers of strata L=2, 3, 4, 5, 6 in which OPS and AOPS are found out for each number of strata, and hence sampling variances of stratified sampling at optimum points and approximately optimum points for each number of strata are calculated. OPS and AOPS are found by successive iterations. We also calculate sampling variances of stratified sampling for equal interval stratification in each of the populations for each of the considered numbers of strata. The relative efficiencies of equations (15) and methods of approximation (16), (17) and (18) with respect to equal interval stratification are calculated and shown in Tables 1, 2, 3, 4, 5 and 6.

The regression function $C(x) = \alpha + \beta x$ is taken to be linear with the slope at 45^0. The constant σ^2 in $V(y \mid x) = \sigma^2 x^2$ is determined in each case in such a way that 90% of the total variation is accounted for by the regression. Secondly, we compare the efficiencies of equations (15) and methods of approximations (16), (17) and (18) with the equations giving OPS and methods of approximation proposed by Gupt and Ahamed (2020) for particular case

g=2, in the three populations of Uniform, Right-triangular and Exponential density functions in which the authors illustrated their proposed methods. These comparisons are given in Tables 7, 8 and 9.

Table 1: Comparison with equal interval stratification, Chi-square distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y_{st}})$	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency
2	4.00	0.841	2.747	0.646	130	2.661	0.646	130	2.747	0.646	130	2.747	0.646	130
3	3.00 5.00	0.795	2.112 3.759	0.349	228	2.055 3.689	0.350	227	2.111 3.758	0.349	227	2.112 3.758	0.349	227
4	2.50 4.00 5.50	0.359	2.015 3.189 4.622	0.275	131	1.735 2.725 4.300	0.277	129	2.015 3.188 4.623	0.275	131	2.016 3.189 4.623	0.275	131
5	2.20 3.40 4.60 5.80	0.279	2.015 3.172 4.504 5.860	0.262	106	1.735 2.725 4.240 5.809	0.259	108	2.015 3.172 4.504 5.860	0.262	106	2.016 3.172 4.505 5.860	0.262	106
6	2.00 3.00 4.00 5.00 6.00	0.249	1.751 2.613 3.631 4.797 6.177	0.219	114	1.729 2.593 3.622 4.789 6.171	0.219	114	1.751 2.612 3.652 4.848 6.202	0.219	114	1.751 2.613 3.652 4.849 6.207	0.219	114

It is seen that in the population of Chi-square distribution, the proposed equations (15) and methods of approximation are found to be more efficient than that of equal interval stratification for all numbers of strata and relative efficiency is remarkably high for stratum 3. Moreover, methods of approximations (16), (17) and (18) are having same efficiencies with that of the proposed equations (15) in almost all the numbers of strata.

Table 2: Comparison with equal interval stratification, Exponential distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y_{st}})$	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency
2	2.50	0.202	2.054	0.167	121	1.986	0.168	120	2.054	0.167	121	2.054	0.167	121
3	2.00 3.00	0.123	1.689 2.572	0.100	123	1.649 2.534	0.101	122	1.688 2.572	0.100	123	1.689 2.572	0.100	123
4	1.75 2.50 3.25	0.089	1.645 2.317 3.151	0.085	105	1.385 1.935 2.765	0.075	119	1.645 2.317 3.148	0.086	103	1.646 2.317 3.150	0.086	103
5	1.60 2.20 2.80 3.40	0.081	1.376 1.859 2.465 3.215	0.066	123	1.333 1.787 2.355 3.145	0.062	131	1.376 1.859 2.465 3.215	0.066	123	1.376 1.859 2.465 3.215	0.066	123
6	1.50 2.00 2.30 2.80 3.50	0.068	1.324 1.690 2.046 2.527 3.215	0.054	127	1.319 1.688 2.043 2.524 3.213	0.052	130	1.324 1.690 2.046 2.527 3.215	0.053	128	1.324 1.690 2.046 2.527 3.215	0.053	128

In the population of Exponential distribution, all the proposed methods are performing with more efficiencies than that of equal interval stratification for all the numbers of strata. Moreover, the proposed approximation methods are found in most cases to be performing with same efficiencies as or higher than that of proposed equations (15) giving OPS.

Table 3: Comparison with equal interval stratification, Normal distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y}_{st})$	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency
2	3.00	0.291	2.966	0.291	100	2.928	0.291	100	2.964	0.291	100	2.967	0.291	100
3	2.366 3.633	0.227	2.348 3.492	0.222	102	2.199 3.276	0.197	115	2.346 3.491	0.222	102	2.349 3.442	0.222	102
4	2.05 3.00 3.95	0.139	2.072 2.939 3.783	0.121	115	2.055 2.927 3.759	0.122	114	2.068 2.939 3.783	0.121	115	2.073 2.939 3.783	0.121	115
5	1.86 2.62 3.38 4.14	0.117	2.068 2.848 3.350 3.884	0.111	106	1.996 2.602 3.085 3.793	0.112	104	2.063 2.847 3.350 3.884	0.111	106	2.069 2.848 3.350 3.884	0.111	106
6	1.733 2.366 3.00 3.633 4.266	0.107	1.629 2.239 2.921 3.539 4.095	0.103	104	1.602 2.215 2.917 3.536 4.091	0.0959	111	1.626 2.239 2.921 3.539 4.095	0.103	104	1.629 2.239 2.921 3.539 4.095	0.101	106

In the population of Normal distribution, the proposed methods perform with higher efficiencies than that of equal interval stratification. At the same time, approximation methods are performing almost same as or slightly better than that of equations (15) although approximation method (16) is performing better than approximation methods (17) and (18) for numbers of strata 4, 5 and 6.

Table 4: Comparison with equal interval stratification, Right Triangular distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y}_{st})$	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency	Points of stratification	$nV(\overline{y}_{st})$	Relative Efficiency
2	1.50	0.028	1.373	0.022	127	1.355	0.022	127	1.373	0.022	127	1.394	0.022	127
3	1.334 1.667	0.014	1.273 1.587	0.013	108	1.269 1.585	0.013	108	1.272 1.587	0.013	108	1.272 1.587	0.013	108
4	1.25 1.50 1.75	0.012	1.180 1.386 1.629	0.009	133	1.178 1.390 1.628	0.009	133	1.179 1.386 1.631	0.010	120	1.179 1.387 1.629	0.010	120
5	1.20 1.40 1.60 1.80	0.010	1.170 1.370 1.579 1.736	0.010	100	1.178 1.380 1.578 1.736	0.009	111	1.179 1.379 1.579 1.736	0.010	100	1.179 1.379 1.579 1.736	0.010	100
6	1.167 1.334 1.499 1.667 1.833	0.009	1.145 1.289 1.435 1.589 1.736	0.009	100	1.121 1.270 1.430 1.588 1.736	0.008	113	1.145 1.290 1.435 1.589 1.736	0.009	100	1.145 1.289 1.435 1.589 1.736	0.009	100

In the population of Right-triangular distribution, all the proposed methods of stratification are found to be having higher efficiencies than that of equal interval stratification except at numbers of strata 5 and 6 at which the proposed equations (15) and approximation methods (17) and (18) have same efficiencies with that of equal interval stratification. But, proposed method (16) performs better than equal interval stratification.

Table 5: Comparison with equal interval stratification, Standard Gamma distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y_{st}})$	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency
2	3.00	0.405	2.614	0.373	109	2.539	0.368	110	2.614	0.373	109	2.616	0.373	109
3	2.33 3.66	0.268	1.882 3.044	0.205	131	1.902 3.076	0.211	127	1.882 3.043	0.205	131	1.896 3.051	0.205	131
4	2.00 3.00 4.00	0.169	1.793 2.780 3.986	0.161	105	1.749 2.673 3.796	0.161	105	1.793 2.780 3.984	0.161	105	1.794 2.781 3.986	0.161	105
5	1.80 2.60 3.40 4.20	0.145	1.753 2.534 3.268 4.153	0.143	102	1.695 2.396 3.105 4.056	0.137	106	1.752 2.534 3.268 4.152	0.143	102	1.753 2.534 3.268 4.153	0.143	102
6	1.66 2.32 2.98 3.64 4.30	0.129	1.660 2.221 2.862 3.625 4.327	0.129	100	1.630 2.170 2.833 3.620 4.326	0.128	102	1.644 2.188 2.845 3.624 4.327	0.131	99	1.645 2.188 2.846 3.625 4.327	0.131	99

In the population of Standard Gamma distribution, all the proposed methods of stratification are stratifying the population more efficiently than equal interval stratification except for number of strata 6 at which approximation methods (17) and (18) are having slightly less efficiencies than that of equal interval stratification and the proposed equations (15) perform with same efficiency as that of equal interval stratification. In this population too, the proposed method of approximation (16) is found to be performing with higher efficiency than that of equal interval stratification as well as all other proposed methods.

Table 6: Comparison with equal interval stratification, Uniform distribution

No. of Strata (L)	Equal Interval Stratification		Stratification due to equations (15)			Stratification due to approx. method (16)			Stratification due to approx. method (17)			Stratification due to approx. method (18)		
	Points of stratification	$nV(\overline{y_{st}})$	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency	Points of stratification	$nV(\overline{y_{st}})$	Relative Efficiency
2	1.50	0.028	1.504	0.028	100	1.498	0.028	100	1.504	0.028	100	1.504	0.028	100
3	1.334 1.667	0.016	1.323 1.667	0.016	100	1.318 1.658	0.016	100	1.323 1.666	0.016	100	1.323 1.666	0.016	100
4	1.25 1.50 1.75	0.013	1.252 1.475 1.732	0.013	100	1.248 1.469 1.727	0.0126	101	1.252 1.475 1.732	0.013	100	1.252 1.475 1.732	0.013	100
5	1.20 1.40 1.60 1.80	0.0124	1.212 1.369 1.529 1.743	0.0119	104	1.211 1.368 1.529 1.742	0.0119	104	1.212 1.369 1.529 1.743	0.0119	104	1.212 1.369 1.529 1.743	0.0119	104
6	1.167 1.334 1.499 1.667 1.833	0.0107	1.169 1.318 1.504 1.716 1.873	0.0108	99	1.162 1.303 1.488 1.708 1.873	0.0107	100	1.169 1.318 1.504 1.716 1.873	0.0108	99	1.169 1.318 1.504 1.716 1.873	0.0108	99

For the population of Uniform distribution, equal interval stratification is generally considered to be the best stratification method. We have observed that all the proposed methods of stratification work with same efficiencies as that of equal interval stratification in almost all the considered numbers of strata except at number of strata 6, the proposed equations (15) and methods of approximation (17) and (18) are performing with slightly less efficiencies than that of equal interval stratification. But at stratum number 5, our proposed methods are performing with slightly higher efficiencies than that of equal interval

stratification. Hence it is seen that all our proposed methods of stratification are efficient in stratifying the population. The proposed approximation method (16) is found to be most credibly stratifying the population to give OPS.

On the other hand, as Yadava and Singh (1984) did not perform numerical illustration, we compare our proposed equations and methods of approximation with that of Gupt and Ahamed (2020) whose special case gives Yadava and Singh's (1984) methods. The comparisons are shown in Tables 7-9. Gupt and Ahamed (2020) illustrated their proposed methods in three generated populations of Exponential, Right-triangular and Uniform distribution in the same range of x variable as taken in this paper too, and therefore the comparison of our methods with respect to their methods are made in these three populations only.

Table 7: Comparison with the methods proposed by Gupt and Ahamed (2020), Exponential distribution

No. of Strata (L)	Equations for OPS proposed by Gupt and Ahamed(2020)	Proposed equations (15)	Comparison of (15) *w.r.t* equations for OPS proposed by Gupt and Ahamed (2020)	Approximation methods proposed by Gupt and Ahamed (2020)	$nV(\overline{y_{st}})$ for proposed Approximation Methods			Comparison of methods (16), (17) and (18) *w.r.t* approximation method proposed by Gupt and Ahamed (2020)		
	$nV(\overline{y_{st}})$	$nV(\overline{y_{st}})$	Relative efficiency	$nV(\overline{y_{st}})$	Method (16)	Method (17)	Method (18)	Relative efficiency of (16)	Relative efficiency of (17)	Relative efficiency of (18)
2	0.1673	0.167	100	0.1683	0.168	0.167	0.167	100	101	101
3	0.1029	0.100	103	0.1045	0.100	0.100	0.100	105	105	105
4	0.0855	0.085	101	0.0729	0.075	0.086	0.086	97	85	85
5	0.0659	0.066	100	0.0605	0.062	0.066	0.066	98	92	92
6	0.0535	0.053	101	0.053	0.052	0.053	0.053	102	100	100

In Exponential distribution, our proposed equations (15) are found to be having same or slightly higher efficiencies than that of equations giving OPS proposed by Gupt and Ahamed (2020). When we compare our proposed methods of approximation with the method of approximation proposed by Gupt and Ahamed (2020), for number of strata 2, 3 and 6 our approximation methods are slightly better or equal to that of their method, but for number of strata 4 and 5, their approximation method is slightly better. But our proposed approximation (16) is having almost same efficiency as that of theirs for numbers of strata 4 and 5.

Table 8: Comparison with the methods proposed by Gupt and Ahamed (2020), Right-triangular distribution

No. of Strata (L)	Equations for OPS proposed by Gupt and Ahamed (2020)	Proposed equations (15)	Comparison of (15) *w.r.t* equations for OPS proposed by Gupt and Ahamed (2020)	Approximation methods proposed by Gupt and Ahamed (2020)	$nV(\overline{y_{st}})$ for proposed Approximation Methods			Comparison of methods (16), (17) and (18) *w.r.t* approximation method proposed by Gupt and Ahamed (2020)		
	$nV(\overline{y_{st}})$	$nV(\overline{y_{st}})$	Relative efficiency	$nV(\overline{y_{st}})$	Method (16)	Method (17)	Method (18)	Relative efficiency of (16)	Relative efficiency of (17)	Relative efficiency of (18)
2	0.0224	0.022	102	0.0234	0.022	0.022	0.022	106	106	106
3	0.0130	0.013	100	0.0130	0.013	0.013	0.013	100	100	100
4	0.0109	0.009	121	0.0102	0.009	0.010	0.010	113	102	102
5	0.0098	0.010	98	0.0103	0.009	0.010	0.010	114	103	103
6	0.0088	0.009	98	0.0087	0.008	0.009	0.009	109	97	97

In the population of Right-triangular distribution, the proposed equations (15) are having slightly more efficiencies than that of their proposed equations giving OPS for numbers of strata 2, 3 and 4. For numbers of strata 5 and 6, the proposed equations (15) are having slightly less efficiencies than that of the equations proposed by them. But in the case of comparison of the methods of approximations (16), (17) and (18) with their proposed method of approximation, our methods are having higher efficiencies except at number of strata 6, our proposed methods (17) and (18) are having slightly less efficiencies than that of their method of approximation. But, the proposed approximation method (16) is performing best of all approximation methods proposed in this paper as well as proposed by them.

Table 9: Comparison with the methods proposed by Gupt and Ahamed (2020), Uniform distribution

No. of Strata (L)	Equations for OPS proposed by Gupt and Ahamed (2020)	Proposed equations (15)	Comparison of (15) *w.r.t* equations for OPS proposed by Gupt and Ahamed (2020)	Approximation methods of Gupt and Ahamed (2020)	$nV(\overline{y_{st}})$ for proposed Approximation Methods			Comparison of methods (16), (17) and (18) *w.r.t* approximation method proposed by Gupt and Ahamed (2020)		
	$nV(\overline{y_{st}})$	$nV(\overline{y_{st}})$	Relative efficiency	$nV(\overline{y_{st}})$	Method (16)	Method (17)	Method (18)	Relative efficiency of (16)	Relative efficiency of (17)	Relative efficiency of (18)
2	0.0278	0.028	99	0.0283	0.0278	0.028	0.0278	102	101	102
3	0.0164	0.0164	100	0.0168	0.0162	0.0164	0.0164	104	102	102
4	0.0127	0.0127	100	0.0128	0.0126	0.0127	0.0127	102	101	101
5	0.0119	0.0119	100	0.0117	0.0119	0.0119	0.0119	98	98	98
6	0.0108	0.0108	100	0.0113	0.0107	0.0108	0.0108	106	105	105

In the population of Uniform distribution, the proposed equations (15) perform with almost same efficiencies as that of their proposed equations giving OPS. The proposed methods of approximation (16), (17) and (18) are performing with higher efficiencies than that of approximation method proposed by them except for number of strata 5 at which these

proposed methods are performing with slightly less efficiencies than that of approximation method proposed by them.

In using the approximation methods, Gupt and Ahamed's (2020) method is restricted to use in stratifying population in which auxiliary variable, based on which stratification is to be done, of the study variable follows a known probability density function, but our three proposed methods of approximation are free from such restriction.

4. Conclusion

In this paper, it is observed that the proposed equations (15) giving OPS and methods of approximation (16), (17) and (18) giving AOPS are found to be different from the equations and methods proposed by Yadava and Singh (1984), which could be obtained as particular case of the generalised methods proposed by Gupt and Ahamed (2020), although allocation used is same in all the cases. The use of different techniques and procedure in the same problem in this paper has yielded a distinct set of equations for giving OPS and a few methods of obtaining AOPS which are very efficient and fairly suitable for practical applications in stratifying various heteroscedastic populations. In all the empirical illustrations, the proposed equations and methods of approximation are found to be performing better than or as good as methods of stratification proposed by other authors considered in this paper. It is also fascinating to learn that all the approximation methods proposed in this paper are found almost as efficient as exact equations (15) giving OPS. The method of approximation (16) - AOPS between any two consecutive strata are given by geometric mean of means of two consecutive strata of the auxiliary variable - is found to be the best of all the proposed methods in terms of efficiently stratifying population of all types and suitability for practical applications. Therefore, this method is recommended for practical application in stratifying populations of the considered level of heteroscedasticity optimally.

Acknowledgements

We, the authors of this paper, express our sincere thankfulness to the anonymous reviewer whose enormous suggestion has led to the elaborate improvement of the paper. We are also thankful to the Chief Editor for his valuable suggestion during the revision of the paper.

References

Dalenius, T. (1950). The problem of optimum stratification-I.*Skandinavisk Aktuaritidskrift*, **33**, 203-213.

Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification-II. *Skandinavisk Aktuaritidskrift*, **34**, 133-148.

Dalenius, T., and Hodges, J. L. (1957). The choice of stratification points. *Skandinavisk Aktuaritidskrift*, **3(4)**, 198-203.

Gupt, B. K. (2003). Sample size allocation for stratified sampling under a correlated superpopulation model. *METRON -International Journal of Statistics, LXI, n,1:*35-52.

Gupt, B. K. (2012). *Allocation of sample size in stratified sampling under superpopulation models.* Saarbrucken, Deutschland/Germany: LAP LAMBERTAcademic Publishing AV Akademikerverlag GmbH & Co.KG.

Gupt, B. K. and Ahamed, M. I. (2020). Optimum stratification for a generalized auxiliary variable proportional allocation under a superpopulation model. *Communications in Statistics-Theory and Methods*, Published online: 25 July 2020. https://doi.org/10.1080/03610926.2020.1793203

Neyman, J. (1934). On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97(4)**, 558-625.

Rao, T. J. (1968). On the allocation of sample size in stratified sampling. *Annals of the Institute of Statistical Mathematics*, **20**, 159-166.

Singh, R. and Sukhatme, B. V. (1969). Optimum stratification. *Annals of the Institute of Statistical Mathematics*, **21**, 515-528.

Singh, R. and Sukhatme, B. V. (1972). Optimum stratification in sampling with varying probabilities. *Annals of the Institute of Statistical Mathematics*, **24**, 485-494.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association*, **66**, 829-833.

Singh, R. (1975a). An alternate method of stratification on the auxiliary variable. *Sankhya* **C37,** 100-108.

Singh, R. (1975b). On optimum stratification for proportional allocation. *Sankhya*, C**37**, Pt.I, 109-115.

Singh, R. (1975c). A note on optimum stratification in sampling with varying probabilities. *Australian Journal of Statistics*, **27(1)**, 12–21.

Singh, R. and Prakash, D. (1975). Optimum stratification for equal allocation. *Annals of the Institute of Statistical Mathematics*, **27**, 273-280.

Taga, Y. (1967). On optimum stratification for the objective variable based onconcomittant variable. *Annals of the Institute of Statistical Mathematics*, **19**, 101-130.

Tschuprow, A. A. (1923). On mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, **2**, 461-493.

Yadava, S. S. and Singh, R. (1984). Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics: Theory and Methods*, **13(22)**, 2793-2806.

Statistics and Application {ISSN 2454-7395 (online)}
Volume 20, No 1, 2022 (New Series), pp 193-211

Chain Ratio-Type and Exponential Chain Ratio-Ratio-Type Estimators in Double Sampling for Stratification

Housila P. Singh and Pragati Nigam
School of Studies in Statistics, Vikram University, Ujjain, (M.P.), India.

Received: 04 October 2020; Revised: 18 March 2021; Accepted: 21 April 2021

Abstract

In this paper, we have suggested chain ratio-type estimator, exponential chain ratio-ratio-type estimator, improved estimator and a general class of estimators in Double sampling for stratification for finite population mean. Different conditions were obtained under which the proposed estimators perform better than unbiased estimator, ratio-type and product-type estimators and ratio and product-type exponential estimators. An empirical study is carried out to demonstrate the performance of the proposed estimators over existing estimators.

Keywords: Auxiliary variable; Bias; Mean squared error.

Mathematics Subject Classification Code: 62D05.

1. Introduction and Notations

When the population is homogeneous, for selecting a representative sample from the population, the practitioner usually uses the simple random sampling scheme. However, in practice, heterogeneous population are also encountered. In such situation, stratification is one of the most widely used procedures in sample survey to provide samples that are representatives of major sub-groups of a population and improve precision of estimators, see Holt and Smith (1979). In stratified random sampling, it is assumed that strata weights as well as sampling frame are available in advance. But there are several situations of practical importance where strata weights are known and the frame within strata is not available. For example, in a household survey in a city, number of households in different colonies may be available, but list of households may not be available, see Tailor *et al*. (2014). In such a situation the technique of post stratification is effectively employed. However, in other situations strata weights may not be known exactly as they become outdated with the passage of time. Further the information on the stratification variable may not be readily available but could be made available by diverting a part of the survey budget, see Tripathi and Bahl (1991, p. 2590). Under these situations that procedure of double sampling for stratification (*DSS*) can be employed.

Let $U = \{U_1, U_2, ..U_N\}$ be a finite population of size *N*. let (*y*, *x*) be the (study, auxiliary) variates respectively. It is desired to estimate the population mean $\overline{Y}$ of study variable *y* and consider it desirable to stratify the population based on the values of an auxiliary character *x* but the frequency distribution of *x* is not known. The sampling frame for different strata, the

Corresponding Author: Pragati Nigam
Email: pragatinigam1@gmail.com

strata weights $W_h = \frac{N_h}{N}; h = 1,2,...L$, (N_h being the size of the h^{th} stratum) are unknown although the strata may be fixed in advance, see Ige and Tripathi (1987, p. 192). In such situations *DSS* scheme is used. In *DSS* scheme we draw a first phase sample $S^{(1)}$ of size n' from the population *U* using simple random sampling without replacement (*SRSWOR*) scheme and observe auxiliary variable *x*. let $x_j, j = 1,2,...,n'$ be the *x*-observations and $\bar{x}' = \frac{1}{n'}\sum_{j=1}^{n'} x_j$, the sample mean. The sample $S^{(1)}$ is then divided into *L* strata on the basis of information gathered for auxiliary variable *x* through $S^{(1)}$. Let n'_h be the number of units in $S^{(1)}$ falling into stratum $h\left(h = 1,2,...,L; \sum_{h=1}^{L} n'_h = n'\right)$, $n' = \{n'_1, n'_2, ..., n'_L\}$ yielding the representation $\bar{x}' = \sum_{h=1}^{L} w_h \bar{x}'_h$, where $\bar{x}'_h = \sum_{j=1}^{n'_h} \frac{x'_{hj}}{n'_h}$ and $w_h = \frac{n'_h}{n'}$ such that $E(w_h) = W_h = \frac{N_h}{N}$. Subsamples of sizes $n_h = \upsilon_h n'_h, 0 < \upsilon_h < 1 (h = 1,2,...,L)$, υ_h is known in advance for each *h*, are then drawn independently, using *SRSWOR* within each stratum and *y*, the study variable is measured.

Let $n = \sum_{h=1}^{L} n_h$, $n = \{n_1, n_2, ..., n_L\}$ and $y_{hj}, j = 1,2,...,n_h; h = 1,2,...,L$ denote *y* observations, $\bar{y}_h = \sum_{j=1}^{n_h} \frac{y_{hj}}{n_h}$. It is assumed throughout the paper that n'_h is large enough so that $\Pr(n'_h = 0) = 0$ for all *h*.

Further we denote

First degree of approximation: *fda*,

$f = \frac{n'}{N}$: Sampling fraction,

$$\bar{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} y_{hj}, \bar{X}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} x_{hj}, \bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{j=1}^{N_h} y_{hj}, \bar{X} = \frac{1}{N}\sum_{h=1}^{L}\sum_{j=1}^{N_h} x_{hj},$$

$$S_{yh}^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(y_{hj} - \bar{Y}_h\right)^2, S_{xh}^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(x_{hj} - \bar{X}_h\right)^2, S_{yxh} = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(y_{hj} - \bar{Y}_h\right)\left(x_{hj} - \bar{X}_h\right),$$

$$S_y^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(y_{hj} - \bar{Y}\right)^2, S_x^2 = \frac{1}{N-1}\sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(x_{hj} - \bar{X}\right)^2, S_{yx} = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}\left(y_{hj} - \bar{Y}\right)\left(x_{hj} - \bar{X}\right),$$

$$\gamma = \left(\frac{1-f}{n'}\right), R = \frac{\bar{Y}}{\bar{X}}, \varphi_h = \left(\frac{1}{\upsilon_h} - 1\right), k = \frac{\theta_{yx}}{R\theta_x}, \theta_y = \sum_{h=1}^{L} W_h \varphi_h S_{yh}^2, \theta_x = \sum_{h=1}^{L} W_h \varphi_h S_{xh}^2,$$

$$\theta_{yx} = \sum_{h=1}^{L} W_h \varphi_h S_{yxh}, \rho = \frac{\theta_{yx}}{\sqrt{\theta_y \theta_x}}.$$

We note that

$\bar{y}_{ds} = \sum_{h=1}^{L} w_h \bar{y}_h$, $\bar{x}_{ds} = \sum_{h=1}^{L} w_h \bar{x}_h$ are unbiased estimators of the population means $\bar{Y}$ and $\bar{X}$ respectively, where $\bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$ and $\bar{x}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} x_{hj}$.

Now, to obtain the biases and mean squared errors (*MSE*s) of various estimators of population mean $\bar{Y}$, we write

$$\bar{y}_{ds} = \bar{Y}(1+\varepsilon_0), \bar{x}_{ds} = \bar{X}(1+\varepsilon_1), \bar{x}' = \bar{X}(1+\varepsilon'_1)$$

such that $E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon'_1) = 0$ and

$$E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2}\left[\gamma S_y^2 + \frac{1}{n'}\theta_y\right],$$

$$E(\varepsilon_1^2) = \frac{1}{\bar{X}^2}\left[\gamma S_x^2 + \frac{1}{n'}\theta_x\right],$$

$$E(\varepsilon'^2_1) = E(\varepsilon_1 \varepsilon'_1) = \frac{1}{\bar{X}^2}\gamma S_x^2,$$

$$E(\varepsilon_0 \varepsilon_1) = \frac{1}{\bar{Y}\bar{X}}\left[\gamma S_{yx} + \frac{1}{n'}\theta_{yx}\right],$$

$$E(\varepsilon_0 \varepsilon'_1) = \frac{1}{\bar{Y}\bar{X}}\gamma S_{yx}.$$

1.1. Reviewing some existing estimators

The conventional unbiased estimator for population mean $\bar{Y}$ [which does not utilize the entire information gathered on the first-phase (preliminary large) sample and the stratified sub samples] is defined by

$$\bar{y}_{ds} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{1}$$

with mean squared error /variance

$$V(\bar{y}_{ds}) = MSE(\bar{y}_{ds}) = \gamma S_y^2 + \frac{1}{n'}\theta_y, \tag{2}$$

is well known [see Rao (1973); Cochran (1977)].

Based on *DSS*, utilizing the auxiliary information obtained on the first phase sample both at the designing as well as at estimation stages, Ige and Tripathi (1987) proposed the ratio-type (*RT*) and product-type (*PT*) estimators for $\bar{Y}$ respectively as

$$\hat{\bar{Y}}_{R(dss)} = \bar{y}_{ds}\left(\frac{\bar{x}'}{\bar{x}_{ds}}\right), \tag{3}$$

$$\hat{\bar{Y}}_{P(dss)} = \bar{y}_{ds}\left(\frac{\bar{x}_{ds}}{\bar{x}'}\right). \tag{4}$$

To the *fda*, the *MSE*s of $\hat{\bar{Y}}_{R(dss)}$ and $\hat{\bar{Y}}_{P(dss)}$ are respectively given by

$$MSE\left(\hat{\bar{Y}}_{R(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + R^2\theta_x(1-2k)\right], \tag{5}$$

$$MSE\left(\hat{\bar{Y}}_{P(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + R^2\theta_x(1+2k)\right]. \tag{6}$$

Further, motivated by Bahl and Tuteja (1991), Tailor *et al*. (2014) suggested *RT* and *PT* exponential estimators respectively as

$$\hat{\bar{Y}}_{\mathrm{Re}(dss)} = \bar{y}_{ds}\exp\left\{\frac{\bar{x}'-\bar{x}_{ds}}{\bar{x}'+\bar{x}_{ds}}\right\} \quad , \tag{7}$$

$$\hat{\bar{Y}}_{Pe(dss)} = \bar{y}_{ds}\exp\left\{\frac{\bar{x}_{ds}-\bar{x}'}{\bar{x}'+\bar{x}_{ds}}\right\}. \tag{8}$$

The *MSE*s of *RT* and *PT* exponential estimators to the *fda* are respectively given by

$$MSE\left(\hat{\bar{Y}}_{\mathrm{Re}(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + \frac{R^2\theta_x}{4}(1-4k)\right], \tag{9}$$

$$MSE\left(\hat{\bar{Y}}_{Pe(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + \frac{R^2\theta_x}{4}(1+4k)\right]. \tag{10}$$

From (2), (5), (6), (9) and (10), it is observed that the *RT* estimator $\hat{\bar{Y}}_{R(dss)}$, the *PT* estimator $\hat{\bar{Y}}_{P(dss)}$, the *RT* exponential estimator $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$ and the *PT* exponential estimator $\hat{\bar{Y}}_{Pe(dss)}$ are better than $\bar{y}_{ds}$ if the conditions $k > \frac{1}{2}, k < -\frac{1}{2}, k > \frac{1}{4}$ and $k < -\frac{1}{4}$ respectively hold good.

2. Proposed Chain-Type Estimators in DSS

2.1. Chain RT estimator

On replacing $\bar{y}_{ds}$ by $\hat{\bar{Y}}_{R(dss)}$ in (3), we get chain *RT* estimator in *DSS* for population mean $\bar{Y}$ as

$$\hat{\bar{Y}}^{C}_{R(dss)} = \hat{\bar{Y}}_{R(dss)}\left(\frac{\bar{x}'}{\bar{x}_{ds}}\right) = \bar{y}_{ds}\left(\frac{\bar{x}'}{\bar{x}_{ds}}\right)^2. \tag{11}$$

Putting $\bar{y}_{ds} = \bar{Y}(1+\varepsilon_0), \bar{x}_{ds} = \bar{X}(1+\varepsilon_1)$ and $\bar{x}' = \bar{X}(1+\varepsilon'_1)$ in (11) we have

$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon'_1)^2(1+\varepsilon_1)^{-2}. \tag{12}$$

We assume that $|\varepsilon_1| < 1$ so that the term $(1+\varepsilon_1)^{-2}$ is expandable. Now, expanding the right hand side (*RHS*) of (12) multiplying out and neglecting terms of ε's having power greater than two, we have

$$\left(\hat{\bar{Y}}^C_{R(dss)} - \bar{Y}\right) \cong \bar{Y}\left[\varepsilon_0 - 2\varepsilon_1 + 2\varepsilon'_1 + 3\varepsilon_1^2 + \varepsilon'^2_1 - 2\varepsilon_0\varepsilon_1 + 2\varepsilon_0\varepsilon'_1 - 4\varepsilon_1\varepsilon'_1\right]. \tag{13}$$

To obtain the bias of $\hat{\bar{Y}}^C_{R(dss)}$ to the *fda*, we take the expectation of both sides of (13) and thus

$$B\left(\hat{\bar{Y}}^C_{R(dss)}\right) = \frac{R\theta_x}{n'\bar{X}}(3-k). \tag{14}$$

which is negligible if sample size n' is sufficiently large.

Squaring both sides of (13), neglecting terms of ε's having power greater than two, we have

$$= \bar{Y}^2\left[\varepsilon_0^2 + 4\varepsilon_1^2 + 4\varepsilon'^2_1 - 4\varepsilon_0\varepsilon_1 + 4\varepsilon_0\varepsilon'_1 - 8\varepsilon_1\varepsilon'_1\right]. \tag{15}$$

Taking expectation of both sides of (15) we get the *MSE* of $\hat{\bar{Y}}^C_{R(dss)}$ to the *fda* as

$$MSE\left(\hat{\bar{Y}}^C_{R(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + 4R^2\theta_x(1-k)\right]. \tag{16}$$

From (2), (5), (9) and (16) it can be shown that

$$MSE\left(\hat{\bar{Y}}^C_{R(dss)}\right) < MSE(\bar{y}_{ds}) \text{ if } k > 1, \tag{17}$$

$$MSE\left(\hat{\bar{Y}}^C_{R(dss)}\right) < MSE\left(\hat{\bar{Y}}_{R(dss)}\right) \text{ if } k > \frac{3}{2}, \tag{18}$$

and

$$MSE\left(\hat{\bar{Y}}^C_{R(dss)}\right) < MSE\left(\hat{\bar{Y}}_{\mathrm{Re}(dss)}\right) \text{ if } k > \frac{5}{4}. \tag{19}$$

Thus, the proposed chain *RT* estimator $\hat{\bar{Y}}^C_{R(dss)}$ is more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(dss)}$ and $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$ if the conditions (17), (18) and (19) are satisfied respectively. It is also observed from (17), (18) and (19) that the condition $k > \frac{3}{2}$ is sufficient for the proposed estimator $\hat{\bar{Y}}^C_{R(dss)}$ to be more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(dss)}$ and $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$.

2.2. Chain PT estimator

On replacing $\bar{y}_{ds}$ by $\hat{\bar{Y}}_{P(dss)}$ in (4), we get a chain *PT* estimator in *DSS* for population mean $\bar{Y}$ as

$$\hat{\bar{Y}}^{C}_{P(dss)} = \hat{\bar{Y}}_{P(dss)}\left(\frac{\bar{x}_{ds}}{\bar{x}'}\right) = \bar{y}_{ds}\left(\frac{\bar{x}_{ds}}{\bar{x}'}\right)^{2}. \tag{20}$$

Inserting $\bar{y}_{ds} = \bar{Y}(1+\varepsilon_0), \bar{x}_{ds} = \bar{X}(1+\varepsilon_1)$ and $\bar{x}' = \bar{X}(1+\varepsilon'_1)$ in (20) we have

$$\hat{\bar{Y}}^{C}_{P(dss)} = \bar{Y}(1+\varepsilon_0)\left\{\frac{\bar{X}(1+\varepsilon_1)}{\bar{X}(1+\varepsilon'_1)}\right\}^{2} = \bar{Y}(1+\varepsilon_0)(1+\varepsilon_1)^{2}(1+\varepsilon'_1)^{-2}. \tag{21}$$

We assume that $|\varepsilon'_1| < 1$ so that the term $(1+\varepsilon'_1)^{-2}$ is expandable. Now, expanding the *RHS* of (21), multiplying out, neglecting terms of ε's having power greater than two and then subtracting $\bar{Y}$ from both sides of (21), we have

$$\left(\hat{\bar{Y}}^{C}_{P(dss)} - \bar{Y}\right) \cong \bar{Y}\left[\varepsilon_0 + 2\varepsilon_1 - 2\varepsilon'_1 + 3\varepsilon'^{2}_1 + \varepsilon_1^{2} + 2\varepsilon_0\varepsilon_1 - 2\varepsilon_0\varepsilon'_1 - 4\varepsilon_1\varepsilon'_1\right]. \tag{22}$$

Taking expectation of both sides of (22) we get the bias of $\hat{\bar{Y}}^{C}_{P(dss)}$ to the *fda* as

$$B\left(\hat{\bar{Y}}^{C}_{P(dss)}\right) = \frac{R\theta_x}{n'\bar{X}}(1+2k). \tag{23}$$

which is negligible if sample size n' is large enough.

Now, squaring both sides of (22) and neglecting terms of ε's having power greater than two we have

$$= \bar{Y}^{2}\left[\varepsilon_0^{2} + 4\varepsilon_1^{2} + 4\varepsilon'^{2}_1 + 4\varepsilon_0\varepsilon_1 - 4\varepsilon_0\varepsilon'_1 - 8\varepsilon_1\varepsilon'_1\right]. \tag{24}$$

Taking expectation of both sides of (24) we get the *MSE* of $\hat{\bar{Y}}^{C}_{P(dss)}$ to the *fda* as

$$MSE\left(\hat{\bar{Y}}^{C}_{P(dss)}\right) = \gamma S_y^{2} + \frac{1}{n'}\left[\theta_y + 4R^{2}\theta_x(1+k)\right]. \tag{25}$$

It can be easily observed from (2), (6), (10) and (25) that the suggested chain *PT* estimator $\hat{\bar{Y}}^{C}_{P(dss)}$ is more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{P(dss)}$ and $\hat{\bar{Y}}_{Pe(dss)}$ respectively if the conditions $k < -1$, $k < -\frac{3}{2}$ and $k < -\frac{5}{4}$ holds good. It is further observed that the condition $k < -1$ is sufficient for the proposed chain *PT* estimator $\hat{\bar{Y}}^{C}_{P(dss)}$ to be more efficient than the estimators $\bar{y}_{ds}, \hat{\bar{Y}}_{P(dss)}$ and $\hat{\bar{Y}}_{PExp(dss)}$.

2.3. Chain RT exponential estimator

Inserting $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$ in place of $\bar{y}_{ds}$ in (7), we obtain the chain *RT* exponential estimator in *DSS* for population mean $\bar{Y}$ as

$$\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)} = \hat{\bar{Y}}_{\mathrm{Re}(dss)} \exp\left\{\frac{(\bar{x}'-\bar{x}_{ds})}{(\bar{x}'+\bar{x}_{ds})}\right\} = \bar{y}_{ds} \exp\left\{\frac{2(\bar{x}'-\bar{x}_{ds})}{(\bar{x}'+\bar{x}_{ds})}\right\}. \tag{26}$$

Expressing (26) in terms of ε's we have

$$\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)} = \bar{Y}(1+\varepsilon_0)\exp\left\{(\varepsilon'_1-\varepsilon_1)\left(1+\frac{(\varepsilon'_1+\varepsilon_1)}{2}\right)^{-1}\right\}. \tag{27}$$

Expanding the *RHS* of (27), multiplying out, neglecting terms of ε's having power greater than two and then subtracting $\bar{Y}$ from both sides, we have

$$\left(\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)} - \bar{Y}\right) = \bar{Y}\left[\varepsilon_0 - \varepsilon_1 + \varepsilon'_1 + \varepsilon_1^2 + \varepsilon_0\varepsilon'_1 - \varepsilon_0\varepsilon_1 - \varepsilon_1\varepsilon'_1\right]. \tag{28}$$

Taking expectation of both sides of (28) we get bias of $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ to the *fda* as

$$B\left(\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}\right) = \frac{R\theta_x}{n'\bar{X}}(1-k) \quad . \tag{29}$$

For sufficiently large n', the bias of $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ at (29) is negligible.

Squaring both sides of (28), neglecting terms of ε's having power higher than two and taking expectation of both sides, we get the *MSE* of $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ to the *fda* as

$$MSE\left(\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left\{\theta_y + R^2\theta_x(1-2k)\right\}. \tag{30}$$

From (5) and (30), it is observed that to the *fda*, the *MSE* of *RT* estimator $\hat{\bar{Y}}_{R(dss)}$ and the *MSE* of chain *RT* exponential estimator $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ are same *i.e.*, $MSE\left(\hat{\bar{Y}}_{\mathrm{Re}(dss)}\right) = MSE\left(\hat{\bar{Y}}_{R(dss)}\right)$.

2.4. Chain PT exponential estimator

On replacing $\bar{y}_{ds}$ by $\hat{\bar{Y}}_{Pe(dss)}$ in (8), we obtain the chain *PT* exponential estimator in *DSS* for population mean $\bar{Y}$ as

$$\hat{\bar{Y}}^{C}_{Pe(dss)} = \hat{\bar{Y}}_{Pe(dss)} \exp\left\{\frac{(\bar{x}_{ds}-\bar{x}')}{(\bar{x}'+\bar{x}_{ds})}\right\} = \bar{y}_{ds} \exp\left\{\frac{2(\bar{x}_{ds}-\bar{x}')}{(\bar{x}'+\bar{x}_{ds})}\right\}. \tag{31}$$

Substituting $\bar{y}_{ds} = \bar{Y}(1+\varepsilon_0), \bar{x}_{ds} = \bar{X}(1+\varepsilon_1)$ and $\bar{x}' = \bar{X}(1+\varepsilon'_1)$ in (31) we have

$$\hat{\bar{Y}}^{C}_{Pe(dss)} = \bar{Y}(1+\varepsilon_0)\exp\left\{(\varepsilon_1 - \varepsilon'_1)\left(1+\frac{(\varepsilon_1+\varepsilon'_1)}{2}\right)^{-1}\right\}. \tag{32}$$

Expanding the *RHS* of (32) neglecting terms of ε's having power greater than two and subtracting $\bar{Y}$ from both sides, we have

$$\left(\hat{\bar{Y}}^{C}_{Pe(dss)} - \bar{Y}\right) = \bar{Y}\left[\varepsilon_0 + \varepsilon_1 - \varepsilon'_1 + \varepsilon'^2_1 + \varepsilon_0\varepsilon_1 - \varepsilon_0\varepsilon'_1 - \varepsilon_1\varepsilon'_1\right]. \tag{33}$$

Taking expectation of both sides of (33) we get the bias of $\hat{\bar{Y}}^{C}_{Pe(dss)}$ to the *fda* as

$$B\left(\hat{\bar{Y}}^{C}_{Pe(dss)}\right) = \frac{\theta_{yx}}{n'\bar{X}}. \tag{34}$$

Squaring both sides of (33), retaining terms of ε's up to second degree and then taking expectation of both sides we get the *MSE* of $\hat{\bar{Y}}^{C}_{Pe(dss)}$ to the *fda* as

$$MSE\left(\hat{\bar{Y}}^{C}_{Pe(dss)}\right) = \gamma S_y^2 + \frac{1}{n'}\left[\theta_y + R^2\theta_x(1+2k)\right]. \tag{35}$$

which equals to the *MSE* of $\hat{\bar{Y}}_{P(dss)}$ *i.e.* $MSE\left(\hat{\bar{Y}}^{C}_{Pe(dss)}\right) = MSE\left(\hat{\bar{Y}}_{P(dss)}\right)$.

2.5. Chain Ratio-RT exponential estimator

Chain ratio-*RT* exponential estimator in *DSS* for population mean $\bar{Y}$ is obtained on replacing $\bar{y}_{ds}$ by $\hat{\bar{Y}}_{R(dss)}$ in (7) given by

$$\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)} = \bar{y}_{ds}\left(\frac{\bar{x}'}{\bar{x}_{ds}}\right)\exp\left(\frac{\bar{x}'-\bar{x}_{ds}}{\bar{x}'+\bar{x}_{ds}}\right). \tag{36}$$

Proceeding as earlier the bias and *MSE* of $\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}$ to the *fda*, are respectively given by

$$B\left(\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}\right) = \frac{R\theta_x}{2n'\bar{X}}\left(\frac{15}{4} - 3k\right), \tag{37}$$

$$MSE\left(\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}\right) = \left[\gamma S_y^2 + \frac{1}{n'}\left\{\theta_y + 3R^2\theta_x\left(\frac{3}{4} - k\right)\right\}\right]. \tag{38}$$

It is observed from (2), (5), (9), (16), (30) and (38) it is observed that the estimator $\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}$ is more efficient than the estimators $\bar{y}_{ds}, \left(\hat{\bar{Y}}_{R(dss)}, \hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}\right), \hat{\bar{Y}}_{\mathrm{Re}(dss)}$ and $\hat{\bar{Y}}^{C}_{R(dss)}$ respectively if

the conditions $k > \frac{3}{4}, k > \frac{5}{4}, k > 1$ and $k < \frac{7}{4}$ are satisfied. We also conclude that if k is lies between $\left(\frac{5}{4}, \frac{7}{4}\right)$, then chain ratio-*RT* exponential estimator $\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}$ performs better than the estimators $\bar{y}_{ds}, \hat{\bar{Y}}_{R(dss)}, \hat{\bar{Y}}_{\mathrm{Re}(dss)}, \hat{\bar{Y}}^{C}_{R(dss)}$ and $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$.

2.6. Chain product-PT exponential estimator

Inserting $\hat{\bar{Y}}_{P(dss)}$ in place of $\bar{y}_{ds}$ in (8), we define a chain product-*PT* exponential estimator for population mean $\bar{Y}$ in *DSS* as

$$\hat{\bar{Y}}^{C}_{PPe(dss)} = \hat{\bar{Y}}_{P(dss)} \exp\left(\frac{\bar{x}_{ds} - \bar{x}'}{\bar{x}' + \bar{x}_{ds}}\right) = \bar{y}_{ds}\left(\frac{\bar{x}_{ds}}{\bar{x}'}\right)\exp\left(\frac{\bar{x}_{ds} - \bar{x}'}{\bar{x}' + \bar{x}_{ds}}\right) \quad . \tag{39}$$

Using the procedure adopted in preceding sections, we get the bias and *MSE* of the $\hat{\bar{Y}}^{C}_{PPe(dss)}$ to the *fda*, respectively as

$$B\left(\hat{\bar{Y}}^{C}_{PPe(dss)}\right) = \frac{3R\theta_{x}}{8n'\bar{X}}(1 + 4k), \tag{40}$$

$$MSE\left(\hat{\bar{Y}}^{C}_{PPe(dss)}\right) = \left[\gamma S_{y}^{2} + \frac{1}{n'}\left\{\theta_{y} + 3R^{2}\theta_{x}\left(\frac{3}{4} + k\right)\right\}\right]. \tag{41}$$

It is observed from (2), (6), (10), (25), (35) and (41) that the proposed estimator $\hat{\bar{Y}}^{C}_{PPe(dss)}$ is better than:

(i) the unbiased estimator $\bar{y}_{ds}$ if $k < -\frac{3}{4}$,

(ii) the *PT* estimator $\hat{\bar{Y}}_{P(dss)}$ and the chain *PT* exponential estimator $\hat{\bar{Y}}^{C}_{Pe(dss)}$ if $k < -\frac{5}{4}$,

(iii) the *PT* exponential estimator $\hat{\bar{Y}}_{Pe(dss)}$ if $k < -1$; and

(iv) the chain *PT* estimator $\hat{\bar{Y}}^{C}_{P(dss)}$ if $k > -\frac{7}{4}$.

It is further observed from (i) to (iv) that the proposed estimator $\hat{\bar{Y}}^{C}_{PPe(dss)}$ is always better than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{P(dss)}$, $\hat{\bar{Y}}_{Pe(dss)}$, $\hat{\bar{Y}}^{C}_{P(dss)}$ and $\hat{\bar{Y}}^{C}_{Pe(dss)}$ if $k \in \left(-\frac{7}{4}, -\frac{5}{4}\right)$.

3. A Class of Chain Ratio-RT Exponential Estimators in DSS

We have suggested a class of chain ratio-*RT* exponential estimators for population mean $\bar{Y}$ in *DSS* as

$$\hat{\bar{Y}}^{P}_{CR(dss)} = \bar{y}_{ds}\left(\frac{\xi\,\bar{x}'+\psi}{\xi\,\bar{x}_{ds}+\psi}\right)^{\lambda}\exp\left\{\frac{\omega\xi(\bar{x}'-\bar{x}_{ds})}{\xi(\bar{x}'+\bar{x}_{ds})+2\psi}\right\}, \tag{42}$$

where $\{\xi(\neq 0),\psi\}$ are real constants or known parameters such as standard deviation S_x, coefficient of variation C_x, coefficient of skewness $\beta_1(x)$, coefficient of kurtosis $\beta_2(x)$ and $\Delta=(\beta_2(x)-\beta_1(x)-1)$ associated with auxiliary variable x or coefficient of variation C_y of study variable y or ρ_{yx}, the coefficient of correlation between y and x; (λ,ω) are suitably chosen design parameters. In particular (λ,ω) are to be determined such that *MSE* of the class of chain ratio-*RT* exponential estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$ is minimum.

We note that for different values of scalars $(\xi,\psi,\lambda,\omega)$ a large number of estimators can be generated from the suggested class of estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$.

To the *fda*, the bias and the *MSE* of the estimator $\hat{\bar{Y}}^{P}_{CR(dss)}$ are respectively given by

$$B\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)=\frac{\eta\tau R\theta_x}{2n'\bar{X}}\left\{\frac{\tau}{4}-k\right\}. \tag{43}$$

$$MSE\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)=\left[\gamma S_y^2+\frac{1}{n'}\left\{\theta_y+\frac{\tau\eta R^2\theta_x}{4}(\tau\eta-4k)\right\}\right]. \tag{44}$$

where $\tau=\dfrac{\xi\bar{X}}{\xi\bar{X}+\psi}$, $\eta=(\omega+2\lambda)$.

We note from (43) that the bias of $\hat{\bar{Y}}^{P}_{CR(dss)}$ is negligible if the sample size n' is sufficiently large. The proof of the results in (43) and (44) are simple so omitted.

The *MSE* of $\hat{\bar{Y}}^{P}_{CR(dss)}$ is minimum when

$$\eta=\frac{2k}{\tau}=\eta_{(opt)}\,\text{say}. \tag{45}$$

Thus, the resulting minimum *MSE* of $\hat{\bar{Y}}^{P}_{CR(dss)}$ is given by

$$MSE_{\min}\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)=\left[\gamma S_y^2+\frac{\theta_y}{n'}(1-\rho^2)\right]. \tag{46}$$

where $\rho=\dfrac{\theta_{yx}}{\sqrt{\theta_y\theta_x}}$.

Now we state the following theorem:

Theorem 1: Up to first order of approximation,

$$MSE\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)\geq\left[\gamma S_y^2+\frac{\theta_y}{n'}\left(1-\rho^2\right)\right]$$

with equality holding if $\eta=\frac{2k}{\tau}$.

3.1. Efficiency comparison

From (2) and (44), we have that

$$MSE\left(\hat{\bar{Y}}^{C}_{CR(dss)}\right)< MSE\left(\bar{y}_{ds}\right)\text{if}$$

either $$k>\frac{\tau\eta}{4},\tau\eta>0, \tag{47}$$

or $$k<\frac{\tau\eta}{4},\tau\eta<0, \tag{48}$$

If we set $(\lambda,\omega)=(\lambda^*,0)$ then the class of estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$ reduces to

$$\hat{\bar{Y}}^{P1}_{CR(dss)}=\bar{y}_{ds}\left(\frac{\xi\bar{x}'+\psi}{\xi\bar{x}_{ds}+\psi}\right)^{\lambda^*} \tag{49}$$

where λ^* is a constant.

Putting $(\lambda,\omega)=(\lambda^*,0)$in (43) and (44) we get the bias and *MSE* of $\hat{\bar{Y}}^{P1}_{CR(dss)}$ to the *fda*, are respectively given by

$$B\left(\hat{\bar{Y}}^{P1}_{CR(dss)}\right)=\frac{\tau R\lambda^*\theta_x}{n'\bar{X}}\left\{\frac{\tau\left(\lambda^*+1\right)}{2}-k\right\}, \tag{50}$$

$$MSE\left(\hat{\bar{Y}}^{P1}_{CR(dss)}\right)=\left[\gamma S_y^2+\frac{1}{n'}\left\{\theta_y+\tau R^2\lambda^*\theta_x\left(\tau\lambda^*-2k\right)\right\}\right]. \tag{51}$$

Now from (44) and (51), we have

$$MSE\left(\hat{\bar{Y}}^{P1}_{CR(dss)}\right)-MSE\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)=\frac{\tau R^2\theta_x}{n'}\left(2\lambda^*-\eta\right)\left[\frac{\tau\left(2\lambda^*+\eta\right)}{4}-k\right]>0,\text{ if}$$

either $$k<\frac{\tau\left(2\lambda^*+\eta\right)}{4},\eta<2\lambda^*, \tag{52}$$

or $$k>\frac{\tau\left(2\lambda^*+\eta\right)}{4},\eta>2\lambda^*. \tag{53}$$

If we set $(\lambda,\omega)=(0,\omega^*)$, then the class of estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$ reduces to

$$\hat{\bar{Y}}^{P2}_{CR(dss)}=\bar{y}_{ds}\exp\left\{\frac{\omega^*\xi(\bar{x}'-\bar{x}_{ds})}{\xi(\bar{x}'+\bar{x}_{ds})+2\psi}\right\}, \tag{54}$$

where ω^* is a constant.

Putting $(\lambda,\omega)=(0,\omega^*)$in (43) and (44) we get bias and *MSE* of the estimator $\hat{\bar{Y}}^{P2}_{CR(dss)}$ respectively as

$$B\left(\hat{\bar{Y}}^{P2}_{CR(dss)}\right)=\frac{\tau\omega^* R\theta_x}{2n'\bar{X}}\left\{\frac{\tau(\omega^*+2)}{4}-k\right\}, \tag{55}$$

$$MSE\left(\hat{\bar{Y}}^{P2}_{CR(dss)}\right)=\left[\gamma S_y^2+\frac{1}{n'}\left\{\theta_y+\frac{\tau R^2\omega^*\theta_x}{4}(\tau\omega^*-4k)\right\}\right]. \tag{56}$$

From (44) and (56), we have that

$$MSE\left(\hat{\bar{Y}}^{P}_{CR(dss)}\right)<MSE\left(\hat{\bar{Y}}^{P2}_{CR(dss)}\right), \text{ if}$$

either $$k<\frac{\tau(\omega^*+\eta)}{4},\eta<\omega^*, \tag{57}$$

or $$k>\frac{\tau(\omega^*+\eta)}{4},\eta>\omega^*. \tag{58}$$

Thus we conclude that $\hat{\bar{Y}}^{P}_{CR(dss)}$ is better than $\bar{y}_{ds}$, $\hat{\bar{Y}}^{P1}_{CR(dss)}$ and $\hat{\bar{Y}}^{P2}_{CR(dss)}$ if the conditions given in equations (47) or (48), (52) or (53) and (57) or (58) respectively are satisfied.

4. Improved Class of Estimators

Motivated by Searls (1964) we consider an improved class of chain *RT* estimators in *DSS* for population mean $\bar{Y}$ as

$$P^I_{(dss)}=\alpha_2\,\hat{\bar{Y}}^{P}_{CR(dss)}$$

$$=\alpha_2\,\bar{y}_{ds}\left(\frac{\xi\bar{x}'+\psi}{\xi\bar{x}_{ds}+\psi}\right)^{\lambda}\exp\left\{\frac{\omega\xi(\bar{x}'-\bar{x}_{ds})}{\xi(\bar{x}'+\bar{x}_{ds})+2\psi}\right\} \tag{59}$$

where α_2 is a constant to be determined such that *MSE* of $P^I_{(dss)}$ is minimum.

To the *fda*, the bias and *MSE* of $P^I_{(dss)}$ are respectively given by

$$B\left(P^I_{(dss)}\right)=\bar{Y}(\alpha_2\theta_5-1), \tag{60}$$

$$MSE\left(P^I_{(dss)}\right)=\bar{Y}^2\left[1+\alpha_2^2\theta_2-2\alpha_2\theta_5\right]. \tag{61}$$

where

$$\theta_2 = \left[1+\frac{1}{\bar{Y}^2}\left\{\gamma S_y^2+\frac{1}{n'}\left(\theta_y+\frac{\tau\eta R^2 A_x}{2}\left[\tau(\eta+1)-4k\right]\right)\right\}\right],$$

$$\theta_5 = \left[1+\frac{\tau\eta\theta_x}{8n'\bar{X}^2}\left\{\tau(\eta+2)-4k\right\}\right].$$

The $MSE\left(P_{(dss)}^{I}\right)$ at (61) is minimized for

$$\alpha = \frac{\theta_5}{\theta_2} = \alpha_{2(opt)}, \text{ say.} \tag{62}$$

Thus, the resulting minimum *MSE* of $P_{(dss)}^{I}$ is given by

$$MSE_{\min}\left(P_{(dss)}^{I}\right) = \bar{Y}^2\left[1-\frac{\theta_5^2}{\theta_2}\right]. \tag{63}$$

Now we arrived at the following theorem.

Theorem 2: Up to terms of order $O\left(n^{-1}\right)$,

$$MSE\left(P_{(dss)}^{I}\right) \geq \bar{Y}^2\left[1-\frac{\theta_5^2}{\theta_2}\right]$$

with equality holding if

$$\alpha_2 = \frac{\theta_5}{\theta_2}.$$

For comparing $\hat{\bar{Y}}_{CR(dss)}^{P}$ with $P_{(dss)}^{I}$, we express (44) in terms of θ_2 and θ_5 as

$$MSE\left(\hat{\bar{Y}}_{CR(dss)}^{P}\right) = \bar{Y}^2\left[1+\theta_2-2\theta_5\right]. \tag{64}$$

Now from (63) and (64), we have

$$MSE\left(\hat{\bar{Y}}_{CR(dss)}^{P}\right) - MSE_{\min}\left(P_{(dss)}^{I}\right) = \bar{Y}^2\,\frac{\left(\theta_2-\theta_5\right)^2}{\theta_2} \geq 0,$$

which follows that

$$MSE_{\min}\left(P_{(dss)}^{I}\right) \leq MSE\left(\hat{\bar{Y}}_{CR(dss)}^{P}\right). \tag{65}$$

Thus from (65) we conclude that the improved estimator $P_{(dss)}^{I}$ is more efficient than the suggested estimator $\hat{\bar{Y}}_{CR(dss)}^{P}$ under the optimum condition (62). Also, the estimators belonging

to the improved family of estimators $P_{(dss)}^{I}$ are more efficient as compared to the estimators belonging to suggested class of estimators $\hat{\bar{Y}}_{CR(dss)}^{P}$.

5. A General Class of Estimators

Following the procedure adopted by Upadhyaya *et al*. (1985), we define a generalized class of estimators in *DSS* for population mean $\bar{Y}$ as

$$P_{(dss)}^{G} = \alpha_1 \bar{y}_{ds} + \alpha_2 \bar{y}_{ds} \left(\frac{\xi \bar{x}' + \psi}{\xi \bar{x}_{ds} + \psi} \right)^{\lambda} \exp\left\{ \frac{\omega \xi (\bar{x}' - \bar{x}_{ds})}{\xi (\bar{x}' + \bar{x}_{ds}) + 2\psi} \right\}, \tag{66}$$

where (α_1, α_2) are constants to be determined such that the *MSE* of $P_{(dss)}^{G}$ is minimum; and the scalars $(\xi, \psi, \omega, \lambda)$ are same as defined earlier.

To the *fda*, the bias and *MSE* of the generalized class of estimators $P_{(dss)}^{G}$ are respectively given by

$$B\left(P_{(dss)}^{G}\right) = \bar{Y}\left[\alpha_1 \theta_4 + \alpha_2 \theta_5 - 1\right], \tag{67}$$

$$MSE\left(P_{(dss)}^{G}\right) = \bar{Y}^2\left[1 + \alpha_1^2 \theta_1 + \alpha_2^2 \theta_2 + 2\alpha_1 \alpha_2 \theta_3 - 2\alpha_1 \theta_4 - 2\alpha_2 \theta_5\right], \tag{68}$$

where

$$\theta_1 = \left[1 + \frac{1}{\bar{Y}^2}\left(\gamma S_y^2 + \frac{1}{n'}\theta_y\right)\right],$$

$$\theta_3 = \left[1 + \frac{1}{\bar{Y}^2}\left(\gamma S_y^2 + \frac{1}{n'}\theta_y\right) + \frac{\tau \eta \theta_x}{8n' \bar{X}^2}\{(\eta + 2)\tau - 8k\}\right],$$

$$\theta_4 = 1.$$

θ_2 and θ_5 are same as defined previously.

Minimization of (68) with respect to (α_1, α_2) gives

$$\begin{bmatrix} \theta_1 & \theta_3 \\ \theta_3 & \theta_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \theta_4 \\ \theta_5 \end{bmatrix}. \tag{69}$$

After simplification of (69), we get the optimum values of (α_1, α_2) respectively as

$$\left.\begin{aligned}\alpha_{1(opt)} &= \frac{\Delta_1}{\Delta_0}\\ \alpha_{2(opt)} &= \frac{\Delta_2}{\Delta_0}\end{aligned}\right\}, \tag{70}$$

where

$$\Delta_0 = \begin{vmatrix}\theta_1 & \theta_3\\ \theta_3 & \theta_2\end{vmatrix} = \left(\theta_1\theta_2 - \theta_3^2\right),$$

$$\Delta_1 = \begin{vmatrix}\theta_4 & \theta_3\\ \theta_5 & \theta_2\end{vmatrix} = \left(\theta_2\theta_4 - \theta_3\theta_5\right),$$

$$\Delta_2 = \begin{vmatrix}\theta_1 & \theta_4\\ \theta_3 & \theta_5\end{vmatrix} = \left(\theta_1\theta_5 - \theta_3\theta_4\right).$$

Thus the resulting minimum *MSE* of $P_{(dss)}^{G}$ is given by

$$MSE_{\min}\left(P_{(dss)}^{G}\right) = \bar{Y}^2\left[1 - \frac{\left\{\theta_2\theta_4^2 - 2\theta_3\theta_4\theta_5 + \theta_1\theta_5^2\right\}}{\left(\theta_1\theta_2 - \theta_3^2\right)}\right]. \tag{71}$$

Theorem 3: Up to terms of order $O\left(n^{-1}\right)$,

$$MSE\left(P_{(dss)}^{G}\right) \geq \bar{Y}^2\left[1 - \frac{\left\{\theta_2\theta_4^2 - 2\theta_3\theta_4\theta_5 + \theta_1\theta_5^2\right\}}{\left(\theta_1\theta_2 - \theta_3^2\right)}\right]$$

with equality if

$$\alpha_{1(opt)} = \frac{\Delta_1}{\Delta_0}, \alpha_{2(opt)} = \frac{\Delta_2}{\Delta_0}.$$

Now from (63) and (71), we have

$$MSE_{\min}\left(P_{(dss)}^{I}\right) - MSE_{\min}\left(P_{(dss)}^{G}\right) = \bar{Y}^2\frac{\left(\theta_2\theta_4 - \theta_3\theta_5\right)^2}{\theta_2\left(\theta_1\theta_2 - \theta_3^2\right)} \geq 0$$

which follows that

$$MSE_{\min}\left(P_{(dss)}^{G}\right) \leq MSE\left(P_{(dss)}^{I}\right). \tag{72}$$

Combining the inequalities (65) and (72) we have

$$MSE_{\min}\left(P_{(dss)}^{G}\right) \leq MSE\left(P_{(dss)}^{I}\right) \leq MSE\left(\hat{\bar{Y}}_{CR(dss)}^{P}\right). \tag{73}$$

From (73), we infer that the generalized estimator $P_{(dss)}^{G}$ is more efficient than the suggested improved estimator $P_{(dss)}^{G}$ and the estimator $\hat{\bar{Y}}_{CR(dss)}^{P}$.

6. Numerical Illustration

To demonstrate the performance of the various estimators of the population mean $\bar{Y}$ of y, we have taken two data sets. Description of the population data sets are given below.

Data 1 [Source: Chouhan (2012)]
y: Productivity (MT/Hectare), x: Production in '000 Tons
N=20, n=8, $n_1 = 4, n_2 = 4, n'_1 = 7, n'_2 = 7, N_1 = 10, N_2 = 10, \bar{Y}_1 = 1.70, \bar{Y}_2 = 3.67, \bar{Y} = 2.685,$
$\bar{X}_1 = 10.41, \bar{X}_2 = 289.14, \bar{X} = 149.705, S_{x1} = 3.53, S_{x2} = 111.61, S_{y1} = 0.50, S_{y2} = 1.41, S_{yx1} = 1.60,$
$S_{yx2} = 144.87, S_y^2 = 2.20, R = 0.018.$

Data 2 [Source: Murthy (1967), p228]
y: Output, x: Fixed capital,
N=10, n=4, $n_1 = 2, n_2 = 2, n'_1 = 4, n'_2 = 4, N_1 = 5, N_2 = 5, \bar{Y}_1 = 1925.8, \bar{Y}_2 = 3115.6, \bar{Y} = 1260.35$
$\bar{X}_1 = 214.4, \bar{X}_2 = 333.8, \bar{X} = 137.05, S_{x1} = 74.87, S_{x2} = 66.35, S_{y1} = 615.92, S_{y2} = 340.38,$
$S_{yx1} = 39360.68, S_{yx2} = 22356.50, S_y^2 = 668351.00, R = 9.196.$

We have computed the percent relative efficiencies (*PRE*s) of estimators $\hat{\bar{Y}}_{R(dss)}, \hat{\bar{Y}}_{\mathrm{Re}(dss)}$, $\hat{\bar{Y}}_{R(dss)}^{C}, \hat{\bar{Y}}_{\mathrm{Re}(dss)}^{C}, \hat{\bar{Y}}_{R\mathrm{Re}(dss)}^{C}$ with respect to usual unbiased estimator $\bar{y}_{ds}$ by using the following formulae:

$$PRE\left(\hat{\bar{Y}}_{R(dss)}, \bar{y}_{ds}\right) = PRE\left(\hat{\bar{Y}}_{\mathrm{Re}(dss)}^{C}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\left[\gamma S_y^2 + \left(\frac{1}{n'}\right)\left\{\theta_y + R^2\theta_x(1-2k)\right\}\right]} * 100, \qquad (74)$$

$$PRE\left(\hat{\bar{Y}}_{\mathrm{Re}(dss)}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\left[\gamma S_y^2 + \left(\frac{1}{n'}\right)\left\{\theta_y + \frac{R^2\theta_x}{4}(1-4k)\right\}\right]} * 100, \qquad (75)$$

$$PRE\left(\hat{\bar{Y}}_{R(dss)}^{C}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\left[\gamma S_y^2 + \left(\frac{1}{n'}\right)\left\{\theta_y + 4R^2\theta_x(1-k)\right\}\right]} * 100, \qquad (76)$$

$$PRE\left(\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\left[\gamma S_y^2 + \left(\frac{1}{n'}\right)\left\{\theta_y + 3R^2\theta_x(0.75-k)\right\}\right]} * 100, \tag{77}$$

and findings are displayed in Table 1.

The *PRE*s of proposed class of estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$, improved class of estimators $P^{I}_{(dss)}$ and generalized class of estimators $P^{G}_{(dss)}$ with respect to $\bar{y}_{ds}$ have been computed by using the formulae:

$$PRE\left(\hat{\bar{Y}}^{P}_{CR(dss)}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\bar{Y}^2\left[1 + \theta_2 - 2\theta_5\right]} * 100, \tag{78}$$

$$PRE\left(P^{I}_{(dss)}, \bar{y}_{ds}\right) = \frac{\theta_2\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\bar{Y}^2\left(\theta_2 - \theta_5^2\right)} * 100, \tag{79}$$

$$PRE\left(P^{G}_{(dss)}, \bar{y}_{ds}\right) = \frac{\left\{\gamma S_y^2 + \left(\frac{\theta_y}{n'}\right)\right\}}{\bar{Y}^2\left[1 - \frac{\left(\theta_2\theta_4^2 - 2\theta_3\theta_4\theta_5 + \theta_1\theta_5^2\right)}{\left(\theta_1\theta_2 - \theta_3^2\right)}\right]} * 100 \tag{80}$$

for different values of $(\xi, \psi, \lambda, \omega)$. Results are shown in Table 2.

Table 1: *PRE* of Different estimators with respect to $\bar{y}_{ds}$

Estimators	Data 1	Data 2
	PRE	***PRE***
$\bar{y}_{ds}$	100	100
$\hat{\bar{Y}}_{R(dss)}$ or $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$	145.18	138.96
$\hat{\bar{Y}}_{\mathrm{Re}(dss)}$	168.40	152.82
$\hat{\bar{Y}}^{C}_{R(dss)}$	41.98	48.01
$\hat{\bar{Y}}^{C}_{R\,\mathrm{Re}(dss)}$	77.83	83.62

Table 2: *PRE*s of $\hat{\bar{Y}}^{P}_{CR(dss)}$, $P^{I}_{(dss)}$ and $P^{G}_{(dss)}$ with respect to $\bar{y}_{ds}$ for different values of scalars $(\xi,\psi,\lambda,\omega)$

Values of scalars				Data 1			Data 2		
ξ	ψ	λ	ω	$PRE\ \hat{\bar{Y}}^{P}_{CR(dss)}$	$PRE\ P^{I}_{(dss)}$	$PRE\ P^{G}_{(dss)}$	$PRE\ \hat{\bar{Y}}^{P}_{CR(dss)}$	$PRE\ P^{I}_{(dss)}$	$PRE\ P^{G}_{(dss)}$
1	0	1	–1	168.4	170.13	177.37	152.82	155.05	160.94
1	0	0.75	–0.75	154.23	155.77	177.14	142.40	144.46	160.70
1	0	0.5	–0.5	136.19	137.64	176.92	128.90	130.87	160.49
1	0	0.25	–0.25	117.44	118.89	176.73	114.28	116.25	160.28
1	0	1	0	145.18	148.54	178.45	138.96	142.81	161.99
1	0	0.25	0.25	154.23	155.77	177.14	142.40	144.46	160.70
1	0	0.5	0.5	172.81	175.27	177.88	157.32	160.24	161.44
1	0	0.75	0.75	126.45	130.17	178.76	124.98	129.27	162.29
1	0	0.75	1	121.21	112.2	179.09	110.33	115.00	162.60
*PRE*s at optimum value				175.54	177.66	177.71	158.62	161.26	161.27

It is observed from Table 1 that the estimators $\hat{\bar{Y}}_{R(dss)}$ and $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$ are equally efficient. The estimators $\hat{\bar{Y}}^{C}_{R(dss)}$, $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ and $\hat{\bar{Y}}^{C}_{R\mathrm{Re}(dss)}$ are more efficient than the conventional unbiased estimator $\bar{y}_{ds}$ with considerable gain in efficiency. The chain estimators $\hat{\bar{Y}}^{C}_{R(dss)}$ and $\hat{\bar{Y}}^{C}_{R\mathrm{Re}(dss)}$ are even inferior to the unbiased estimator $\bar{y}_{ds}$.

Table 2 shows that the proposed estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$, $P^{I}_{(dss)}$ and $P^{G}_{(dss)}$ are more efficient than the usual unbiased estimator $\bar{y}_{ds}$ for selected values of $(\xi,\psi,\lambda,\omega)$. The proposed chain estimator $\hat{\bar{Y}}^{P}_{CR(dss)}$ (at optimum value of η) is more efficient than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(dss)}$, $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$, $\hat{\bar{Y}}^{C}_{R(dss)}$, $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ and $\hat{\bar{Y}}^{C}_{R\mathrm{Re}(dss)}$.

Table 2 also exhibits that the generalized estimator $P^{G}_{(dss)}$ is the best (in the sense of having least *MSE*) among all the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(dss)}$, $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$, $\hat{\bar{Y}}^{C}_{R(dss)}$, $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$, $\hat{\bar{Y}}^{C}_{R\mathrm{Re}(dss)}$, $\hat{\bar{Y}}^{P}_{CR(dss)}$ and $P^{I}_{(dss)}$ discussed here. Thus there is enough scope of selecting the scalars $(\xi,\psi,\lambda,\omega)$ involved in the proposed class of estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$, $P^{I}_{(dss)}$ and $P^{G}_{(dss)}$ obtaining better than the estimators $\bar{y}_{ds}$, $\hat{\bar{Y}}_{R(dss)}$, $\hat{\bar{Y}}_{\mathrm{Re}(dss)}$, $\hat{\bar{Y}}^{C}_{R(dss)}$, $\hat{\bar{Y}}^{C}_{\mathrm{Re}(dss)}$ and $\hat{\bar{Y}}^{C}_{R\mathrm{Re}(dss)}$.

So the proposed estimators $\hat{\bar{Y}}^{P}_{CR(dss)}$, $P^{I}_{(dss)}$ and $P^{G}_{(dss)}$ are recommended for their use in practice.

Acknowledgement

Authors are grateful to Professor Vinod Gupta, Chair Editor, Statistics and Applications and the learned referee for their valuable suggestions regarding improvement of the paper.

References

Bahl, S. and Tuteja, R. K. (1991). Ratio and product-type exponential estimator. *Journal of Information and Optimization Sciences*, **12**(**1**), 159-164.

Chouhan, S. (2012). Improved estimation of parameters using auxiliary information in sample surveys. *Unpublished Ph.D. Thesis, Vikram University, Ujjain, M.P. India*.

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

Holt, D. and Smith, T. M. F. (1979). Post Stratification. *Journal of Royal Statistical Society*, **A142**, 33-46.

Ige, A. F. and Tripathi, T. P. (1987). On doubling for stratification and use of auxiliary information. *Journal of the Indian Society of Agricultural Statistics*, **39**, 191-201.

Murthy, M. N. (1967). *Sampling Theory and Method*. Statistical Publishing Society, Calcutta, India, 228.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, **6**(**1**), 125-133.

Searls, D. T. (1964). The utilization of known coefficient of variation in the estimation procedure. *Journal of the American Statistical Association*, **59**, 1225-1226.

Tailor, R., Chouhan, S. and Kim, J. M. (2014). Ratio and product type exponential estimators of population mean in double sampling for stratification. *Communications for Statistical Applications and Methods,* **21**(**1**), 1-9.

Tripathi, T. P. and Bahl, S. (1991). Estimation of mean using double sampling for stratification and multivariate auxiliary information. *Communications in Statistics-Theory and Methods*, **20**(**9**), 2589-2602.

Upadhyaya, L. N., Singh, H. P. and Vos, J. W. E. (1985). On the estimation of population means and ratio using supplementary information. *Statistica Neerlandica*, **39**(**3**), 309-318.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 213-227

Cost and Profit Analysis of State-dependent Feedback Queue with Impatient Customer Subject to Catastrophes

Ankita Roy Chowdhury and Indra
Department of Statistics and Operational Research,
Kurukshetra University, Kurukshetra, Haryana-136119, India

Received: 29 January 2021; Revised: 21 March 2021; Accepted: 23 April 2021

Abstract

The paper analyses a single-server Markovian queueing system having state dependent service rates with customer's balking and feedback subject to catastrophes. Using matrix-geometric solution method, we obtain steady-state solution for the system. Various measurable indicators have been evaluated with the assistance of Maple software and based on these measures; we have presented an expected cost and profit analysis.

Key words: Matrix-geometric method; Balking; Feedback; Catastrophes; State-dependent.

1. Introduction

In such an intense market condition, where attracting and finding a potential customer is very difficult, no one wants to bear the cost of customer loss. So, providing quality of service at faster rate is very important factor in this fast paced life. Considering parameters like balking, feedback and state-dependent service rate provides more pliability for optimal design and finds its applicability in communication network, production system, and in various congestion problems. Queueing analysis presents an optimal solution by providing suitable suggestions to reduce congestion.

In this paper, we have considered the parameters; balking, feedback, catastrophe and state dependent service altogether, to analyse the system performance. These parameters inordinately affect the system and its cost function.

There are many practical situations, where service rates depend on the size of the system. Such situations can be seen in hotels or restaurants during rush hours, where waiters and cooks work with a faster rate to cope up with the demand, or in hospitals for patients coming to the emergency ward *etc*. Many authors have contributed in the study of state dependent service rates. Davingon and Disney (1976) considered single server state dependent feedback queue. Doshi and Jangerman (1986) obtained some important performance measures for an M/G/1 queue where, balking depended on system size using supplementary variable technique. Abou-El-Ata (1991) extended the model of Ancker and Gafarian (1963) to study the state dependent finite queue with impatient customers. Such system may also get affected by balking, where customer doesn't want to join the system due to long queue.

Corresponding Author: Indra
E-mail: indra@kuk.ac.in

Queueing systems incorporating balking, feedback or both have attracted many researchers. They are useful in designing and managing systems like transmission of data, emergency ward of health sector where balking is common and chances of rework is more. The perception of customer impatience was first appeared in the work of Haight (1957). Tackas (1963) analysed a single server queue with feedback. For conceivable uses, the history and contributions of researchers on queueing systems with balking and feedback, one may see articles by Santhakumaran and Thangaraj (2000), Choudhury and Paul (2005), Kumar *et al.* (2013), Varalakshmi *et al.* (2018), Bouchentouf *et al.* (2019).

Queueing models with catastrophes have gained importance during last few decades because of its relevancy in many area *viz.* computers and telecommunications, health sector, production sector, disaster management. Queues with catastrophes have attracted many researchers due to the fact; they are very unpredictable in nature and force the customers to leave the system immediately. So, including them in modelling makes the model more pragmatic. Thangaraj and Vanitha (2009) obtained transient solution of M/M/1 feedback queue with catastrophe using continued fractions. Kumar *et al.* (2014) studied queueing systems subjected to catastrophes and customer's impatience and obtained time-dependent and steady-state probabilities when system is operational as well when under repair process. Bura and Bura (2015) analysed finite, single-server markovian-catastrophic queueing system with restorative effects.

The primary objectives of this paper are:

i. To obtain steady state solutions to aforesaid queueing system using matrix geometric method.
ii. To evaluate important performance measures such as mean number of customers in the system and in the queue, probability of ideal, probability of busy, mean balking rate *etc.* and to perform sensitivity analysis.
iii. To formulate an expected cost and profit functions based on measures obtained.
iv. Graphical representations showing effect of different parameters on expected cost and expected profit functions.

2. Model Assumptions and Descriptions

We consider a markovian queueing system of infinite capacity, where the arrivals and departures both follow Poisson process with mean inter-arrival time $\frac{1}{\lambda}$ and mean inter-service times $\frac{1}{\mu_1}$ or $\frac{1}{\mu_2}$ depending upon the system size. Arriving customer may join the queue with probability 'β' if he finds the server non-empty or balk with probability $'1-\beta'$ according to some predetermined norms. The server decides to operate with two different service rates; 'slow and fast' subjected to the length of the queue. If it finds the system size is less than or equal to the critical value 'r', it serves with a slower rate 'μ_1'; otherwise with a faster rate 'μ_2'. If customer, on service completion is satisfied by the service, the customer leaves the system with probability 'q'. On contrary the customer re-joins the queue with probability 'p' if one finds the service dissatisfactory. Occurrence of catastrophes ejects all the customers from the system instantly and system becomes inactive momentarily. Catastrophes occur according to Poisson process with rate

of occurrence′ ξ', when the system is non-empty.

The infinitesimal generator matrix Q of the system is given by:

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ (q\mu_1+\xi) & -(q\mu_1+\beta\lambda+\xi) & \beta\lambda & \dots & \dots & \dots & \dots & \dots & \dots \\ \xi & q\mu_1 & -(q\mu_1+\beta\lambda+\xi) & \dots & \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots \\ \xi & 0 & 0 & q\mu_1 & -(q\mu_1+\beta\lambda+\xi) & \beta\lambda & 0 & \dots & \dots \\ \xi & 0 & 0 & \dots & q\mu_1 & -(q\mu_1+\beta\lambda+\xi) & \beta\lambda & \dots & \dots \\ \xi & 0 & 0 & \dots & \dots & q\mu_2 & -(q\mu_2+\beta\lambda+\xi) & \beta\lambda & \dots \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \dots \end{pmatrix} \begin{matrix} 0 \\ 1 \\ 2 \\ \vdots \\ \dots \\ r-1 \\ r \\ r+1 \\ \dots \\ \dots \end{matrix}$$

Let *n(t)* ≡ number of customers in the system at time '*t*'. Let '*n*' be the stationary random variable for the number of the customers in the system. We define π_i=$\{n = i\}$= $\lim_{t\to\infty} P\{n(t) = i\}$, where $i \in \mathbb{W}$ and π_i represents the stationary probability of *i* customers in the system. The stationary probability vector is given by,

$$\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \dots\dots\dots\dots\dots, \pi_r, \pi_{r+1}, \dots\dots\dots) \tag{1}$$

The steady-state probabilities π_i are related geometrically to each other as $\pi_i = \pi_r R^{i-r}\ \forall i \geq r$. Here, R is called the rate element and for this system it is given by:

$$R = \frac{(\beta\lambda + q\mu_2 + \xi) - \sqrt{(\beta\lambda + q\mu_2 + \xi)^2 - 4\beta\lambda q\mu_2}}{2q\mu_2} \tag{2}$$

The steady-state probabilities are obtained by solving the following equations

$$\boldsymbol{\pi Q} = \mathbf{0} \tag{3}$$

$$\boldsymbol{\pi e} = \mathbf{1} \tag{4}$$

3. Performance Measures

We calculate some performance indicators using the probabilities; obtained by employing equation (3) and equation (4), for the system as follows.

i) "Expected number of customers in the system:"

$$\boldsymbol{MNS} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{r}} \boldsymbol{n\pi_n} + \sum_{\boldsymbol{n=r}}^{\infty} \boldsymbol{n\pi_r R^{n-r}} \tag{5}$$

ii) "Expected number of customers in the queue:"

$$\boldsymbol{MNQ} = \sum_{n=1}^{r-1} \boldsymbol{n\pi_{n+1}} + \sum_{n=r}^{\infty} \boldsymbol{n\pi_r R^{n+1-r}} \tag{6}$$

iii) Mean Balking Rate (B.R):

$$\boldsymbol{B.R} = (\mathbf{1} - \boldsymbol{\beta})\boldsymbol{\lambda}(\mathbf{1} - \boldsymbol{\pi_0}) \tag{7}$$

iv) Probability that the server is busy:

$$\boldsymbol{P_b} = (\mathbf{1} - \boldsymbol{\pi_0}) \tag{8}$$

v) Probability that the server is ideal:

$$\boldsymbol{P_I} = \boldsymbol{\pi_0} \tag{9}$$

vi) Expected waiting time in the system:

$$\boldsymbol{MWS} = \frac{\boldsymbol{MNS}}{\boldsymbol{\lambda}} \tag{10}$$

vii) Expected waiting time in the queue:

$$\boldsymbol{MWQ} = \boldsymbol{MWS} - \frac{\mathbf{1}}{\boldsymbol{\mu_2}} \tag{11}$$

Special Case

If we put $\beta = 1, q = 1, \xi = 0$ and consider only one service rate throughout *i.e.* μ, then the rate element reduces to

$$\boldsymbol{R} = \frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}$$

and $\boldsymbol{\pi_n}$is given by:

$$\boldsymbol{\pi_n} = \boldsymbol{R^n}(\mathbf{1} - \boldsymbol{R})$$

which is same as the probability of *n* customers in the system, for classical M/M/1 queue.

Particular Cases

We obtain stationary probabilities when *r* = 1 and *r* =2 in the following section.

Case-I: When *r* = 1

The infinitesimal generator matrix Q of the system is given by:

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & \dots & \dots \\ (q\mu_1 + \xi) & -(q\mu_1 + \beta\lambda + \xi) & \beta\lambda & 0 & \dots & \dots \\ \xi & q\mu_2 & -(q\mu_2 + \beta\lambda + \xi) & \beta\lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \dots & \dots & \dots \end{pmatrix}$$

Using (3) and (4) we have,

$$\pi_0 = \frac{q\mu_1(1-R)+\xi}{q\mu_1(1-R)+\lambda+\xi} \tag{12}$$

$$\pi_1 = \frac{\lambda(1-R)}{q\mu_1(1-R)+\lambda+\xi} \tag{13}$$

The other steady state probabilities are obtained by $\pi_i = \pi_1 R^{i-1}\ \forall\ i \geq 2$

Case-II: When *r* = 2

The infinitesimal generator matrix Q of the system is given by:

$$\boldsymbol{Q} = \begin{pmatrix} -\boldsymbol{\lambda} & \boldsymbol{\lambda} & \boldsymbol{0} & \dots & \dots & \dots \\ (\boldsymbol{q\mu_1}+\boldsymbol{\xi}) & -(\boldsymbol{q\mu_1}+\boldsymbol{\beta\lambda}+\boldsymbol{\xi}) & \boldsymbol{\beta\lambda} & \dots & \dots & \dots \\ \boldsymbol{\xi} & \boldsymbol{q\mu_1} & -(\boldsymbol{q\mu_1}+\boldsymbol{\beta\lambda}+\boldsymbol{\xi}) & \dots & \dots & \dots \\ \boldsymbol{\xi} & \boldsymbol{0} & \boldsymbol{q\mu_2} & -(\boldsymbol{q\mu_2}+\boldsymbol{\beta\lambda}+\boldsymbol{\xi}) & \boldsymbol{\beta\lambda} & \dots \\ \vdots & \vdots & \vdots & \dots & \dots & \dots \end{pmatrix}$$

Using (3) and (4) we have,

$$\pi_0 = \frac{(q\mu_1+\xi)(q\mu_1+\beta\lambda+\xi-q\mu_2R)+\beta\lambda\xi(1-R)^{-1}}{(q\mu_1+\xi+\lambda)(q\mu_1+\beta\lambda+\xi-q\mu_2R)+\beta\lambda\xi(1-R)^{-1}+\beta\lambda^2(1-R)^{-1}} \tag{14}$$

$$\pi_1 = \frac{(q\mu_1+\beta\lambda+\xi-q\mu_2R)\lambda}{(q\mu_1+\xi+\lambda)(q\mu_1+\beta\lambda+\xi-q\mu_2R)+\beta\lambda\xi(1-R)^{-1}+\beta\lambda^2(1-R)^{-1}} \tag{15}$$

$$\pi_2 = \frac{\beta\lambda^2}{(q\mu_1+\xi+\lambda)(q\mu_1+\beta\lambda+\xi-q\mu_2R)+\beta\lambda\xi(1-R)^{-1}+\beta\lambda^2(1-R)^{-1}} \tag{16}$$

The remaining probabilities are obtained by $\pi_i = \pi_2 R^{i-2}\ \forall\ i \geq 3$.

4. Cost Model and Profit Model

Constructing an expected cost function for a system which not only get affected by varying arrival and service rates but also by balking, feedback, and catastrophes is very difficult. Here, we confine ourselves in determining the optimum value of '*r*' which minimizes the cost. Let C_1 be the cost associated with a customer present in the queue, C_2 be the cost associated with a customer when server is busy, C_3 be the cost associated with a customer loss, and C_4 be the cost associated with server when it is ideal. So, we have the expected cost function as,

$$\text{Total Expected Cost } (\mathbf{TEC}) = \boldsymbol{C_1} * \mathbf{MNQ} + \boldsymbol{C_2} * \boldsymbol{P_b} + \boldsymbol{C_3} * \mathbf{B.R} + \boldsymbol{C_4} * \boldsymbol{P_I} \quad (17)$$

Similarly, for an expected profit function, we have

$$\text{Total Expected Profit } (\mathbf{TEP}) = \boldsymbol{\rho} * \mathbf{MNS} - \mathbf{TEC} \quad (18)$$

where $\boldsymbol{\rho}$ is the revenue.

Though the cost function may appeared to be simple but it is highly non-linear and complex in nature which makes it difficult in optimizing the value of '*r*'. In order to arrive at a decision, we carry out sensitivity analyses by substituting different values for the parameters.

5. Sensitivity Analysis

Sensitivity analyses have been performed to compare the systems $r = 1$ and $r = 2$, by changing values of the parameters involved. For calculation, let C_1 =100, C_2 =150, C_3 =200, and C_4 =250. The measurable indicators are computed coupled with total expected cost and total expected profit. These measures have guided in deciding the optimal value of '*r*' in order to minimize its expected cost and maximize the expected profit. Different Cost and profit graphs have been plotted by varying the parameters under consideration. These graphs are illustrated and discussed below.

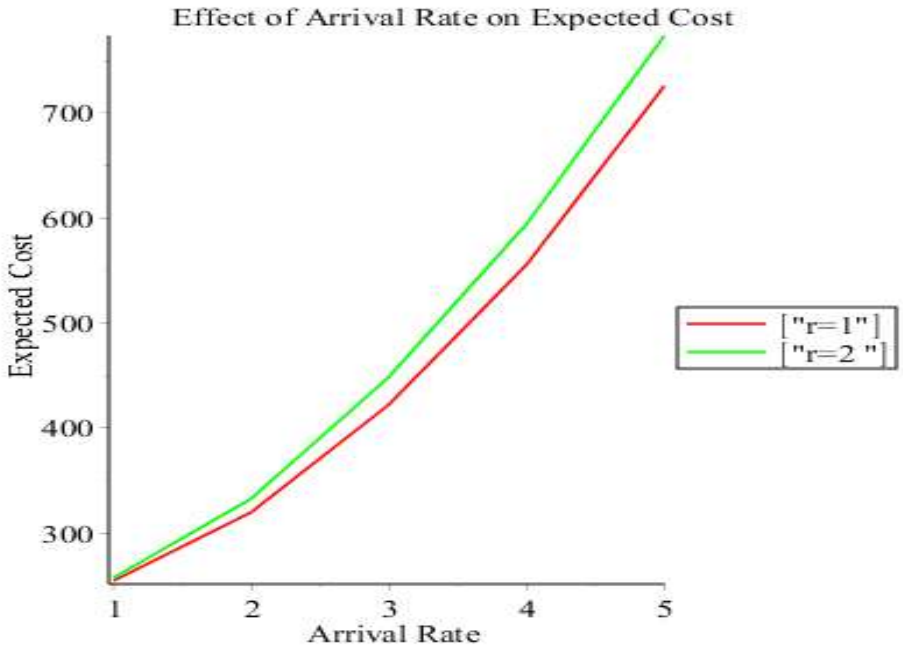


Figure 1(a)

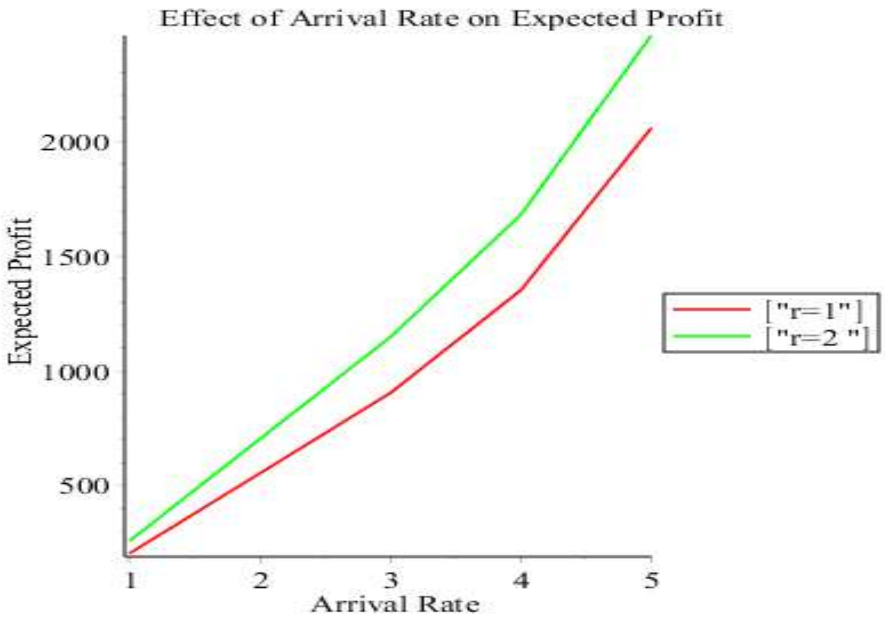


Figure 1(b)

In figures 1(a) and 1(b), we fix $\mu_1 = 3, \mu_2 = 5, = 0.5, \xi = 0.01$, and $q = 0.8$ and display the expected cost and expected profit by varying arrival rates for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems are almost same and increases as arrival increases. Same trend can be seen for profit as well and if arrival rate becomes same or greater than the slow service rate, it is beneficial to use the faster rate to maximize the profit.

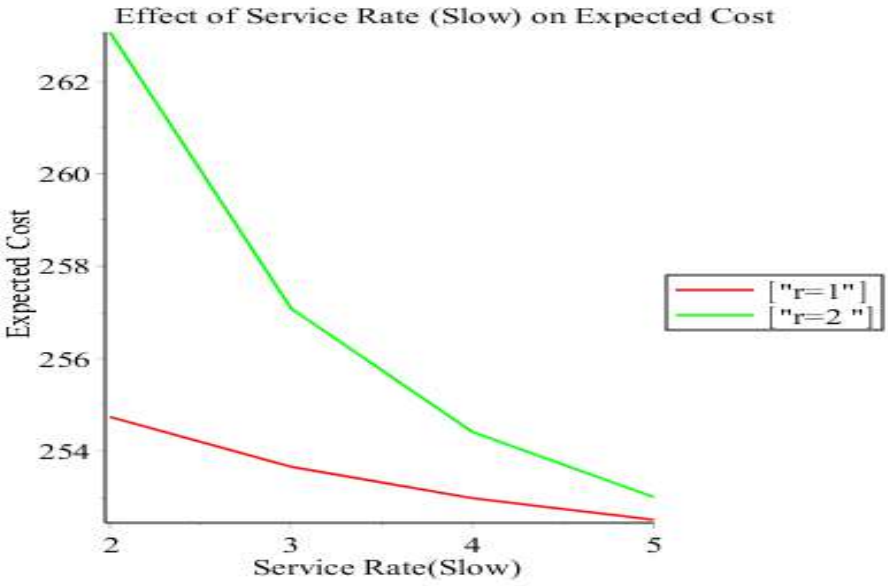


Figure 2(a)

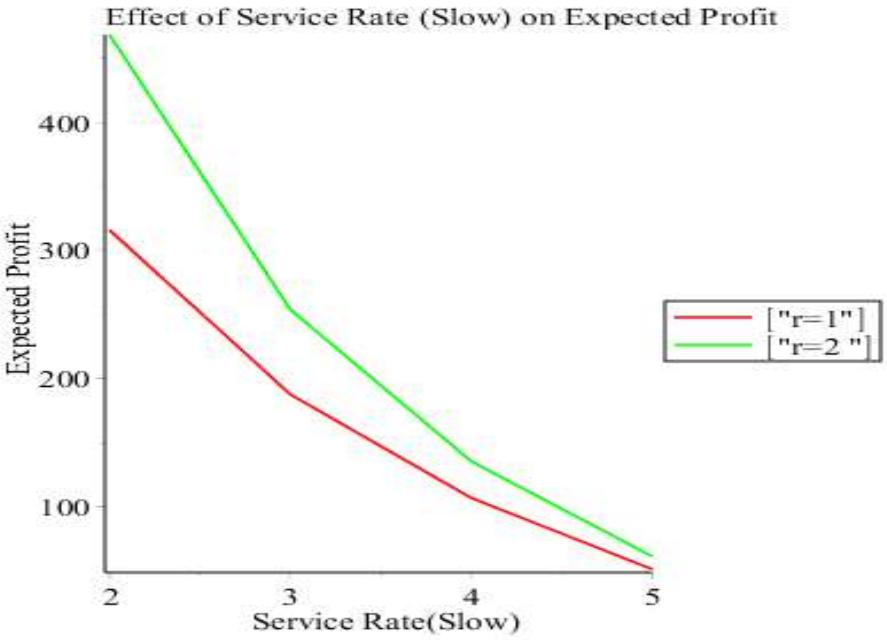


Figure 2(b)

In figures 2(a) and 2(b), we fix $\lambda = 1$, $\mu_2 = 6$, $\beta = 0.5$, $\xi = 0.01$, and $q = 0.8$ and display the expected cost and expected profit by varying slow service rate for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems decreases as service rate increases. But the decrement is more rigorous for system $r = 2$ than for $r = 1$. Same trend can be seen for profit as well. This is because the server remains ideal for rest of the time.

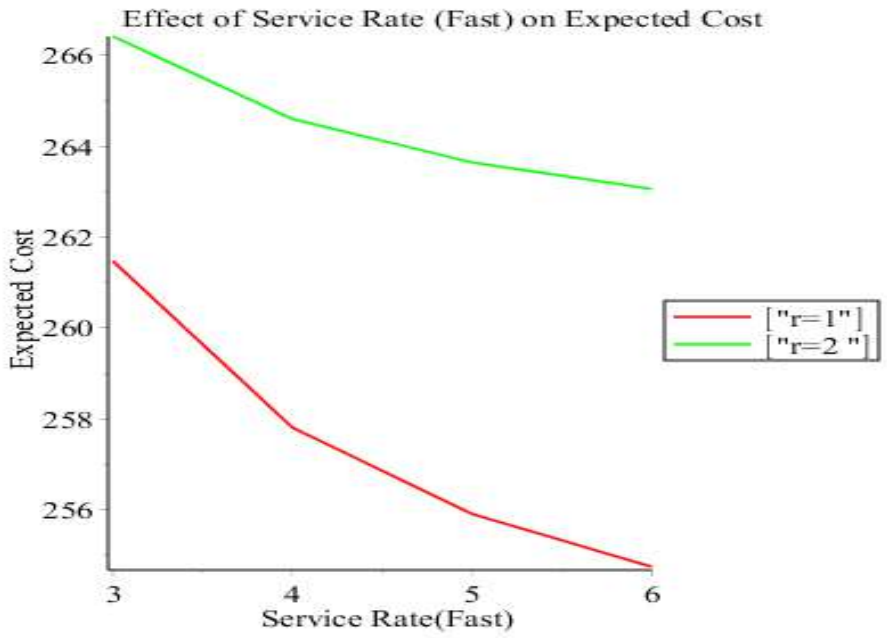


Figure 3(a)

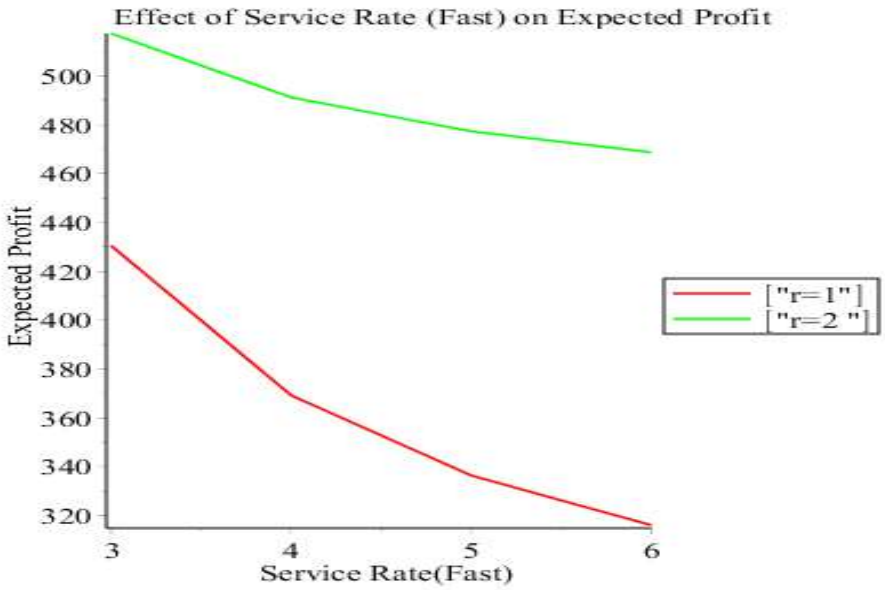


Figure 3(b)

In figures 3(a) and 3(b), we fix $\lambda = 1$, $\mu_1 = 2$, $\beta = 0.5$, $\xi = 0.01$, and $q = 0.8$ and display the expected cost and expected profit by varying fast service rate for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems decreases as service rate increases. Same trend can be seen for profit as well. Varying fast service rate rarely affects the expected cost and slightly affects the expected profit.

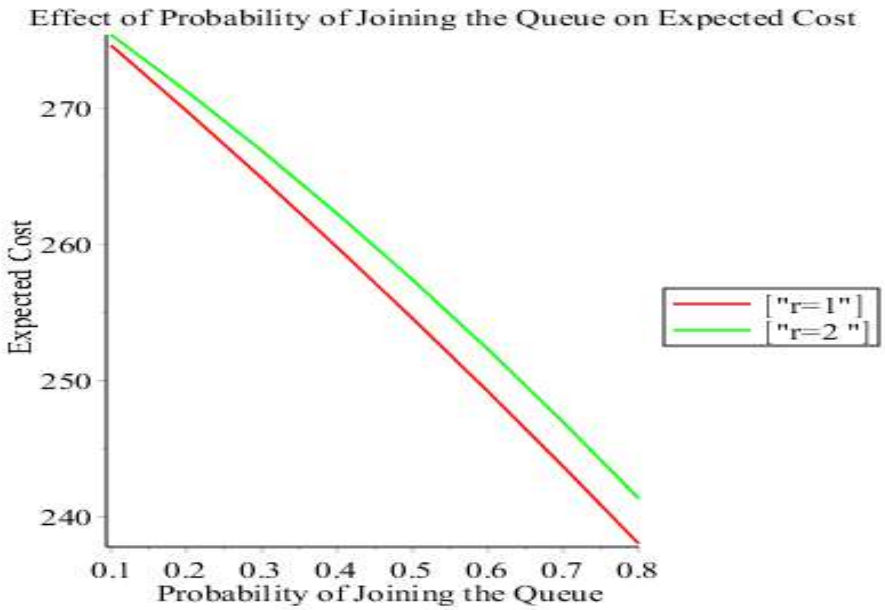


Figure 4(a)

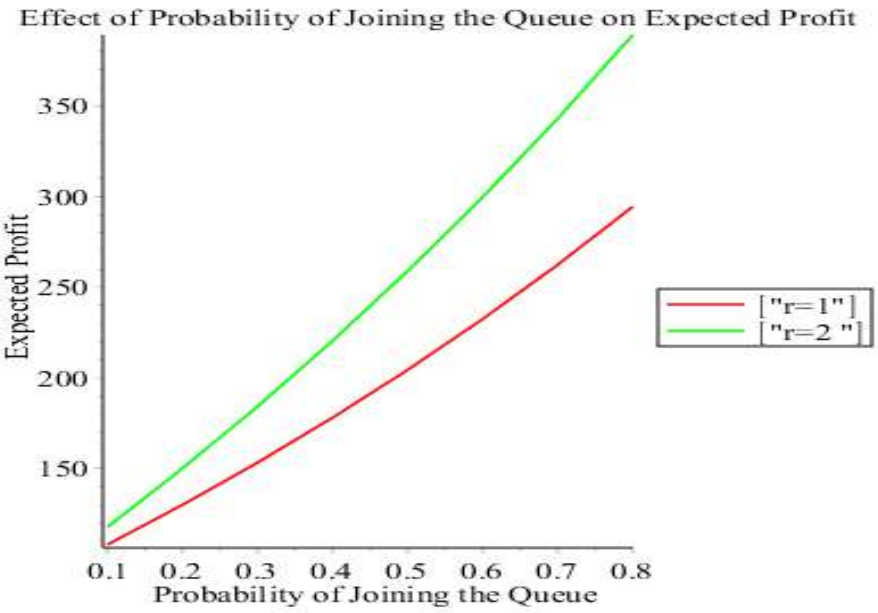


Figure 4(b)

In figures 4(a) and 4(b), we fix $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 5$, $\xi = 0.01$, and $q = 0.8$ and display the expected cost and expected profit by varying joining probability for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems decreases as joining probability increases, whereas expected profit increases as joining probability increases. Thus, profit could be maximized by encouraging the customers to join the system.

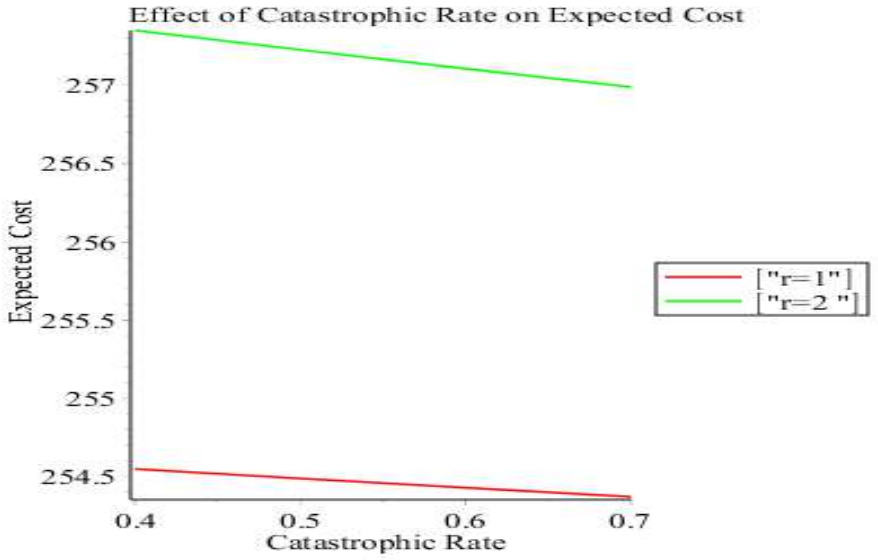


Figure 5(a)

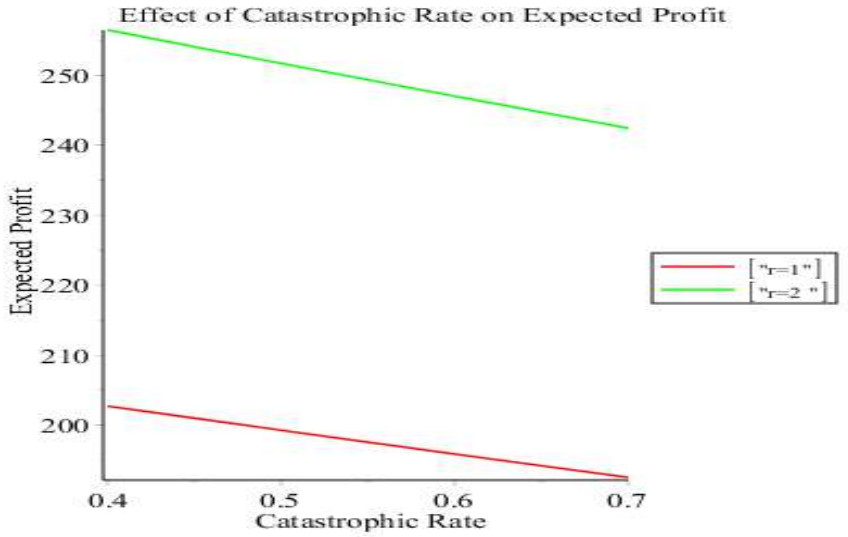


Figure 5(b)

In figures 5(a) and 5(b), we fix $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 5$, $\beta = 0.5$, and $q = 0.8$ and display the expected cost and expected profit by varying catastrophic rate for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems decreases as catastrophic rate increases. Same trend can be seen for profit as well. Increasing catastrophic rate barely affects the cost and profit function.

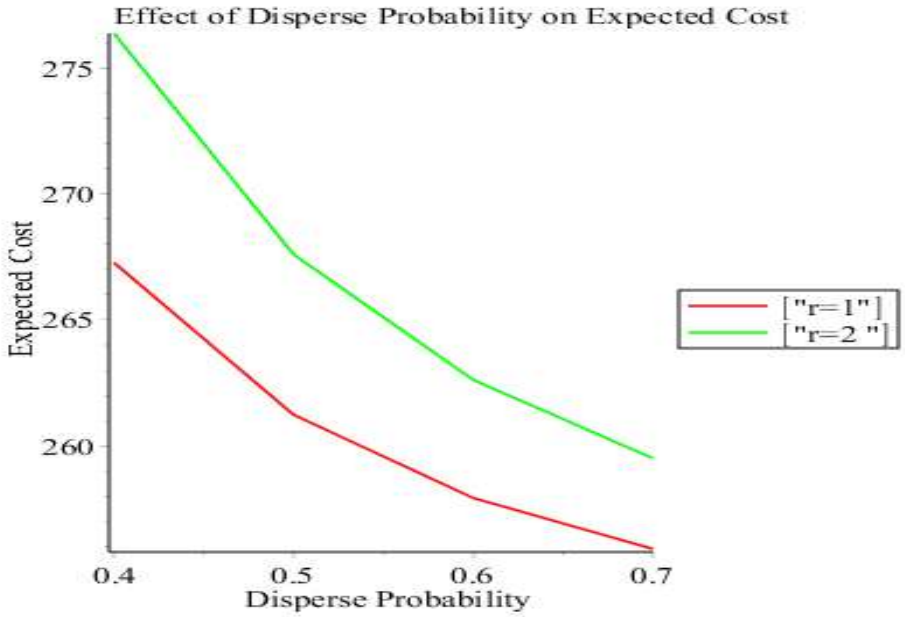


Figure 6(a)

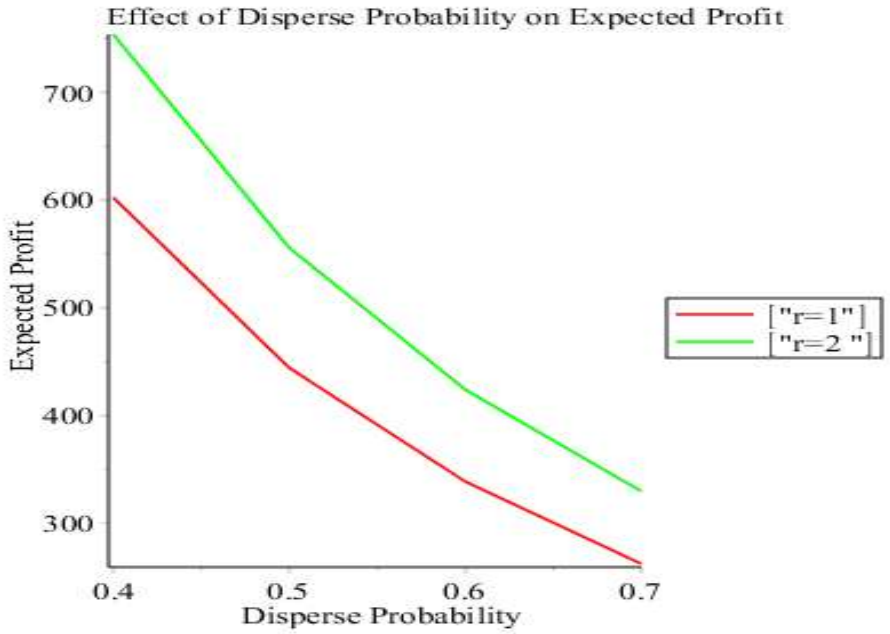


Figure 6(b)

In figures 6(a) and 6(b), we fix $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 5$, $\beta = 0.5$, and $\xi = 0.01$ and display the expected cost and expected profit by varying disperse probability for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected cost for both the systems decreases as probability of leaving the system increases. Same trend can be seen for profit as well. Intuitively, increment in feedback probability will increase the cost.

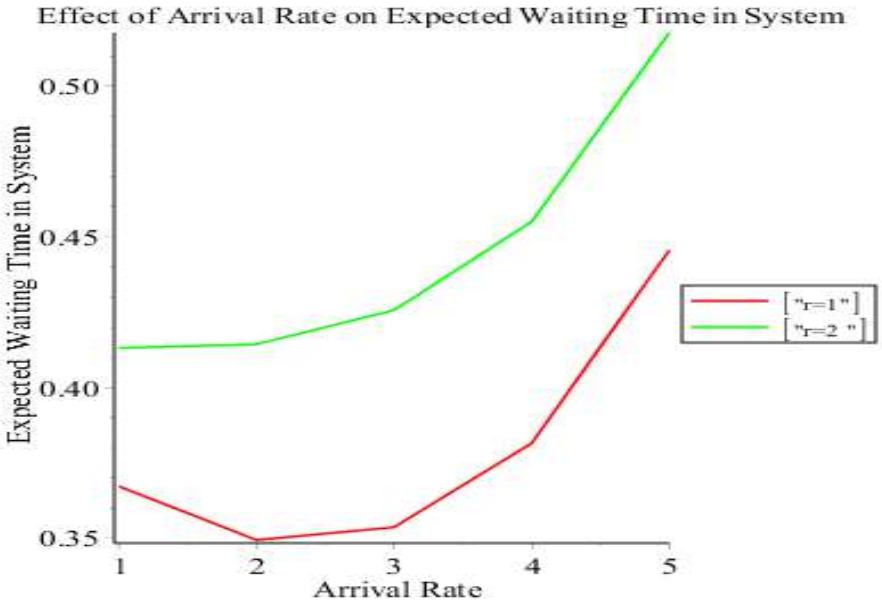


Figure 7(a)

In figure 7(a), we fix $\mu_1 = 3$, $\mu_2 = 5$, $\beta = 0.5$, $\xi = 0.01$, and $q = 0.8$ and display the expected waiting time by varying arrival rates for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected waiting time for both the systems increases as arrival increases.

Figure 7(b)

In figure 7(b), we fix $\lambda = 1$, $\mu_2 = 6$, $\beta = 0.5$, $\xi = 0.01$, and $q = 0.8$ and display the expected waiting time by varying slow service rate for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected waiting time for both the systems decreases as service rate increases. But the decrement is more rigorous for system $r = 2$ than for $r = 1$.

Figure 7(c)

In figure 7(c), we fix $\lambda = 1$, $\mu_1 = 2$, $\beta = 0.5$, $\xi = 0.01$, and $q = 0.8$ and display the expected waiting time by varying fast service rate for both the systems $r = 1$, and $r = 2$. It is clear from the graph that expected waiting time for both the systems decreases as service rate increases.

6. Conclusions

We have presented a detailed study of a queueing system with various parameters. We come across many situations where customer's impatience, dissatisfaction or sudden occurrence of any calamity may cause customer loss and affect the system profit as well. We have incorporated balking, catastrophes, feedback and state dependent service rate altogether to make the model more applicable in real life situations. Many practical congestion situations that we normally encounter such as manufacturing system, call center, communication and telecommunication systems, and health sector may remodel their systems to improve the output by using the results so obtained as tools. Using matrix-geometric solution method, we have analysed the steady-state behaviour of the system and evaluated various performance indicators for the same. An expected cost and profit analysis for the system has been presented and discussed with different set of parameters. From the graphs, it is clear that the optimal value for 'r' is 1. Also, we conclude that server can opt for a faster rate if the arrival rate dominates the initial service rate.

7. Future Considerations

Many real life congestion problems which have special structural properties can be easily solved using matrix-geometric technique even if the dimensions are of higher order. The work can be further extended for markovian and non-markovian queueing networks by considering different parameters along with their transient solutions.

Acknowledgments

We thank the Editors and the reviewers for their constructive comments that helped to improve this paper.

References

Abou-El-Ata, M. O. (1991). The state-dependent queue: M/M/1/N with reneging and general balk functions. *Microelectronics Reliability*, **31 (5)**, 1001-1007.

Ancker Jr, C. J., and Gafarian, A.V. (1963). Some queuing problems with balking and reneging-I. *Operations Research*, **11 (1)**, 88-100.

Bouchentouf, A.A., Cherfaoui, M. and Boualem, M. (2019). Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers. *Opsearch*, **56 (1)**, 300-323.

Bura, G. S., and Bura, R.N. (2015). Time dependent analysis of a queueing system incorporating the effect of environment, catastrophe and restoration. *Journal of Reliability and Statistical Studies*, **8 (2)**, 29-40.

Choudhury, G. and Paul, M. (2005). A two phase queueing system with Bernoulli feedback. *International Journal of Information and Management Sciences*, **16 (1)**, 35-52.

D'Avignon, G. and Disney, R. (1976). Single-server queues with state-dependent

feedback. *INFOR: Information Systems and Operational Research*, **14 (1)**, 71-85.

Doshi, B. T. and Jagerman, D. L. (1986). An M/G/1 queue with class dependent balking (reneging). In *Proc. of the international seminar on Teletraffic analysis and computer performance evaluation*, 225-243.

Haight, F. A. (1957). Queueing with balking. *Biometrika*, **44(3/4),** 360-369.

Kumar, B. K., Madheswari, S. P. and Lakshmi, S. A. (2013). An M/G/1 Bernoulli feedback retrial queueing system with negative customers. *Operational Research*, **13 (2),** 187-210.

Kumar, B.K., Anantha Lakshmi, S. R., Anbarasu, S. and Pavai Madheswari, S. (2014). Transient and steady-state analysis of queueing systems with catastrophes and impatient customers. *International Journal of Mathematics in Operational Research*, **6 (5)**, 523-549.

Santhakumaran, A. and Thangaraj, V. (2000). A single server queue with impatient and feedback customers. *International Journal of Information and Management Sciences*, **11 (3)**, 71-80.

Takacs, L. (1963). A single-server queue with feedback. *Bell System Technical Journal*, **42 (2)**, 505-519.

Thangaraj, V. and Vanitha, S. (2009). On the analysis of M/M/1 feedback queue with catastrophes using continued fractions. *International Journal of Pure and Applied Mathematics*, **53 (1)**, 131-151.

Varalakshmi, M., Chandrasekaran, V. M. and Saravanarajan, M. C. (2018). A single server queue with immediate feedback, working vacation and server breakdown. *International Journal of Engineering and Technology*, **7 (4.10)**, 476-479.

Statistics and Applications {ISSN 2454-7395(online)}
Volume 20, No. 1, 2022 (New Series), pp 229-246

Bayesian Inference of Progressive Type - II Censored Data using Mixture of Log Logistic Distributions

S.S. Bhavsar and M.N. Patel

Department of Statistics, School of Sciences, Gujarat University, Ahmedabad, Gujarat, India

Received: 28 March 2021; Revised: 28 April 2021; Accepted: 01 May 2021

ABSTRACT

Bayesian approach is considered by several authors in mixture models under Type – I, Type – II and other censoring schemes in the area of life testing and survival analysis. In this paper we consider the estimation of parameters of a mixture of log logistic distributions under classical and Bayesian setup. The estimation is done based on progressive Type – II censored sample and the squared error loss function, K- loss function and precautionary loss function are used as loss functions under Bayesian approach. A simulation study is conducted to examine the performance of the proposed estimators based on mean squared error. Bayes estimators under the three types of loss functions are compared using posterior risk too. The results are also compared based on Progressive Type – II censoring and Type – II censoring schemes. Additionally a real life data is considered to determine whether the estimators have similar behavior as seen in simulation study.

Key words: Maximum likelihood; Gamma prior; Log logistic distribution; Posterior risk; Importance sampling.

1. Introduction

The concept of censoring is generally used in life testing experiments and survival analysis since partial or complete destruction of the testing unit becomes quite expensive and time consuming. In literature there are many censoring schemes available, which are mainly based on Type – I or Type – II censoring schemes. When life test is terminated as soon as the pre-determined time is observed, such censoring scheme is known as Type – I censoring, where as in Type - II censoring the test is terminated as soon as pre-determined number of failures observed.

One of the significant shortcoming of conventional censoring schemes is that other than the terminal point of the experiment they do not allow removal of the experimental units at any other points. A censoring scheme in which some of the experimental units are withdrawn during the test and test is continued after the withdrawal is known as progressive censoring scheme. Based on Type – I & Type – II censoring schemes progressive censoring schemes can be formulated as progressive Type – I & progressive Type – II censoring schemes. For detailed study of progressive censoring scheme one may refer Balakrishnan and Aggarwala (2000). The progressive Type – II censoring scheme became very popular among the researchers. Some of the references are Wu *et al.* (2006), Patel and Patel (2007), Gajjar and Patel (2008), Saraçoğlu *et al.* (2010) and Ahmed (2014).

Corresponding author: S. S. Bhavsar
Email: snehbhavsar1@gmail.com

Various types of lifetime models are available in the literature like Exponential, Weibull, Rayleigh, Power function, *etc.* Some of the works available in literature under progressive censoring for above mentioned lifetime models are considered by Fernández (2004), Jung and Chung (2011), Kim and Han (2009), *etc.* In the recent years, the estimation under Bayesian setup for log logistic distribution for progressive censoring is studied by Abbas and Tang (2016), Al-Shomrani *et al.* (2016), Kumar (2018), Yahaya and Ibrahim (2019), *etc.*

In life testing experiments failure of unit may occur due to more than one causes for *e.g.* failure of an electrical component may occur due to fluctuations in voltage or its operating environment or a mechanical shock. Similarly death of a person may occur due to heart attack or old age or any other reason. In such situations a lifetime model based on multiple causes is considered which is known as mixture model. Suppose there are k causes of failure of an experimental unit and p_i be the probability of failure of a unit due to i^{th} cause such that $\sum_{i=1}^{k} p_i = 1$ and $f_i(x)$ be the probability density function or probability mass function of lifetime of the failure units due to i^{th} cause then the mixture model is defined as

$$f(x) = \sum_{i=1}^{k} p_i \, f_i(x) \tag{1}$$

The mixture model is found useful in engineering, medical, agriculture, and many more fields.

Pearson (1894) introduced a statistical model based on finite mixtures of distributions to analyse crab morphometry data. Based on their causes of failures Mendenhall and Hader (1958) studied a population of failures by dividing them in two sub-populations. For estimating the parameters of a mixture of Rayleigh distribution Saleem and Aslam (2009) have used Bayesian procedure. Bayesian estimation was considered by Kazmi *et al.* (2012) for a mixture of Maxwell distribution under Type-I censoring scheme. Bayesian estimation of mixture of power function distributions using Type – II censored sample was discussed by Bhavsar and Patel (2019). Complete and Type – I censored sample are considered by Saleem *et al.* (2010) for estimation of parameters of the mixture of power function distributions. Very few works are available in the area of estimation of mixture model of the log logistic distribution under Bayesian setup based on progressive censoring. This has motivated us to consider a problem of estimation for mixture of log logistic distribution under progressive Type – II censoring scheme.

In this article, an estimation of the parameters of mixture of two log logistic distributions is carried out using the progressive Type – II censored sample considering the maximum likelihood estimation and Bayesian approach, and their respective mean squared errors and posterior risks are studied. The prior considered for the parameters β_1 and β_2 is gamma prior and uniform prior is considered for proportion parameter p of the mixture model. The squared error loss function, K – loss function and precautionary loss function are considered to obtain the Bayes estimates and a comparison between them based on MSE & posterior risk is done. A simulation study is carried out to obtain some interesting conclusions and a real life data is also considered. The rest of the article is structured as follows. In section 2, a two-component mixture model for log logistic distribution and likelihood function under progressive Type – II censoring is described. In section 3, the parameter estimation is carried out using the maximum likelihood estimation approach and the estimators are derived along with their asymptotic variances. Section 4 covers the estimation carried out under the Bayesian setup considering three different loss functions. MSE and posterior risks are obtained for the Bayes estimators. A simulation study is conducted to compare the performance of the proposed estimators in section 5. In section 6, some discussion on the numerical results are presented. Section 7 gives

an idea about the behavior of the estimators for real life data under classical and Bayesian setup. The final conclusion is given in Section 8.

2. Mixture Model

The two - component mixture model for log logistic distribution is defined as follows

$$f(x) = pf_1(x) + (1-p)f_2(x) \tag{2}$$

where $f_i(x) = \frac{\beta_i x^{\beta_i - 1}}{(1+x^{\beta_i})^2}$, $x > 0$, $\beta_i > 0$,$0 < p < 1$;

is the probability density function of log logistic distribution and corresponding distribution function is

$$F_i(x) = \frac{x^{\beta_i}}{(1+x^{\beta_i})}; \qquad i = 1,2 \tag{3}$$

Here β_1, β_2 are unknown parameters of the log logistic distributions and p is unknown mixing proportion with mixing weight $p : 1-p$.

The life testing experiment under progressive censoring can be conducted as follows. Let us suppose that n experimental units are put on test and as soon as the m^{th} failure is observed the test is terminated. Considering a mixture model an experimental unit may fail due to cause 1 or cause 2. The failed unit can easily be identified whether it is from sub population 1 (which failed due to cause 1) or sub population 2 (which failed due to cause 2). Since this is progressive censoring scheme, as soon as the 1st failure occurs R_1 units are removed from the test which has remaining $(n-1)$ units on the test and the test is continued with $(n-1-R_1)$ units. Similarly on the $(m-1)^{th}$ failure R_m units are withdrawn from the remaining units on the test and test is continued with $(n-m-1-\sum_{i=1}^{m-1} R_i)$ units. The test is finally terminated as soon as the m^{th} failure is observed.

Thus depending upon the cause of failure, we can identify the number of failures m_1 due to cause 1 and m_2 due to cause 2 from the $m(=m_1+m_2)$ observed failures. The remaining $(n-m-\sum_{i=1}^{m-1} R_i)$ units are censored which provide no information about the sub population and survive beyond the time $X_{(m)}$, the observed time of the m^{th} failure.

To produce precise inferences a mixture model must be identifiable and in our model we have only shape parameters β_1 and β_2. Suppose x_{1i} and x_{2i} are the i^{th} failure time due to cause 1 and cause 2 respectively. The general form of likelihood function for the two – component mixture distribution under progressive Type – II censoring is given by:

$$L(\beta_1, \beta_2, p|x) \propto \prod_{i=1}^{m_1} p.f_1(x_{1i}) \prod_{i=1}^{m_2} (1-p) f_2(x_{2i}) \prod_{i=1}^{m} \left[1 - px_i^{\beta_1} - (1-p)x_i^{\beta_2}\right]^{R_i} \tag{4}$$

3. Maximum Likelihood Estimation (MLE)

Using Eq. (2) and (4), the likelihood function under progressive Type – II censoring for mixture model is obtained as,

$$L \propto p^{m_1} {\beta_1}^{m_1} \prod_{i=1}^{m_1} \frac{(x_{1i})^{\beta_1 - 1}}{\left(1+x_{1i}^{\beta_1}\right)^2} (1-p)^{m_2} {\beta_2}^{m_2} \prod_{i=1}^{m_2} \frac{(x_{2i})^{\beta_2 - 1}}{(1+x_{2i}^{\beta_2})^2} \times$$

$$\prod_{i=1}^{m}\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]^{R_i} \tag{5}$$

$$\begin{aligned}\log L \propto\ & m_1 log p + m_1 log\beta_1 + (\beta_1-1)\sum_{i=1}^{m_1} log x_{1i} - 2\sum_{i=1}^{m_1}\log\left(1+{x_{1i}}^{\beta_1}\right) + \\ & m_2\log(1-p) + m_2 log\beta_2 + (\beta_2-1)\sum_{i=1}^{m_2} log x_{2i} - 2\sum_{i=1}^{m_1}\log\left(1+{x_{2i}}^{\beta_2}\right) + \\ & \sum_{i=1}^{m} R_i \log\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]\end{aligned} \tag{6}$$

The first derivatives of equation Eq. (6) with respect to β_1, β_2 and p are

$$\begin{aligned}\frac{\partial log L}{\partial\beta_1} =\ & \frac{m_1}{\beta_1} + \sum_{i=1}^{m_1} log x_{1i} - 2\sum_{i=1}^{m_1}\frac{{x_{1i}}^{\beta_1}\log x_{1i}}{1+{x_{1i}}^{\beta_1}} \\ & + \sum_{i=1}^{m} R_i \frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]}\left(-p\frac{{x_i}^{\beta_1}\, log x_i}{(1+{x_i}^{\beta_1})^2}\right)\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{\partial log L}{\partial\beta_2} =\ & \frac{m_2}{\beta_2} + \sum_{i=1}^{m_2} log x_{2i} - 2\sum_{i=1}^{m_2}\frac{{x_{2i}}^{\beta_2}\log x_{2i}}{1+{x_{2i}}^{\beta_2}} \\ & + \sum_{i=1}^{m} R_i \frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]}\left(-(1-p)\frac{{x_i}^{\beta_2}\, log x_i}{(1+{x_i}^{\beta_2})^2}\right)\end{aligned} \tag{8}$$

$$\frac{\partial log L}{\partial p} = \frac{m_1}{p} - \frac{m_2}{(1-p)} - \sum_{i=1}^{m} R_i\left(\frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]} + \frac{{x_i}^{\beta_1}}{(1+{x_i}^{\beta_1})} - \frac{{x_i}^{\beta_2}}{(1+{x_i}^{\beta_2})}\right) \tag{9}$$

On equating the Eq. (7), (8) and (9) with respect to 0, we get the likelihood equations for β_1, β_2 and p as

$$\beta_1 = \frac{m_1}{-\sum_{i=1}^{m_1} log x_{1i} + 2\sum_{i=1}^{m_1}\frac{{x_{1i}}^{\beta_1}\log x_{1i}}{1+{x_{1i}}^{\beta_1}} - \sum_{i=1}^{m} R_i \frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]}\left(-p\frac{{x_i}^{\beta_1}\, log x_i}{(1+{x_i}^{\beta_1})^2}\right)} \tag{10}$$

$$\beta_2 = \frac{m_2}{-\sum_{i=1}^{m_2} log x_{2i} + 2\sum_{i=1}^{m_2}\frac{{x_{2i}}^{\beta_2}\log x_{2i}}{1+{x_{2i}}^{\beta_2}} - \sum_{i=1}^{m} R_i \frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]}(-(1-p)\frac{{x_i}^{\beta_2}\, log x_i}{(1+{x_i}^{\beta_2})^2})} \tag{11}$$

$$p = \frac{m_1-(p-p^2)+\sum_{i=1}^{m} R_i\left(\frac{1}{\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]+\frac{{x_i}^{\beta_1}}{\left(1+{x_i}^{\beta_1}\right)}-\frac{{x_i}^{\beta_2}}{\left(1+{x_i}^{\beta_2}\right)}}\right)}{m} \tag{12}$$

which can be solved by any method of iteration and we get the MLEs $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\hat{p}$.

To obtain Variance-Covariance matrix of ML estimators, we find second derivatives of $logL$ with respect to the parameters β_1, β_2 and p as

$$\frac{\partial^2 log L}{\partial {\beta_1}^2} = \frac{-m_1}{{\beta_1}^2} - 2\sum_{i=1}^{m_1}\log x_{1i}\left[\frac{(1+{x_{1i}}^{\beta_1}){x_{1i}}^{\beta_1}\log x_{1i} - {x_{1i}}^{\beta_1}{x_{1i}}^{\beta_1}\log x_{1i}}{\left(1+{x_{1i}}^{\beta_1}\right)^2}\right] -$$

$$p\ \sum_{i=1}^{m} R_i \log x_i \left[\left(1 - p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)\left(1+{x_i}^{\beta_1}\right)^2 {x_i}^{\beta_1}\ logx_i - \right.$$

$${x_i}^{\beta_1}\left\{\begin{matrix} 2\left(1+{x_i}^{\beta_1}\right){x_i}^{\beta_1}\ logx_i - p\left({x_i}^{\beta_1}\ logx_i + {x_i}^{2\beta_1} 2\ logx_i\right) - \\ (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\ {x_i}^{\beta_1}\ logx_i\ 2\left(1+{x_i}^{\beta_1}\right) \end{matrix}\right\} \times$$

$$\left.(1 - p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}})^{-2}\left(1+{x_i}^{\beta_1}\right)^{-4}\right] \tag{13}$$

$$\frac{\partial^2 logL}{\partial {\beta_2}^2} = \frac{-m_2}{{\beta_2}^2} - 2\sum_{i=1}^{m_2} \log x_{2i} \left[\frac{\left(1+{x_{2i}}^{\beta_2}\right){x_{2i}}^{\beta_2} \log x_{2i} - {x_{2i}}^{\beta_2}{x_{2i}}^{\beta_2} \log x_{2i}}{\left(1+{x_{2i}}^{\beta_2}\right)^2}\right] -$$

$$(1-p)\sum_{i=1}^{m} R_i \log x_i \left[\left(1 - p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)\left(1+{x_i}^{\beta_2}\right)^2 {x_i}^{\beta_2}\ logx_i\ - \right.$$

$${x_i}^{\beta_2}\left\{\begin{matrix} 2\left(1+{x_i}^{\beta_2}\right){x_i}^{\beta_2}\ logx_i - (1-p)\left({x_i}^{\beta_2}\ logx_i + {x_i}^{2\beta_2} 2\ logx_i\right) - \\ (1-p)\left(\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}\ {x_i}^{\beta_2}\ logx_i\ 2\left(1+{x_i}^{\beta_2}\right)\right) \end{matrix}\right\} \times$$

$$\left.\left(1 - p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)^{-2}\left(1+{x_i}^{\beta_1}\right)^{-4}\right] \tag{14}$$

$$\frac{\partial^2 logL}{\partial\beta_1\partial\beta_2} = -p\ (1-p)\ \sum_{i=1}^{m}\left[\frac{R_i\ {x_i}^{\beta_1}\ {x_i}^{\beta_2}\ (\log x_i)^2}{\left(1+{x_i}^{\beta_1}\right)^2\left(1+{x_i}^{\beta_2}\right)^2\left(1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)^2}\right] \tag{15}$$

$$\frac{\partial^2 logL}{\partial\beta_1\partial p} = \sum_{i=1}^{m}\frac{R_i}{\left(1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)^2} - \frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} + \frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}} - \frac{{x_i}^{\beta_1}\log x_i}{\left(1+{x_i}^{\beta_1}\right)^2} \tag{16}$$

$$\frac{\partial^2 logL}{\partial\beta_2\partial p} = \sum_{i=1}^{m}\frac{R_i}{\left(1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)^2} - \frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} + \frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}} - \frac{{x_i}^{\beta_2}\log x_i}{\left(1+{x_i}^{\beta_2}\right)^2} \tag{17}$$

$$\frac{\partial^2 logL}{\partial p^2} = -\frac{m_1}{p^2} - \frac{m_2}{(1-p)^2}\sum_{i=1}^{m}\frac{R_i}{\left(1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} - (1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right)^2} - \frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}} + \frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}} \tag{18}$$

The Variance – Covariance matrix of MLEs of the parameters is given by,

$$V = \begin{bmatrix} -E\left[\frac{\partial^2 logL}{\partial {\beta_1}^2}\right] & -E\left[\frac{\partial^2 logL}{\partial\beta_2\partial\beta_1}\right] & -E\left[\frac{\partial^2 logL}{\partial p\partial\beta_1}\right] \\ -E\left[\frac{\partial^2 logL}{\partial\beta_2\partial\beta_1}\right] & -E\left[\frac{\partial^2 logL}{\partial {\beta_2}^2}\right] & -E\left[\frac{\partial^2 logL}{\partial p\partial\beta_2}\right] \\ -E\left[\frac{\partial^2 logL}{\partial p\partial\beta_1}\right] & -E\left[\frac{\partial^2 logL}{\partial p\partial\beta_2}\right] & -E\left[\frac{\partial^2 logL}{\partial p^2}\right] \end{bmatrix}^{-1}$$

According to Lawless (2003) the estimate of variance covariance matrix is given as

$$\hat{V} = \left[\begin{matrix} -\frac{\partial^2 logL}{\partial {\beta_1}^2} & -\frac{\partial^2 logL}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 logL}{\partial p \partial \beta_1} \\ -\frac{\partial^2 logL}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 logL}{\partial {\beta_2}^2} & -\frac{\partial^2 logL}{\partial p \partial \beta_2} \\ -\frac{\partial^2 logL}{\partial p \partial \beta_1} & -\frac{\partial^2 logL}{\partial p \partial \beta_2} & -\frac{\partial^2 logL}{\partial p^2} \end{matrix}\right]^{-1}_{(\beta_1,\beta_2,p)=(\hat{\beta}_1,\hat{\beta}_2,\hat{p})} \tag{19}$$

The variances of $\widehat{\beta_1}$, $\widehat{\beta_2}$ and $\hat{p}$ are given by diagonal elements of the matrix V.

The mean squared error is calculated for each of the above parameter using the below equation

$$\text{Mean squared error}\left(\hat{\theta}\right) = \text{Variance}\left(\hat{\theta}\right) + (\hat{\theta} - \theta)^2, \quad \theta = (\beta_1, \beta_2, p) \tag{20}$$

4. Bayes Estimation

The Bayesian approach considers prior information along with the information available from the data to form a posterior distribution which is used for Bayesian inference. Comparatively less sample data is required in Bayesian method than in classical sampling theory, which makes it more preferable in life testing and reliability estimation where sample data is costly and hard to obtain.

Under Bayesian estimation, a joint distribution function $\phi(\theta, \underline{x})$ is obtained using the likelihood function and the specified prior distribution of the unknown parameters. A marginal distribution $m(\underline{x})$ is derived on integrating the joint distribution function over the range of its parameters. The joint posterior distribution $g(\theta|\underline{x})$ is obtained by taking a ratio of joint distribution of β_1, β_2, p and $\underline{x}$ and marginal distribution $m(\underline{x})$. The marginal posterior distribution is derived by integrating joint posterior distribution over the range of its parameters. In this section, Bayes estimates of the parameters are obtained using the marginal posterior distributions of the parameters and their corresponding mean squared errors and posterior risks are also obtained.

Consider the gamma priors for the parameter β_1 and β_2, and uniform prior for the parameter p.

$$\Pi_1(\beta_1) = \frac{{a_1}^{b_1} {\beta_1}^{b_1-1} e^{-a_1\beta_1}}{\Gamma b_1}, \ \beta_1 > 0; \ a_1, b_1 > 0 \tag{21}$$

$$\Pi_2(\beta_2) = \frac{{a_2}^{b_2} {\beta_2}^{b_2-1} e^{-a_2\beta_2}}{\Gamma b_2}, \ \beta_2 > 0; \ a_2, b_2 > 0 \tag{22}$$

$$\Pi_3(p) = 1, \ \ 0 < p < 1 \tag{23}$$

Using the likelihood function in Eq. (5) and prior distributions in Eq. (21), (22) and (23), the joint distribution of parameters and sample becomes

$$\phi(\beta_1, \beta_2, p, \underline{x}) \propto L \, \Pi_1(\beta_1) \, \Pi_2(\beta_2) \, \Pi_3(p) \tag{24}$$

$$\phi(\beta_1, \beta_2, p, \underline{x}) \propto p^{m_1} {\beta_1}^{m_1} \prod_{i=1}^{m_1} \frac{(x_{1i})^{\beta_1 - 1}}{(1 + {x_{1i}}^{\beta_1})^2} (1-p)^{m_2} {\beta_2}^{m_2} \prod_{i=1}^{m_2} \frac{(x_{2i})^{\beta_2 - 1}}{(1 + {x_{2i}}^{\beta_2})^2}$$

$$\prod_{i=1}^{m}\left[1-p\frac{{x_i}^{\beta_1}}{1+{x_i}^{\beta_1}}-(1-p)\frac{{x_i}^{\beta_2}}{1+{x_i}^{\beta_2}}\right]^{R_i}\ \frac{{a_1}^{b_1}\,{\beta_1}^{b_1-1}\,e^{-a_1\beta_1}}{\Gamma b_1}\ \frac{{a_2}^{b_2}\,{\beta_2}^{b_2-1}\,e^{-a_2\beta_2}}{\Gamma b_2} \tag{25}$$

The joint posterior distribution of β_1, β_2 and p can be obtained as

$$g(\beta_1,\beta_2,p|\underline{x}) = \frac{\phi(\beta_1,\beta_2,p,\underline{x})}{m(\underline{x})} \tag{26}$$

where $m(\underline{x})$ is the marginal distribution of $\underline{x}$ that can be derived from the joint distribution as

$$m(\underline{x}) = \int_{\beta_1}\int_{\beta_2}\int_{p}\phi(\beta_1,\beta_2,p|\underline{x})\,dp\,d\beta_2\,d\beta_1 \tag{27}$$

Using the equations Eq. (25) and (27) the joint posterior distribution can be written as

$$g(\beta_1,\beta_2,p|\underline{x}) \propto p^{m_1}(1-p)^{m_2}{\beta_1}^{m_1+b_1-1}e^{-\beta_1\left(a_1-\sum_{i=1}^{m_1}logx_{1i}\right)}{\beta_2}^{m_2+b_2-1}$$
$$e^{-\beta_2\left(a_2-\sum_{i=1}^{m_2}logx_{2i}\right)}\,e^{\sum_{i=1}^{m}R_i\log\left[1-p\frac{x_i^{\beta_1}}{1+x_i^{\beta_1}}-(1-p)\frac{x_i^{\beta_2}}{1+x_i^{\beta_2}}\right]}\,e^{-2\sum_{i=1}^{m_1}\log\left(1+x_{1i}^{\beta_1}\right)}$$
$$e^{-2\sum_{i=1}^{m_2}\log\left(1+x_{2i}^{\beta_2}\right)}e^{-\sum_{i=1}^{m_1}log(x_{1i})}e^{-\sum_{i=1}^{m_2}log(x_{2i})}\ \frac{{a_1}^{b_1}}{\Gamma b_1}\ \frac{{a_2}^{b_2}}{\Gamma b_2} \tag{28}$$

The marginal posterior distribution of β_1, β_2 and p can be determined by integrating with respect to the other parameters. The marginal posterior distribution of prior β_1 is given by

$$h_1(\beta_1|\underline{x}) = \int_{\beta_2}\int_{p}p^{m_1}\,(1-p)^{m_2}\,{\beta_1}^{m_1+b_1-1}\,e^{-\beta_1\left(a_1-\sum_{i=1}^{m_1}logx_{1i}\right)}\,{\beta_2}^{m_2+b_2-1}$$
$$e^{-\beta_2\left(a_2-\sum_{i=1}^{m_2}logx_{2i}\right)}\ e^{\sum_{i=1}^{m}R_i\log\left[1-p\frac{x_i^{\beta_1}}{1+x_i^{\beta_1}}-(1-p)\frac{x_i^{\beta_2}}{1+x_i^{\beta_2}}\right]}\,e^{-2\sum_{i=1}^{m_1}\log\left(1+x_{1i}^{\beta_1}\right)}$$
$$e^{-2\sum_{i=1}^{m_2}\log\left(1+x_{2i}^{\beta_2}\right)}e^{-\sum_{i=1}^{m_1}log(x_{1i})}e^{-\sum_{i=1}^{m_2}log(x_{2i})}\ \frac{{a_1}^{b_1}}{\Gamma b_1}\ \frac{{a_2}^{b_2}}{\Gamma b_2}\ dp\,d\beta_2 \tag{29}$$

Similarly the marginal posterior distribution of prior β_2 and p are given by

$$h_2(\beta_2|\underline{x}) = \int_{\beta_1}\int_{p}p^{m_1}\,(1-p)^{m_2}\,{\beta_1}^{m_1+b_1-1}\,e^{-\beta_1\left(a_1-\sum_{i=1}^{m_1}logx_{1i}\right)}\,{\beta_2}^{m_2+b_2-1}$$
$$e^{-\beta_2\left(a_2-\sum_{i=1}^{m_2}logx_{2i}\right)}\,e^{\sum_{i=1}^{m}R_i\log\left[1-p\frac{x_i^{\beta_1}}{1+x_i^{\beta_1}}-(1-p)\frac{x_i^{\beta_2}}{1+x_i^{\beta_2}}\right]}\,e^{-2\sum_{i=1}^{m_1}\log\left(1+x_{1i}^{\beta_1}\right)}$$
$$e^{-2\sum_{i=1}^{m_2}\log\left(1+x_{2i}^{\beta_2}\right)}\,e^{-\sum_{i=1}^{m_1}log(x_{1i})}e^{-\sum_{i=1}^{m_2}log(x_{2i})}\ \frac{{a_1}^{b_1}}{\Gamma b_1}\ \frac{{a_2}^{b_2}}{\Gamma b_2}\ dp\,d\beta_1 \tag{30}$$

$$h_3(p|\underline{x}) = \int_{\beta_1}\int_{\beta_2}p^{m_1}\,(1-p)^{m_2}\,{\beta_1}^{m_1+b_1-1}\,e^{-\beta_1\left(a_1-\sum_{i=1}^{m_1}logx_{1i}\right)}\,{\beta_2}^{m_2+b_2-1}$$
$$e^{-\beta_2\left(a_2-\sum_{i=1}^{m_2}logx_{2i}\right)}\ e^{\sum_{i=1}^{m}R_i\log\left[1-p\frac{x_i^{\beta_1}}{1+x_i^{\beta_1}}-(1-p)\frac{x_i^{\beta_2}}{1+x_i^{\beta_2}}\right]}\,e^{-2\sum_{i=1}^{m_1}\log\left(1+x_{1i}^{\beta_1}\right)}$$
$$e^{-2\sum_{i=1}^{m_2}\log\left(1+x_{2i}^{\beta_2}\right)}\,e^{-\sum_{i=1}^{m_1}log(x_{1i})}e^{-\sum_{i=1}^{m_2}log(x_{2i})}\ \frac{{a_1}^{b_1}}{\Gamma b_1}\ \frac{{a_2}^{b_2}}{\Gamma b_2}\ d\beta_2\,d\beta_1 \tag{31}$$

It is not possible to obtain the above mentioned marginal posterior distributions in closed form, which makes it difficult to obtain Bayes estimators directly using marginal posterior distributions. In literature, there are various methods like numerical integration method, Lindley approximation, importance sampling, MCMC technique *etc.* that are useful in such cases. We have used here the importance sampling method used by Kundu and Pradhan (2009) to obtain Bayes estimates of the parameters in any kind of loss functions such as SELF, KLF, PLF *etc.* This method is discussed in many other articles also, some of them are Sultana *et al.* (2020), Madi and Raqab (2009) and Sultana *et al.* (2018).

Based on theory of Bayes estimation a loss function gauges the difference between the estimate $\hat{\theta}$ and the parameter θ and there is no particular procedures to select any loss functions. A posterior risk is the expected value of loss function and the posterior risks associated with the estimators are compared to evaluate the performances of the Bayes estimators. The loss functions used in this paper are described below:

Squared Error loss function (SELF): The Squared error loss function is given by $l(\hat{\theta},\ \theta) = (\hat{\theta} - \theta)^2$.

The Bayes estimate and the posterior risk are defined as

$$\hat{\theta} = E(\theta|x) \tag{32}$$

and

$$\rho(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{33}$$

K – loss function (KLF): The K – loss function was proposed by Wasan (1970), is defined as

$l(\hat{\theta},\ \theta) = (\hat{\theta} - \theta)^2 / \hat{\theta}\theta.$

The Bayes estimate and the posterior risk are defined as

$$\hat{\theta} = \sqrt{E(\theta|x)/E(\theta^{-1}|x)} \tag{34}$$

and

$$\rho(\hat{\theta}) = 2\{E(\theta|x)E(\theta^{-1}|x) - 1\} \tag{35}$$

respectively.

Precautionary loss function (PLF): The Precautionary loss function was proposed by Norstrom (1996), is defined as $l(\hat{\theta},\ \theta) = (\theta - \hat{\theta})^2 / \hat{\theta}$.

The Bayes estimate and the posterior risk are defined as

$$\hat{\theta} = \{E(\theta^2|x)\}^{\frac{1}{2}} \tag{36}$$

and

$$\rho(\hat{\theta}) = 2[\{E(\theta^2|x)\}^{\frac{1}{2}} - E(\theta|x)] \tag{37}$$

respectively.

To employ importance sampling method for Bayes estimation we rewrite the joint posterior distribution given in Eq. (28) as

$$g(\beta_1, \beta_2, p|\underline{x}) \propto p^{m_1}(1-p)^{m_2}\ {\beta_1}^{m_1+b_1-1}\ e^{-\beta_1 A_1}\ {\beta_2}^{m_2+b_2-1}\ e^{-\beta_2 A_2}$$

$$e^{-2\sum_{i=1}^{m_1}\log(1+x_{1i}\beta_1)}\ e^{-2\sum_{i=1}^{m_2}\log(1+x_{2i}\beta_2)}\ e^{\sum_{i=1}^{m} R_i \log\left[1-p\frac{x_i\beta_1}{1+x_i\beta_1}-(1-p)\frac{x_i\beta_2}{1+x_i\beta_2}\right]}$$

$$e^{-\sum_{i=1}^{m_1} log(x_{1i})} e^{-\sum_{i=1}^{m_2} log(x_{2i})} \frac{{a_1}^{b_1}}{\Gamma b_1} \frac{{a_2}^{b_2}}{\Gamma b_2} \tag{38}$$

where $A_1 = a_1 - \sum_{i=1}^{m_1} \log x_{1i}$ and $A_2 = a_2 - \sum_{i=1}^{m_2} \log x_{2i}$

The above form can also be written as

$$g(\beta_1, \beta_2, p|\underline{x}) \propto g_1(p \mid m_1, m_2)\, g_2(\beta_1 \mid \underline{x}, m_1)\, g_3(\beta_2 \mid \underline{x}, m_2)\, \Psi(p, \beta_1, \beta_2 \mid \underline{x}, m) \tag{39}$$

where

$g_1(p \mid m_1, m_2)$ is the probability density function of $beta\ (m_1 + 1, m_2 + 1)$ distribution

$g_2(\beta_1 \mid \underline{x}, m_1)$ is the probability density function of $gamma\ (m_1 + b_1, A_1)$ distribution

$g_3(\beta_2 \mid \underline{x}, m_2)$ is the probability density function of $gamma\ (m_2 + b_2, A_2)$ distribution

$\Psi(p, \beta_1, \beta_2 \mid \underline{x}, m) =$

$e^{-2\sum_{i=1}^{m_1} \log(1+x_{1i}{}^{\beta_1})}\ e^{-2\sum_{i=1}^{m_2} \log(1+x_{2i}{}^{\beta_2})}\ e^{\sum_{i=1}^{m} R_i \log\left[1-p\frac{x_i{}^{\beta_1}}{1+x_i{}^{\beta_1}}-(1-p)\frac{x_i{}^{\beta_2}}{1+x_i{}^{\beta_2}}\right]}$, a function of β_1, β_2, p and $\underline{x}$.

The hyper parameters used in the prior distributions are determined as follows:

- Find means & variances of the MLEs of parameters β_1, β_2 and p and considered them as prior information of the parameters.
- These estimates are compared with theoretical mean & variance of the prior distribution.
- Solving them we obtain estimates of the hyper parameters.

Algorithm-1

The steps of importance sampling to obtain Bayes estimates are as follows:

Step - 1: Decide the values of β_1, β_2, p and $R_1, R_2, \ldots R_m$ such that $\sum_{i=1}^{m} R_i = (n - m)$.

Step - 2: Generate

$x_{1i}\ ; i = 1, 2, \ldots\ldots\ldots, m_1$ and $x_{2i}; i = 1, 2, \ldots\ldots\ldots, m_2$

Step - 3: Generate

N values of p from $beta\ (m_1 + 1, m_2 + 1)$ as $(p_1, p_2, \ldots, p_N)$

N values of β_1 from $gamma\ (m_1 + b_1, A_1)$ as $(\beta_{11}, \beta_{12}, \ldots, \beta_{1N})$

N values of β_2 from $gamma\ (m_2 + b_2, A_2)$ as $(\beta_{21}, \beta_{22}, \ldots, \beta_{2N})$

Step - 4: Calculate the $E(\theta|\underline{x})$ using the formula:

$$E(\theta|\underline{x}) = \frac{\sum_{i=1}^{N} \theta_i . \Psi(\theta_i|\underline{x}, m)}{\sum_{i=1}^{N} \Psi(\theta_i|\underline{x}, m)} \tag{40}$$

Step - 5: Calculate $E(p|x)$, $E(p^2|x)$, $E\left(\frac{1}{p}|x\right)$, $E(\beta_1|x)$, $E({\beta_1}^2|x)$, $E\left(\frac{1}{\beta_1}|x\right)$, $E(\beta_2|x)$, $E({\beta_2}^2|x)$, $E\left(\frac{1}{\beta_2}|x\right)$ using the Eq. (40) to calculate the estimate, PR, and MSE using the squared error loss function, K – loss function and precautionary loss function.

5. Simulation Study

A simulation study is setup to check the performance of ML estimators and Bayes estimators obtained in the earlier sections. We have used the following inputs.

To simulate samples from 2 component mixture of log logistic distributions, we have used the following algorithm with the inputs: S=5000 which is number of simulations, n = 60 and m = 15 (25% censored) with censoring scheme R = (0, 5, 2, 8, 0, 2, 5, 3, 0, 6, 4, 9, 1, 0, 0); m = 25 (42% censored) with censoring scheme R = (0, 2, 5, 3, 2, 3, 0, 0, 0, 1, 4, 1, 1, 1, 2, 0, 0, 5, 2, 0, 1, 0, 0, 2, 0) and m = 35 (58% censored) with censoring scheme R = (0, 0, 2, 0, 1, 0, 0, 3, 0, 1, 1, 1, 1, 1, 2, 0, 0, 2, 2, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0, 1, 0, 0, 0, 2, 0). The values of the prior parameters are considered as β_1= (0.9, 1.9), β_2 = (2.0, 3.5) and p = (0.7, 0.35) based on the method described before the algorithm-1. A Type – II censored sample is also generated using the above mentioned inputs to observe a comparison between performances of estimators using the progressive Type – II censoring and Type – II censoring.
To generate the progressive Type – II censored sample for the mixture model we use the following algorithm.

Algorithm-2

Step - 1: A uniform random number (u) is generated from $U(0,1)$ and if $u \leq p$ (mixture proportion parameter) then select first sub-population $f_1(x)$ having parameter β_1, otherwise second sub-population $f_2(x)$ having parameter β_2.
Step - 2: To generate say r observations from first or second sub-population ($r = m_1 \; or \; m_2$), generate r uniform random numbers $u_1, u_2, \dots, u_r \sim U(0,1)$
Step - 3: Set $\xi_i = \ln(1 - u_i)\,; i = 1, \dots, r$
Step - 4: Let $y_1 = \frac{\xi_1}{r}$ and $y_i = y_{i-1} + \frac{\xi_i}{n - \sum_{j=1}^{i-1} R_j - i + 1}; i = 2, 3, \dots, r$ with $\sum_{i=1}^{r} R_i = n - r$
Step - 5: $x_i = (1 - e^{-y_i})^{\frac{1}{\beta}}; i = 1, 2, \dots, r$ where β = β_1 or β_2.
Step - 6: Calculate ML estimates and Bayes estimates of parameters p, β_1 and β_2 using the respective formulas from section.
Step - 7: Repeat the steps 1 - 5 for S times, thus we have $\widehat{\beta_{1l}}$, $\widehat{\beta_{2l}}$ and $\widehat{p_l}$, $i = 1,2, \dots, S$.
Step - 8: Calculate Bayes estimates of β_1, β_2, and p by taking average of the S values in step 7.
Step - 9: Calculate Root Mean Square Error and PR, using the formula,

$$MSE = \frac{\sum_{i=1}^{S} (\widehat{\theta_i} - \theta)^2}{S}$$

The outputs obtained from the simulations are presented in Table A.1 to Table A.6.

6. Discussion of Numerical Results

From Table A.1 to Table A.6 we observe the following conclusions:

i. The Bayes estimates are better compared to ML estimates based on MSE for both the censoring schemes that are considered.
ii. For the considered values of the parameters β_1, β_2 and p as m increases, the MSE of the estimates decreases for both the censoring schemes for MLEs as well as Bayes estimators.

iii. The Bayes estimator under the K-loss function performs better with respect to MSE compared to squared error loss function for all the values of β_1 and β_2 under both types of censoring schemes adopted.
iv. As the values of the parameters β_1 and β_2 increases, the MSE also increases, for Progressive Type - II censoring scheme.
v. ML estimates and Bayes estimates give almost similar amount of bias.

7. Real Life Example

In this section the analysis of real-life dataset of failure of electrical cables is performed which is presented by Lawless (2003). The test involved 20 cables each with two types of insulations which are considered as Population – I and Population – II respectively. The purpose is to determine whether the estimators have the similar behavior for real life data as it was for simulated data.

The Kolmogorov-Smirnov test is performed to determine whether the data follows log logistic distribution. The calculated value of KS test statistic is 0.1868 for Population – I and 0.0715 for Population – II. The degrees of freedom for Population – I and Population – II are 20 and the test is performed at 5% level of significance. The KS tabulated value for one sample test at 5% level of significance and 20 degrees of freedom is 0.294. The results clearly indicate that the Population – I and Population – II fits well to the log logistic distribution.

From the original data we have prepared progressive type – II censored sample with $n = 40,\ p = 0.5, m_1 = 13, m_2 = 12,\ R = (1,0,0,0,1,1,0,2,0,0,1,0,0,0,0,2,0,0,0,2,0,0,0,0,5)$. The censored mixture data is:

Population – I: $32.0, 35.4, 39.8, 41.2, 45.5, 46.0, 46.2, 46.5, 47.3, 47.3, 49.2, 50.4, 56.3$
and
Population – II: $45.3, 49.2, 51.3, 53.2, 53.2, 55.5, 57.1, 57.5, 59.2, 62.4, 63.8, 67.7.$

The results are obtained using Bayes and Maximum Likelihood Estimation approaches for the above mentioned real-life dataset and are given in the Table A.7.

The analysis under real-life data supports the findings obtained from the simulation study. The Bayes estimates are better compared to the ML estimates and all the three loss functions SELF, KLF and PLF give similar results. This gives us more confidence to suggest the use of Bayes estimation for Progressive Type - II Censored Data using Mixture of Log Logistic Distributions.

8. Conclusion

In this paper, a two component mixture model based on log logistic distributions has been proposed. Maximum likelihood and Bayesian estimation have been used to estimate the parameters of mixture model under progressive Type – II censoring. Three types of loss functions namely, SELF, KLF and PLF are used. The posterior likelihood based on progressive Type – II censoring has no closed form due to which it is not possible to apply numerical integration. Importance sampling method was used to solve this. Finally, we observe that for precise estimation of the unknown parameters of log logistic distribution, Bayes estimation is preferable over maximum likelihood estimation under all the three types of loss functions and this holds true for real life data as well.

References

Abbas, K. and Tang, Y. (2016). Objective Bayesian Analysis for Log-logistic Distribution. *Communications in Statistics - Simulation and Computation*, **45** (**8**), 2782-2791.

Ahmed, E. A. (2014). Bayesian estimation based on progressive Type-II censoring from two-parameter bathtub-shaped lifetime model: a Markov chain Monte Carlo approach. *Journal of Applied Statistics*, **41** (**4**), 752-768.

Al-Shomrani, A.A., Shawky, A.I., Arif, O.H. and Aslam, M. (2016). Log-logistic distribution for survival data analysis using MCMC. *SpringerPlus* **5** (**1774**), 1–16.

Balakrishnan, N. and Aggarwala, R. (2000). *Progressive Censoring – Theory, Methods and Applications*, Birkhauser Publishers, Boston.

Bhavsar, S. S. and Patel, M. N. (2019). Bayes estimation of the mixture of power function distribution using Type – II censored sample. *International Journal of Agriculture Statistical Science*, **15** (**2**), 881 – 893.

Fernández, A.J. (2004). On estimating exponential parameters with general type II progressive censoring. *Journal of Statistical Planning and Inference*, **121** (**1**), 135-147.

Gajjar, K. A. and Patel, M. N. (2008). Estimation for a mixture of exponential distribution based on progressively Type – II censored sample. *International Journal of Agriculture Statistical Science*, **4** (*1*), 169 - 176.

Kazmi, S.M.A., Aslam, M. and Ali, S. (2012). On the Bayesian estimation for two component mixture of Maxwell distribution, assuming Type - I censored data. *International Journal of Applied Science and Technology*, **2** (**1**), 197 – 218.

Kim, C., Jung, J. and Chung, Y. (2011). Bayesian estimation for the exponentiated Weibull model under Type II progressive censoring. *Statistical Papers*, **52**, 53–70.

Kim, C., and Han, K. (2009). Estimation of the scale parameter of the Rayleigh distribution under general progressive censoring. *Journal of the Korean Statistical Society*, **38** (**3**), 239-246.

Kumar, K. (2018). Classical and Bayesian estimation in log-logistic distribution under random censoring. *International Journal of System Assurance Engineering and Management*, **9**, 440–451.

Kundu, D. and Pradhan, B. (2009). Estimating the Parameters of the Generalized Exponential Distribution in Presence of Hybrid Censoring. *Communications in Statistics - Theory and Methods*, **38** (**12**), 2030-2041.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* 2nd Edition. New York: Wiley.

Madi, M. T. and Raqab, M. Z. (2009). Bayesian Inference for the Generalized Exponential Distribution on Progressively Censored Data. *Communications in Statistics – Theory and Methods*, **38** (**12**), 2016 - 2029.

Mendenhall, W. and Hader, R.A. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504 – 520.

Norstrom, J.G. (1996). The use of precautionary loss functions in risk analysis. *IEEE Transactions on Reliability*, **45** (**3**), 400 – 403.

Patel, N. W. and Patel, M. N. (2007). Progressively Type – II censored sample from the Geometric lifetime model. *Journal of Probability and Statistical Science*, **5** (**1**), 81 – 95.
Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London – A*, **185**, 71 – 110.

Saleem, M. and Aslam, M. (2009). On prior selection for the mixture of Rayleigh distribution using predictive Intervals. *Pakistan Journal of Statistics*, **24**, 21 – 35.

Saleem, M., Aslam, M. and Economou, P. (2010). On the Bayesian analysis of the mixture of Power function distribution using the complete and the censored sample. *Journal of Applied Statistics*, **37** (**1**), 25 – 40.

Saraçoğlu, B., Kinaci, I. and Kundu, D. (2012). On estimation of R=P(Y<X) for exponential distribution under progressive Type - II censoring. *Journal of Statistical Computation and Simulation*, **82** (**5**), 729-744.

Sultana, F., Tripathi, Y. M., Rastogi, M. K. and Wu, S. J. (2018). Parameter Estimation for the Kumaraswamy Distribution Based on Hybrid Censoring. *American Journal of Mathematical and Management Sciences*, **37** (**3**), 243 – 261.

Sultana, F., Tripathi, Y. M., Sen, T. and Wu, S. J. (2020). Inference for Kumaraswamy Distribution Based on Type I Progressive Hybrid Censoring. *Annals of Data Science*, doi: 10.1007/s40745-020-00283-z.

Wasan, M. (1970). *Parametric Estimation*. New York : McGraw-Hill Book Company.

Wu, S. J., Chen, D. H. and Chen, S. T. (2006). Bayesian inference for Rayleigh distribution under progressive censored sample. *Applied stochastic models in Business and Industry*, **22**, 269 – 279.

Yahaya, A. and Ibrahim, I.A. (2019). Bayesian Approach of Estimating Shape Parameter of Log Logistic Distribution. *Nigerian Journal of Technological Research*, **14** (**1**), doi: 10.4314/njtr.v14i1.7.

Appendix A

Table A.1: MLE & Bayes Estimates, PR and MSE for $(\boldsymbol{n}, \boldsymbol{m}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{p}) = (\mathbf{60}, \mathbf{15}, \mathbf{0.9}, \mathbf{2.0}, \mathbf{0.7})$

MLE / Bayes	Statistic	$\widehat{\boldsymbol{\beta_1}}$	$\widehat{\boldsymbol{\beta_2}}$	$\widehat{\boldsymbol{p}}$	Type of Censoring
MLE	Estimate	1.14	2.98	0.60	Progressive Type - II
MLE	MSE	0.11	1.18	0.02	Progressive Type - II
SELF	Estimate	1.11	2.86	0.59	Progressive Type - II
SELF	PR	0.05	0.20	0.03	Progressive Type - II
SELF	MSE	0.07	0.80	0.03	Progressive Type - II
KLF	Estimate	1.08	2.82	0.56	Progressive Type - II
KLF	PR	0.08	0.05	0.20	Progressive Type - II
KLF	MSE	0.06	0.76	0.03	Progressive Type - II
PLF	Estimate	1.13	2.90	0.61	Progressive Type - II
PLF	PR	0.04	0.07	0.04	Progressive Type - II
PLF	MSE	0.07	0.84	0.02	Progressive Type - II
MLE	Estimate	1.03	3.69	0.49	Type - II
MLE	MSE	0.05	3.11	0.06	Type - II
SELF	Estimate	0.97	3.48	0.53	Type - II
SELF	PR	0.04	0.24	0.02	Type - II
SELF	MSE	0.02	2.31	0.04	Type - II
KLF	Estimate	0.95	3.45	0.50	Type - II
KLF	PR	0.08	0.04	0.21	Type - II
KLF	MSE	0.02	2.25	0.05	Type - II
PLF	Estimate	0.99	3.52	0.55	Type - II
PLF	PR	0.04	0.07	0.04	Type - II
PLF	MSE	0.03	2.36	0.04	Type - II

Table A.2: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (60,\ 15,\ 1.9,\ 3.5,\ 0.35)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\hat{p}$	Type of Censoring
MLE	Estimate	2.63	4.66	0.28	Progressive Type - II
MLE	MSE	0.68	1.94	0.01	Progressive Type - II
SELF	Estimate	2.55	4.52	0.31	Progressive Type - II
SELF	PR	0.15	0.60	0.02	Progressive Type - II
SELF	MSE	0.46	1.32	0.01	Progressive Type - II
KLF	Estimate	2.52	4.45	0.26	Progressive Type - II
KLF	PR	0.05	0.06	0.84	Progressive Type - II
KLF	MSE	0.43	1.24	0.01	Progressive Type - II
PLF	Estimate	2.58	4.59	0.35	Progressive Type - II
PLF	PR	0.06	0.13	0.06	Progressive Type - II
PLF	MSE	0.49	1.39	0.01	Progressive Type - II
MLE	Estimate	2.25	5.22	0.25	Type - II
MLE	MSE	0.27	3.44	0.01	Type - II
SELF	Estimate	2.19	4.96	0.28	Type - II
SELF	PR	0.14	0.48	0.02	Type - II
SELF	MSE	0.12	2.40	0.01	Type - II
KLF	Estimate	2.16	4.91	0.24	Type - II
KLF	PR	0.06	0.04	0.77	Type - II
KLF	MSE	0.11	2.33	0.02	Type - II
PLF	Estimate	2.22	5.01	0.31	Type - II
PLF	PR	0.06	0.10	0.06	Type - II
PLF	MSE	0.14	2.47	0.01	Type - II

Table A.3: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (60,\ 25,\ 0.9,\ 2.0,\ 0.7)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\hat{p}$	Type of Censoring
MLE	Estimate	1.02	2.40	0.68	Progressive Type - II
MLE	MSE	0.04	0.45	0.01	Progressive Type - II
SELF	Estimate	1.01	2.34	0.67	Progressive Type - II
SELF	PR	0.02	0.25	0.02	Progressive Type - II
SELF	MSE	0.02	0.20	0.01	Progressive Type - II
KLF	Estimate	1.00	2.29	0.66	Progressive Type - II
KLF	PR	0.04	0.09	0.08	Progressive Type - II
KLF	MSE	0.02	0.18	0.01	Progressive Type - II
PLF	Estimate	1.02	2.40	0.68	Progressive Type - II
PLF	PR	0.02	0.11	0.02	Progressive Type - II
PLF	MSE	0.02	0.23	0.01	Progressive Type - II
MLE	Estimate	1.01	2.62	0.66	Type - II
MLE	MSE	0.03	0.81	0.01	Type - II
SELF	Estimate	0.98	2.50	0.66	Type - II
SELF	PR	0.02	0.33	0.02	Type - II
SELF	MSE	0.02	0.38	0.01	Type - II
KLF	Estimate	0.97	2.43	0.65	Type - II
KLF	PR	0.04	0.11	0.08	Type - II
KLF	MSE	0.01	0.33	0.01	Type - II
PLF	Estimate	0.99	2.56	0.67	Type - II
PLF	PR	0.02	0.13	0.02	Type - II
PLF	MSE	0.02	0.42	0.01	Type - II

Table A.4: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (60,\ 25,\ 1.9,\ 3.5,\ 0.35)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{p}$	Type of Censoring
MLE	Estimate	2.41	3.97	0.34	Progressive Type - II
MLE	MSE	0.44	0.69	0.01	Progressive Type - II
SELF	Estimate	2.32	3.92	0.35	Progressive Type - II
SELF	PR	0.16	0.40	0.02	Progressive Type - II
SELF	MSE	0.22	0.34	0.01	Progressive Type - II
KLF	Estimate	2.29	3.87	0.32	Progressive Type - II
KLF	PR	0.06	0.05	0.36	Progressive Type - II
KLF	MSE	0.20	0.31	0.01	Progressive Type - II
PLF	Estimate	2.35	3.97	0.37	Progressive Type - II
PLF	PR	0.07	0.10	0.04	Progressive Type - II
PLF	MSE	0.25	0.37	0.01	Progressive Type - II
MLE	Estimate	2.25	4.16	0.33	Type - II
MLE	MSE	0.29	0.98	0.01	Type - II
SELF	Estimate	2.18	4.03	0.35	Type - II
SELF	PR	0.15	0.44	0.02	Type - II
SELF	MSE	0.13	0.50	0.01	Type - II
KLF	Estimate	2.14	3.98	0.32	Type - II
KLF	PR	0.06	0.06	0.35	Type - II
KLF	MSE	0.11	0.46	0.01	Type - II
PLF	Estimate	2.21	4.08	0.37	Type - II
PLF	PR	0.07	0.11	0.04	Type - II
PLF	MSE	0.14	0.53	0.01	Type - II

Table A.5: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (60,\ 35,\ 0.9,\ 2.0,\ 0.70)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{p}$	Type of Censoring
MLE	Estimate	0.96	2.31	0.70	Progressive Type - II
MLE	MSE	0.02	0.40	0.00	Progressive Type - II
SELF	Estimate	0.96	2.32	0.69	Progressive Type - II
SELF	PR	0.01	0.27	0.01	Progressive Type - II
SELF	MSE	0.01	0.24	0.01	Progressive Type - II
KLF	Estimate	0.96	2.27	0.68	Progressive Type - II
KLF	PR	0.03	0.10	0.05	Progressive Type - II
KLF	MSE	0.01	0.22	0.01	Progressive Type - II
PLF	Estimate	0.97	2.38	0.70	Progressive Type - II
PLF	PR	0.01	0.12	0.02	Progressive Type - II
PLF	MSE	0.01	0.26	0.01	Progressive Type - II
MLE	Estimate	0.97	2.24	0.73	Type - II
MLE	MSE	0.02	0.58	0.00	Type - II
SELF	Estimate	0.96	2.14	0.72	Type - II
SELF	PR	0.01	0.35	0.01	Type - II
SELF	MSE	0.01	0.16	0.01	Type - II
KLF	Estimate	0.96	2.06	0.71	Type - II
KLF	PR	0.03	0.16	0.05	Type - II
KLF	MSE	0.01	0.14	0.01	Type - II
PLF	Estimate	0.97	2.22	0.72	Type - II
PLF	PR	0.01	0.16	0.01	Type - II
PLF	MSE	0.01	0.18	0.01	Type - II

Table A.6: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (60,\ 35,\ 1.9,\ 3.5,\ 0.35)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{p}$	Type of Censoring
MLE	Estimate	2.07	3.83	0.35	Progressive Type - II
MLE	MSE	0.14	0.43	0.01	Progressive Type - II
SELF	Estimate	2.07	3.84	0.36	Progressive Type - II
SELF	PR	0.10	0.29	0.01	Progressive Type - II
SELF	MSE	0.06	0.25	0.01	Progressive Type - II
KLF	Estimate	2.04	3.81	0.34	Progressive Type - II
KLF	PR	0.05	0.04	0.23	Progressive Type - II
KLF	MSE	0.06	0.24	0.01	Progressive Type - II
PLF	Estimate	2.09	3.88	0.38	Progressive Type - II
PLF	PR	0.05	0.07	0.03	Progressive Type - II
PLF	MSE	0.07	0.27	0.01	Progressive Type - II
MLE	Estimate	2.14	3.69	0.37	Type - II
MLE	MSE	0.16	0.39	0.01	Type - II
SELF	Estimate	2.11	3.64	0.38	Type - II
SELF	PR	0.09	0.30	0.01	Type - II
SELF	MSE	0.07	0.13	0.01	Type - II
KLF	Estimate	2.09	3.60	0.36	Type - II
KLF	PR	0.04	0.05	0.23	Type - II
KLF	MSE	0.06	0.12	0.01	Type - II
PLF	Estimate	2.13	3.68	0.40	Type - II
PLF	PR	0.04	0.08	0.03	Type - II
PLF	MSE	0.07	0.14	0.01	Type - II

Table A.7: MLE & Bayes Estimates, PR and MSE for $(n, m, \beta_1, \beta_2, p) = (40,\ 25,\ 0.4054896,\ 0.3840652,\ 0.5)$

MLE / Bayes	Statistic	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{p}$
MLE	Estimate	0.27960	0.26342	0.51278
SELF	Estimate	0.41468	0.39288	0.52065
SELF	PR	0.00005	0.00007	0.00228
KLF	Estimate	0.41462	0.39280	0.51853
KLF	PR	0.00058	0.00090	0.01640
PLF	Estimate	0.41474	0.39298	0.52284
PLF	PR	0.00013	0.00018	0.00438

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 247–263

Estimation in Shifted Lindley Distribution

Saran Ishika Maiti, Sourav Rana and Arindom Chakraborty
Department of Statistics, Visva-Bharati, Santiniketan, India

Received: 18 December 2020; Revised: 06 May 2021; Accepted: 09 May 2021

Abstract

In this article, we propose a shifted version of widely-used Lindley distribution. Some statistical properties such as stochastic ordering, moment generating function, reliability characteristic etc. are studied for this new distribution. For estimating unknown parameters, two types of estimation method viz. method of moments and maximum likelihood method are explored. A simulation study for several choices of parameters is executed. Finally, a real data application illustrates the performance of our proposed distribution.

Key words: Lindley distribution; Stochastic ordering; Parameter estimation; Continuous distribution; Maximum likelihood estimate.

AMS Subject Classifications: 60E05; 62G30; 62E10

1. Introduction

Lifetime distribution tries to capture, mathematically, the length of the life of a system or a device. These distributions have relevance in the fields like environmental sciences, medicine, engineering etc. To analyze lifetime data, gamma, Weibull, Rayleigh etc., distributions are widely used in statistical literature. Chief advantage of these distributions is that they only have more general mathematical closed form compared to the exponential distribution with one additional parameter. Some applied areas such as finance, lifetime analysis and insurance sometimes demand the extended forms of these distributions because there still remain many important problems involving real data in these areas, which do not fit to any of the existent classical statistical models. As a consequence, several classes of generalized distributions have been formed by extending well-known continuous distributions. These generalized distributions entend more flexibility by adding new parameters to the baseline model.

Since last decade, Lindley distribution, proposed by Lindley (1958), has been abruptly acknowledged in different setup by many authors. Pretty recently, in the context of Bayesian statistics as a counter example of fiducial statistics, the Lindley distribution has bagged considerable attention because of its flexibility. Ghitany *et al.*(2008) discussed the various statistical properties of Lindley distribution and showed its applicability over the exponential distribution. They established that in reliability analysis Lindley distribution performs better than exponential model. One of the main reasons to consider the Lindley distribution over the exponential distribution is its time dependent/increasing hazard rate.

Corresponding Author: Arindom Chakraborty
Email: arindom.chakraborty@visva-bharati.ac.in

A random variable X is said to have Lindley distribution with parameter θ if its probability density function (PDF) is defined as:

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

and corresponding (cumulative density function) (CDF) is given by

$$F(x;\theta) = 1 - \frac{\theta+1+\theta x}{1+\theta}e^{-\theta x},\ x>0,\ \theta>0 \tag{2}$$

Of late a lot of research articles came out on the extension of Lindley distribution. The motivation for all these extension stems on the flexibility of the distribution to accommodate more complex data. Some of the advances in the literature of Lindley distribution are given by Ghitany*et al.* (2011) who has introduced a two-parameter weighted Lindley distribution. Generalized Poisson Lindley distribution has been proposed by Mahmoudi *et al.* (2010). Bakouch *et al.* (2012) came up with extended Lindley (EL) distribution, Adamidis *et al.* (1998) introduced exponential geometric (EG) distribution. Shanker *et al.* (2013) introduced two-parameter Lindley distribution. Following a footstep Ghitany (2013) proposed inferential problems stemmed from power Lindley. Zakerzadeh *et al.* (2012) idealized a new two parameter lifetime distribution: model and properties. Hassan (2014) introduced convolution of Lindley distribution. Ghitany *et al.* (2015) worked on the estimation of the reliability of a stress-strength system from power Lindley distribution. Elbatal *et al.*(2013) proposed a new generalized Lindley distribution. However all these extensions were based on introducing more parameters in the constant part of the base Lindley.

The paper is organized as follows: Section 2 introduces a shifted Lindley distribution and presents its basic properties including the behaviour of the density and some results on stochastic orderings, moments, reliability characteristics. Distribution of the sum of iid random variables has also been discussed. In Section 3, estimation process of parameters is demonstrated at length. Monte Carlo simulation study is carried out in Section 4 followed by a real data analysis in Section 5. This paper concludes with some discussions in Section 6.

2. The Shifted Lindley distribution

The extension, proposed in this article, is completely different. We define.

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

as a Shifted Lindley distribution with parameters (θ,μ). It will be denoted by $SL(\theta,\mu)$. The CDF of a Shifted Lindley distribution with parameters (θ,μ) is given by

$$F(x;\theta,\mu) = 1 - \frac{1+\theta(1+x)}{1+\theta(1+\mu)}e^{-\theta(x-\mu)}, x>\mu>0 \tag{4}$$

Note that if we put $\mu=0$ in equations (3) and in (4), these equations become the PDF and CDF, respectively, of a Lindley distribution with a single parameter θ.

The shape of shifted Lindley distribution depends on its parameters. Figure 1 shows the pdf and cfd of shifted Lindley distribution for some choices of μ and θ. This figure reveals that for smaller θ shifted Lindley pdf exhibits right skewed while for larger θ it looks as an inverted J, more specifically tapering to a standard exponential curve.

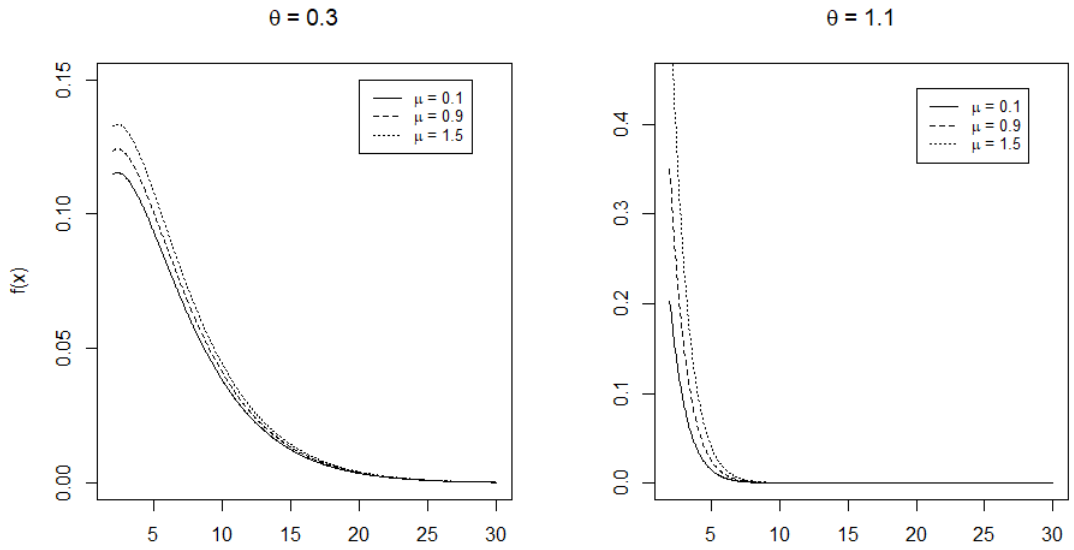


Figure 1: The PDF's of various Shifted Lindley distributions for different values of parameters

2.1. Stochastic orders

One of the main objectives of statistics is the comparison of random quantities. These comparisons are mainly based on the comparison of some measures associated to these random quantities. Stochastic ordering of positive continuous random variables is an important tool for judging such comparative behavior. Suppose X_i is distributed as $SL(\mu_i, \theta_i)$, $i = 1, 2$. Let F_i denote the cumulative distribution of X_i and f_i denote the probability density function of X_i. A random variable X_1 is said to be smaller than a random variable X_2 in the

- Stochastic order ($X_1 \leq_{st} X_2$) if $F_1(x) \geq F_2(x)$ for all x.
- Hazard rate order ($X_1 \leq_{hr} X_2$) if $h_1(x) \geq h_2(x)$ for all x.
- Likelihood ratio order ($X_1 \leq_{Lr} X_2$) if $\frac{f_1(x)}{f_2(x)}$ decreases in x.

In order to establish stochastic ordering of distributions we refer the following result from Shaked *et al.* (1994).

$$X_1 \leq_{LR} X_2 \Longrightarrow X_1 \leq_{hr} X_2 \Longrightarrow X_1 \leq_{st} X_2.$$

Taking a cue from this above-mentioned result, a pair of theorems are proposed regarding the stochastic ordering pattern of $SL(\theta, \mu)$ for different choices of (θ, μ).

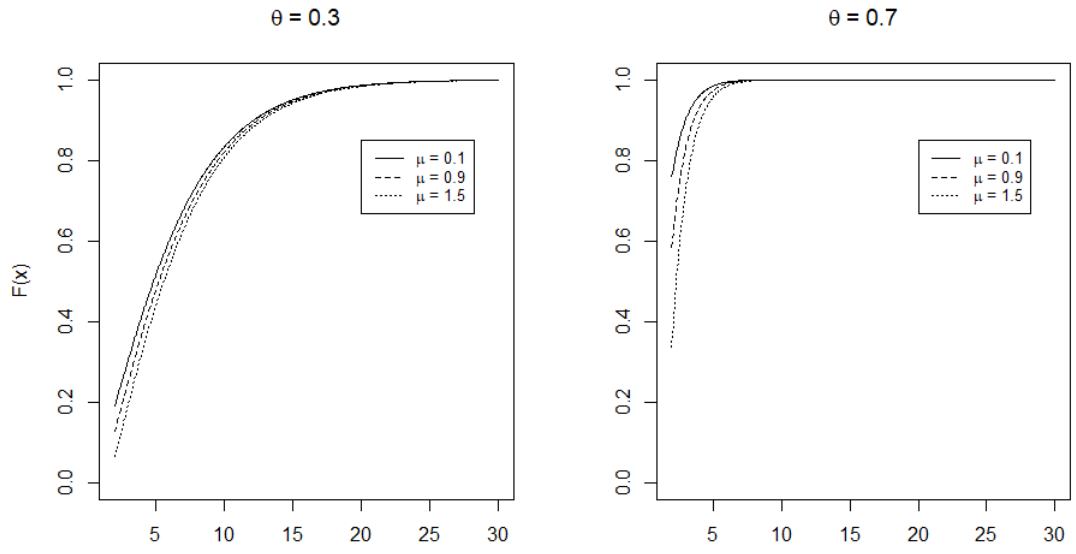


Figure 2: The CDF's of various Shifted Lindley distributions for different values of parameters

Theorem 1: Let $X_1 \sim SL(\theta_1, \mu_1)$ and $X_2 \sim SL(\theta_2, \mu_2)$. If $\mu_1 = \mu_2$ and $\theta_2 < \theta_1$, then $X_1 \leq_{Lr} X_2$ and hence $X_1 \leq_{hr} X_2$ and $X_1 \leq_{st} X_2$.

Proof:

Assume $\mu_1 = \mu_2$. Then $\frac{\delta}{\delta x} ln \frac{f_1(x)}{f_2(x)} = \theta_2 - \theta_1$. So $\frac{\delta}{\delta x} ln \frac{f_1(x)}{f_2(x)} < 0$ if $\theta_2 < \theta_1$ implying $\frac{f_1(x)}{f_2(x)} \downarrow x$.
This means that $X_1 \leq_{Lr} X_2$ and hence $X_1 \leq_{hr} X_2$ and $X_1 \leq_{st} X_2$.

Theorem 2: Let $X_1 \sim SL(\theta_1, \mu_1)$ and $X_2 \sim SL(\theta_2, \mu_2)$. If $\theta_1 = \theta_2 = \theta > 0$ and $\mu_1 > \mu_2$; then $X_1 \geq_{st} X_2$.

Proof:

The ratio of two pdf's does not involve x. So the technique adopted in checking of likelihood ratio ordering fails. Therefore, we would head to investigate via ratio of two corresponding distribution functions and hence directly infering on stochastic ordering of the distribution.

$$\frac{F_1(x)}{F_2(x)} = \frac{1 - \frac{1+\theta(1+x)}{1+\theta(1+\mu_1)} e^{-\theta(x-\mu_1)}}{1 - \frac{1+\theta(1+x)}{1+\theta(1+\mu_2)} e^{-\theta(x-\mu_2)}} = \frac{1 - \left[1 + \frac{\theta(x-\mu_1)}{1+\theta(1+\mu_1)}\right] e^{-\theta(x-\mu_1)}}{1 - \left[1 + \frac{\theta(x-\mu_2)}{1+\theta(1+\mu_2)}\right] e^{-\theta(x-\mu_2)}}.$$

Assume $\mu_1 > \mu_2$. Then $1 + \theta(1 + \mu_1) > 1 + \theta(1 + \mu_2)$ and $\theta(x - \mu_1) < \theta(x - \mu_2)$

$$\begin{aligned} \frac{\theta(x-\mu_1)}{1+\theta(1+\mu_1)} &< \frac{\theta(x-\mu_2)}{1+\theta(1+\mu_2)} \\ 1 + \frac{\theta(x-\mu_1)}{1+\theta(1+\mu_1)} &< 1 + \frac{\theta(x-\mu_2)}{1+\theta(1+\mu_2)}. \end{aligned} \tag{5}$$

Also

$$-e^{-\theta(x-\mu_1)} < -e^{-\theta(x-\mu_2)} \tag{6}$$

Combining (5) and (6) we have

$$1-\left[1+\frac{\theta(x-\mu_1)}{1+\theta(1+\mu_2)}\right]e^{-(x-\mu_1)} < 1-\left[1+\frac{\theta(x-\mu_2)}{1+\theta(1+\mu_2)}\right]e^{-(x-\mu_2)}$$

which results $F_1 < F_2$ upon further simplification. Consequently $X_1 \geq_{st} X_2$. Therefore if $\mu_1 > \mu_2$, $X_1 \geq_{st} X_2$ and vice-versa.

2.2. Moments

In applications, moments are necessary and very important. Through moments, it is possible to study many of the interesting characteristics and features of a distribution. The mean of the distribution can be obtained as:

$$\begin{aligned}\mu_1' = E(X) &= \frac{\theta^2}{1+\theta(1+\mu)}\int_\mu^\infty x(1+x)e^{-\theta(x-\mu)}dx \\ &= \mu+\frac{2}{\theta}-\frac{1+\mu}{1+\theta(1+\mu)}\end{aligned} \tag{7}$$

To find all higher order moment we will use the following result:

Theorem 3: For $k \geq 0$, the recurrence relation for the higher order moments are

$$\mu_{k+1}' = \mu_1'\mu_k' - \frac{d}{d\theta}\mu_1' \tag{8}$$

Proof:

$$\begin{aligned}\frac{d}{d\theta}\mu_k' &= \int_\mu^\infty x^k(1+x)\left[\frac{\theta^2}{1+\theta(1+\mu)}(-1)(x-\mu)e^{-\theta(x-\mu)}\right. \\ &+ e^{-\theta(x-\mu)}\left.\left[\frac{2\theta}{1+\theta(1+\mu)}-\frac{\theta^2(1+\mu)}{(1+\theta(1+\mu))^2}\right]\right] \\ &= -\frac{\theta^2}{1+\theta(1+\mu)}\int_\mu^\infty x^{k+1}(1+x)\ e^{-\theta(x-\mu)}\,dx \\ &+ \mu\frac{\theta^2}{1+\theta(1+\mu)}\int_\mu^\infty x^k(1+x)\ e^{-\theta(x-\mu)}\,dx \\ &+ \left[\frac{2\theta}{1+\theta(1+\mu)}-\frac{\theta^2(1+\mu)}{(1+\theta(1+\mu))^2}\right]\int_\mu^\infty x^k(1+x)\ e^{-\theta(x-\mu)}\,dx \\ &= -\mu_{k+1}' + \mu_1'\mu_k' \\ i.e.& \\ \mu_{k+1}' &= \mu_1'\mu_k' - \frac{d}{d\theta}\mu_k'\end{aligned}$$

Hence the proof.

Putting $k = 1$ we get,

$$\begin{aligned}\mu_2' &= \Big(\mu + \frac{2}{\theta} - \frac{1+\mu}{1+\theta(1+\mu)}\Big)\mu_1' - \frac{d}{d\theta}\mu_1' \\ &= \Big(\mu + \frac{2}{\theta} - \frac{1+\mu}{1+\theta(1+\mu)}\Big)^2 - \Big[-\frac{2}{\theta^2} + \frac{(1+\mu)^2}{(1+\theta(1+\mu))^2}\Big]\end{aligned}$$

and hence

$$\begin{aligned}\mu_2 &= \mu_2' - \mu_1'^2 \\ &= \frac{2}{\theta^2} - \frac{(1+\mu)^2}{(1+\theta(1+\mu))^2}\end{aligned}$$

Putting $\mu = 0$ will imply

$$\mu_2 = \frac{2}{\theta^2} - \frac{1}{(1+\theta)^2}$$

which is the variance of a Lindley distribution with parameter θ. Similarly it can be shown that, for a $SL(\mu, \theta)$ distribution

$$\mu_3 = \frac{4}{\theta^3} - \frac{2(1+\mu)^3}{(1+\theta(1+\mu))^3}$$

and

$$\mu_4 = \frac{24}{\theta^4} - \frac{3(1+\mu)^4}{(1+\theta(1+\mu))^4} - \frac{12(1+\mu)^2}{\theta^2(1+\theta(1+\mu))^2}$$

In reference to the moments of shifted Lindley distribution, next we present heat plot (Figure 3) which unravels the intertwining effect of parameters μ and θ on mean, variance, skewness and kurtosis. Heat plot (or heatmap) is a data visualization technique that shows impact of variables in terms of intensity of color in two dimensions. The variation in color exhibits obvious visual clues about the relationship between two categories. From the matrix layout with color and shading of heat plot furnished here, it comes up that the mean and variance drop as both the parameters increase while the skewness and kurtosis shoot up with the increase of both parameters.

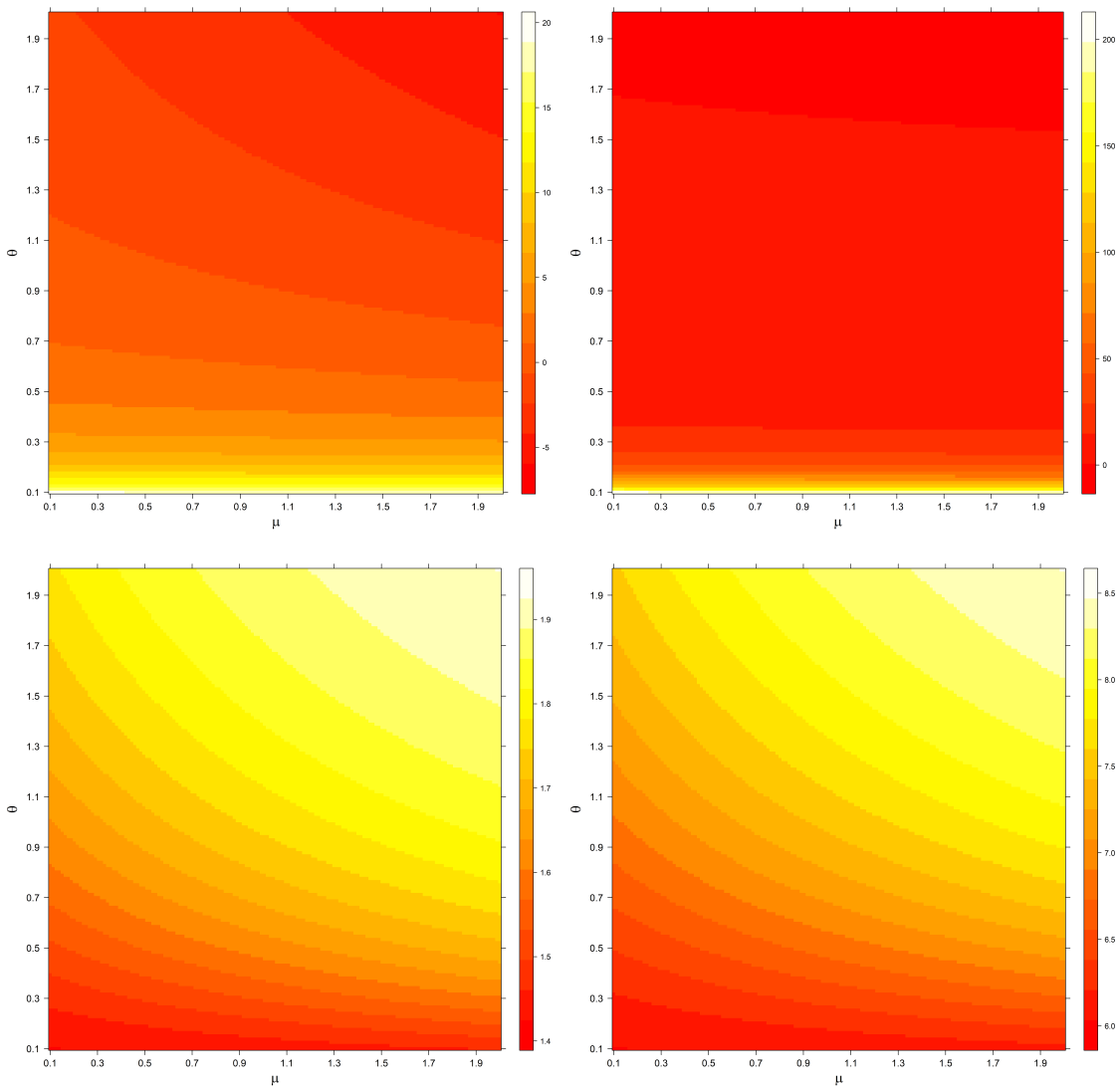


Figure 3: The mean, variance, skewness and kurtosis (from left to right) of the shifted Lindley distributions with respect to the parameters μ and θ.

2.3. Moment generating function (MGF)

In this subsection, we derived the MGF of $SL(\mu, \theta)$ distribution.

Theorem 4: If $X \sim SL(\mu, \theta)$, then the moment generating function $M_X(t)$ has the following form:

$$M_X(t) = \theta^2 \frac{[1 + (\theta - t)(1 + \mu)]}{(\theta - t)^2[1 + \theta(1 + \mu)]} e^{t\mu}, \ |t| < \theta \tag{9}$$

Proof:

$$\begin{aligned}
M_X(t) &= E(e^{tX}) \\
&= \frac{\theta^2}{1+\theta(1+\mu)} \int_{\mu}^{\infty} e^{tX}(1+x)e^{-\theta(x-\mu)}dx \\
&= \frac{\theta^2}{1+\theta(1+\mu)} \int_{\mu}^{\infty} (1+x)e^{-\theta(x-\mu)+tx}dx \\
&= \frac{\theta^2}{1+\theta(1+\mu)} \int_{\mu}^{\infty} (1+x)e^{-(\theta-t)(x-\mu)}e^{t\mu}dx \\
&= \frac{\theta^2}{1+\theta(1+\mu)} e^{t\mu} \int_{\mu}^{\infty} (1+x)e^{-(\theta-t)(x-\mu)}dx \\
&= \frac{\theta^2}{1+\theta(1+\mu)} e^{t\mu} \frac{1+(\theta-t)(1+\mu)}{(\theta-t)^2} \\
&= \theta^2 \frac{[1+(\theta-t)(1+\mu)]}{(\theta-t)^2[1+\theta(1+\mu)]} e^{t\mu}
\end{aligned}$$

So, when $\mu = 0$

$$\begin{aligned}
M_X(t) &= \frac{\theta^2[1+(\theta-t)]}{(\theta-t)^2(1+\theta)} \\
&= \frac{1}{1+\theta} \quad \left[\frac{\theta^2}{(\theta-t)^2} + \frac{\theta^2}{(\theta-t)}\right]
\end{aligned}$$

which coincides the MGF of a Lindley distribution with parameter θ.

In the same way the characteristic function of the shifted Lindley distribution becomes as follows.

$$\phi_X(t) = M_X(it) = \frac{\theta^2}{1+\theta(1+\mu)} \left[\frac{1}{(\theta-it)^2} + \frac{(1+\mu)}{(\theta-it)}\right] e^{it\mu} \tag{10}$$

where $i = \sqrt{-1}$ is the unit imaginary number.

2.4. Quantile function

Let X denotes a random variable with the probability distribution function Eq. (4). The quantile function , say $Q(p)$, defined by $F(Q(p)) = p$ is the root of the equation

$$\frac{1+\theta(1+Q(p))}{1+\theta(1+\mu)} e^{-\theta(Q(p)-\mu)} = 1-p \tag{11}$$

for $0 < p < 1$. On further simplification of the equation (11),

$$\begin{aligned}
1+\theta(1+Q(p))e^{-\theta(Q(p)-\mu)} &= [1+\theta(1+\mu)](1-p) \\
(1+Q(p))e^{-\theta Q(p)} &= \left[\{1+\theta(1+\mu)\}\frac{1-p}{\theta} - \frac{1}{\theta}\right] e^{-\theta\mu}
\end{aligned}$$

Substituting $Z(p) = 1 + Q(p)$ in the above

$$\begin{aligned} Z(p)e^{-\theta Z(p)}e^{\theta} &= \left[\{1+\theta(1+\mu)\}\frac{1-p}{\theta} - \frac{1}{\theta}\right] e^{-\theta\mu} \\ (-\theta Z(p))e^{-\theta Z(p)} &= -\theta e^{-\theta}\left[\{1+\theta(1+\mu)\}\frac{1-p}{\theta} - \frac{1}{\theta}\right] e^{-\theta\mu} \end{aligned}$$

So, the solution for $Z(p)$ is

$$Z(p) = -\frac{1}{\theta}W\Big(-\theta e^{-\theta}\left[\{1+\theta(1+\mu)\}\frac{1-p}{\theta} - \frac{1}{\theta}\right] e^{-\theta\mu}\Big)$$

for $0 < p < 1$, where $W(.)$ is the Lambert W function (see Corless *et al.* (1996)). Inverting the above equation in $Z(p)$

$$Q(p) = -\frac{1}{\theta}W\Big(\left[1 - \{1+\theta(1+\mu)\}(1-p)\right] e^{-\theta(\mu+1)}\Big) - 1,$$

for $0 < p < 1$.

2.5. Distribution of the sum of iid RVs

Theorem 5: If $X_1, X_2, \cdots, X_n$ are IID RVs from $SL(\mu, \theta)$, then the pdf of $Z = X_1 + X_2 + \cdots + X_n$ is

$$g(z; n, \theta, \mu) = \sum_{k=0}^{n} p_{k,n}(\theta, \mu) f_{SG}(z; 2n-k, \mu, \theta) \tag{12}$$

where $p_{k,n}(\theta,\mu) = \binom{n}{k}\frac{(\theta(1+\mu))^k}{(1+\theta(1+\mu))^n}$ and $f_{SG}(z; k, \mu, \theta) = \frac{\theta[\theta(x-\mu)]^{k-1}}{\Gamma k}e^{-\theta(x-\mu)}$, the pdf of a shifted gamma (SG) distribution (or Pearson type III) with parameters (k, μ, θ).

Proof: Recall that, if $X_1 \sim SL(\mu, \theta)$ then the pdf of X_1 is

$$f_{X_1}(x; \theta, \mu) = \frac{\theta^2}{1+\theta(1+\mu)}(1+x)e^{-\theta(x-\mu)} = \frac{(1+\mu)\theta}{1+\theta(1+\mu)}f_{SG(x;1,\theta,\mu)} + \frac{1}{1+\theta(1+\mu)}f_{SG(x;2,\theta,\mu)}$$

Next let us have the following lemma to expedite the proof.

Lemma 6: If $X \sim SG(k, \mu, \theta)$ then moment generating function of X is given by

$$M_X(t) = (1 - t/\theta)^{-k}\, e^{t\mu}.$$

Proof:

$$\begin{aligned} M_X(t) &= \frac{\theta^k}{\Gamma k}\int_{\mu}^{\infty}(x-\mu)^{k-1} \quad e^{tx-\theta x+\theta\mu}dx \\ &= \frac{\theta^k}{\Gamma k}\int_{0}^{\infty} z^{k-1} \quad e^{-\theta z+tz+t\mu}dz \qquad \text{put}(x-\mu)=z \\ &= \frac{\theta^k}{\Gamma k} \quad e^{t\mu}\int_{0}^{\infty} z^{k-1} \quad e^{-(\theta-t)z}dz \\ &= \frac{\theta^k}{\Gamma k} \quad e^{t\mu}\frac{\Gamma k}{(\theta-t)^k} = \frac{\theta^k}{(\theta-t)^k} \quad e^{t\mu} \\ &= \frac{1}{(1-t/\theta)^k} \quad e^{t\mu} = (1-t/\theta)^{-k} \quad e^{t\mu} \end{aligned}$$

The moment generating function (MGF) for X_1 for $|t|<\theta$ is

$$M_{X_1(t)} = E(e^{tX_1}) = \left(\frac{\theta(1+\mu)}{1+\theta(1+\mu)}(1-t/\theta)^{-1} + \frac{1}{1+\theta(1+\mu)}(1-t/\theta)^{-2}\right)e^{(t\mu)}$$

Hence the MGF of Z for $|t|<\theta$ is

$$\begin{aligned} M_z(t) = E(e^{tZ}) &= \left(\frac{\theta(1+\mu)}{1+\theta(1+\mu)}(1-t/\theta)^{-1} + \frac{1}{1+\theta(1+\mu)}(1-t/\theta)^{-2}\right)^n e^{(nt\mu)} \\ &= \left[\frac{1}{(1+\theta(1+\mu))}\right]^n\Big[\theta(1+\mu)(1-t/\theta)^{-1}+(1-t/\theta)^{-2}\Big]^n e^{(nt\mu)} \\ &= \left[\frac{1}{(1+\theta(1+\mu))}\right]^n(1-t/\theta)^{-2n}\Big[\theta(1+\mu)(1-t/\theta)+1\Big]^n e^{(nt\mu)} \\ &= \left[\frac{1}{(1+\theta(1+\mu))}\right]^n(1-t/\theta)^{-2n}\sum_{k=0}^{n}\binom{n}{k}\Big[\theta(1+\mu)(1-t/\theta)\Big]^k e^{(nt\mu)} \\ &= \left[\frac{1}{(1+\theta(1+\mu))}\right]^n\sum_{k=0}^{n}\binom{n}{k}(\theta(1+\mu))^k(1-t/\theta)^{-(2n-k)}e^{(tn\mu)} \end{aligned}$$

Using the Lemma stated above, Theorem 5 follows.

2.6. Reliability characteristics of shifted Lindley distribution

In present section, we consider shifted Lindley distribution as a lifetime model and study different reliability characteristics. The reliability function of the $SL(\mu,\theta)$ distribution is given by:

$$R(t) = P(X>t) = 1-F(t) \tag{13}$$

The mean time to system failure (MTSF) is same as: We know that the hazard function $h(x)$ can be computed as

$$h(t) = \frac{f(t;\theta,\mu)}{1-F(t;\theta,\mu)}$$

which implies

$$h(t) = \theta^2(1+t) \tag{14}$$

The cumulative hazard function $H(x)$ is defined as

$$H(x) = -\log(1 - F(x;\theta,\mu)) = -\log\Big(\frac{\theta+1+\theta x}{1+\theta}e^{-\theta x}\Big) = -\log(R(x))$$

and the failure rate average (fra) is defined by $FRA(x) = H(x)/x$, where $x > \mu$. The conditional survival of t is:

$$R(x|t) = \frac{R(x+t)}{R(t)};\ \theta, R(.) > 0;\ t, x > \mu,\ \mu > 0$$

2.7. Rényi entropy

Entropy is used to measure the randomness of systems, and it is widely used in areas like physics, molecular imaging of tumors and sparse kernel density estimation. If X has the probability distribution function $f(.)$, Rényi entropy is defined by

$$I_\delta(x) = \frac{1}{1-\delta}\log\Big(\int_0^\infty f^\delta(x)\,dx\Big),\ \delta > 0,\ \delta \neq 1.$$

Using equation (3), it is observed that

$$\begin{aligned} f^\delta(x) &= \frac{\theta^{2\delta}}{[1+\theta(1+\mu)]^\delta}(1+x)^\delta e^{-\theta\delta(x-\mu)} \\ &= \frac{\theta^{2\delta}}{[1+\theta(1+\mu)]^\delta}\sum_{i=0}^{\delta}\binom{\delta}{i}x^i e^{-\theta\delta(x-\mu)} \end{aligned}$$

After some algebra, the Rényi entropy of X is reduces to

$$I_\delta(x) = \frac{1}{1-\delta}\log(\sum_{i=0}^{\delta} e_i)$$

where, $e_i = \frac{1}{(\theta\delta)^{i+1}}\Gamma(i+1, \theta\delta\mu)$, $\Gamma(.)$ is the incomplete Gamma function.

3. Estimation

Here, we consider two estimation methods: the methods of moments and maximum likelihood estimation. We provide expressions for the associated Fisher information matrix. Suppose $X_1, X_2, \cdots, X_n$ is a random sample from equation (3). For the moments estimation,

let $m_1 = (1/n)\sum_{i=1}^{n} X_i$ and $m_2 = (1/n)\sum_{i=1}^{n}(X_i - m_1)^2$. By equating the theoretical moments of equation (3) with the sample moments, the following equations are obtained.

$$m_1 = \mu + \frac{2}{\theta} - \frac{1+\mu}{1+\theta(1+\mu)} \tag{15}$$

$$m_2 = \frac{2}{\theta^2} - \frac{(1+\mu)^2}{[1+\theta(1+\mu)]^2} \tag{16}$$

Solving (15), (16) we can estimate the parameters μ and θ.

3.1. Maximum likelihood (ML) estimation of parameters

The likelihood function for a random sample $X_1, X_2, ..., X_n$ which is taken from $SL(\mu, \theta)$ distribution is:

$$L(\boldsymbol{X}, \mu, \theta) = \frac{\theta^{2n}}{(1+\theta(1+\mu))^n}\Big[\prod_{i=1}^{n}(1+x_i)\Big]e^{-\theta\sum_{i=1}^{n}(x_i-\mu)} \tag{17}$$

It is to be noted that mle of μ is

$$\hat{\mu}_{mle} = \min_i X_i = X_{(1)} \tag{18}$$

Differentiating the log-likelihood w.r.t. θ, we get the following equation:

$$\frac{2n}{\theta} - \frac{n(1+\hat{\mu}_{mle})}{1+\theta(1+\hat{\mu}_{mle})} - \sum_{i=1}^{n}(X_i - \hat{\mu}_{mle}) = 0 \tag{19}$$

which needs to be solved using some iterative procedure.

4. Simulation Study

It may be noted that it is not possible to generate samples from shifted Lindley distribution using the inversion of the CDF. It has already been established that an $SL(\mu, \theta)$ distribution can be viewed as a mixture of two shifted gamma distributions. This property is devised to generate random sample from simulation study. In R, function *rgamma3* of package *FAdist* generates samples from a shifted gamma distribution. Using the aforesaid function a convenient sampling scheme for data generation can be framed as follows.

To estimate the parameters μ and θ, we have generated $10,000$ samples from the shifted Lindley distribution. We have considered four different combinations of the parameter to study their influence. Then using the sample moments and equations (15) and (16), we obtain the moment estimates of μ and θ. We have replicated these processes 50, 100 and 500 times and computed standard error of corresponding estimates. The ML estimates are obtained by using the equations (18) and (19) and respective standard errors have been computed using the above techniques.

Algorithm

1. Select values of θ and μ
2. Calculate weight $w = \frac{(1+\mu)\theta}{1+\theta(1+\mu)}$
3. Generate U from $U(0,1)$
4. If $U < w$, generate a sample from $f_{SG(x;1,\theta,\mu)}$ else from $f_{SG(x;2,\theta,\mu)}$ using $rgamma3(1, shape, scale, thres)$

The simulation study is carried out with $N = 10,000$ sample size for $(\mu, \theta) = (0.5, 0.3)$, $(1.5, 1.1)$, $(0.5, 1.1), (1.5, 0.3)$ and replication $n = (50, 100, 500)$. The following measures are calculated to asses the simulation results:
$\hat{\theta}$ and $\hat{\mu}$, estimates obtained through both of the case along with the corresponding standard error of estimates (SE), $Bias_\mu = \sum_{i=1}^{n} \frac{\widehat{\mu}_j - \mu}{n}$, magnitude of relative error $= MRE_\mu = \sum_{i=1}^{n} \frac{\widehat{\mu}_j/\mu}{n}$, mean square error $MSE_\mu = \sum_{i=1}^{n} \frac{(\widehat{\mu}_j - \mu)^2}{n}$, $Bias_\theta = \sum_{i=1}^{n} \frac{\widehat{\theta}_j - \theta}{n}$, $MRE_\theta = \sum_{i=1}^{n} \frac{\widehat{\theta}_j/\theta}{n}$, $MSE_\theta = \sum_{i=1}^{n} \frac{(\widehat{\theta}_j - \theta)^2}{n}$. Results against the parameter θ are shown in Table 3 and that of parameter μ in Table 4. Both of the tables are placed in Appendix at the end of this article.
From Table 3 and Table 4, it may be observed that moment estimators for both are also performing well in terms of small biases. As expected the MRE values are found close to 1, whereas the MSE values are tending close to 0. This study also reveals that moment and ML estimators are equally efficient.

5. Real Data Analysis

The proposed distribution is fitted for a data set available in Duffy *et al.*(1993). The data consists of measurements on strength of the sintered silicon nitride after four-point bend system is applied. On four point bend specimen, the support span of test fixture was 40.373 mm and the inner load span of 19.622 mm. All specimens are subjected to pure four-point bending. Number of complete specimens in the data set is found to be 27. We apply Lindley and shifted Lindley in order to fit this data. Subject to the fitting of shifted Lindley distribution on the data we figure out estimates of the parameters θ and μ by both moment and maximum likelihood method. Estimates alongwith standard errors (SE) are given in Table 1. For the shifted Lindley distribution,it can be seen that both the methods are producing different estimates.

Table 1: Parameter estimates for the four point bend data

Distribution	$\widehat{\mu}_{ML}$ (SE)	$\widehat{\mu}_{MOM}$ (SE)	$\widehat{\theta}_{ML}$ (SE)	$\widehat{\theta}_{MOM}$ (SE)
Lindley	0	0	0.0027	0.0027
	(-)	(-)	(0.0001)	(0.0001)
shifted Lindley	613.9	654.873	0.0096	0.014
	(0.0006)	(0.0531)	(0.0001)	(0.0001)

For further comparison between two distributions fitted to the data, we also report some model selection criteria— Akaike information criterion (AIC), Bayesian information Crite-

Table 2: Model selection criteria for the four point bend data

Distribution	KS(p-value)	AIC	BIC	$CAIC$	$HQIC$
Lindley	2.4(0.823)	391.7628	393.0586	391.9109	392.1481
Shifted Lindley	0.632(0.001)	315.8982	318.49	316.36	316.669

rion (BIC), Corrected AIC (CAIC) and Hannan and Quinn information criterion (HQIC). The definitions used for these selection tools are as: AIC= $-2ln\,L(\theta)+2k$, CAIC= $-2ln\,L(\theta)+2k\frac{n}{n-k-1}$; BIC= $-2ln\,L(\theta)+k\,ln(n)$; and HQIC= $-2ln\,L(\theta)+2k\,ln\{ln(n)\}$, where $lnL(\theta)$ denotes log likelihood, n being the number of observations and k being the number of parameters of the distribution. These are reported in Table 2. Considering all the model selection criteria, reported in Table 2, we found that shifted Lindley fits the data well compared to Lindley distribution. The Kalmogorov-Smirnoff (KS) statistic for shifted Lindley is found to be 0.632 with a p-value of 0.001 confirming the claim of better fit.

6. Conclusion

In this study we have proposed a new distribution called shifted Lindley distribution. Some mathematical properties along with estimation issues are addressed. The hazard rate function of shifted Lindley distribution shows that the subject distribution can be used to model reliability data as well. We derived the moment and maximum likelihood estimates of the parameters along with the biases, mean square error and mean relative errors. A real data application of the shifted Lindley distribution projects that it could provide a meaningful fit than a set of usual statistical distributions, while being considered specially in life time data analysis. A further extension of shifted Lindley might be thought in the context of power Lindley distribution, thereby a comparative study on relative quality of statistical models for a given set of data can be delved into.

7. Acknowledgement

We would like to thank the editor and an anonymous referee for pointing out some important issues which have undoubtedly enhanced the quality of the work. The authors are indebted to Dr. Kashinath Chatterjee for his inputs.

References

Adamidis, K. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics and Probability Letters*, **39**, 35-42.

Bakouch, H. S., Al-Zahrani, B. M., Al-Shomrani, A. A., Marchi, V. A. and Louzada, F. (2012). An Extended Lindley distribution. *Journal of the Korean Statistical Society*, **41**, 75-85.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey D. J. and Knuth D. J. (1996). On the Lambert W function. *Advanced Computational Math*, **5**, 329-59.

Duffy, F. S., Powers, L. M. and Starlinger, A. (1992). Reliability analysis of structural ceramic components using a three parameter Weibull distribution, *Civil and Environmental Engineering faculty publications*, **3**.

Elbatal, I., Merovci, F. and Elgarhy, M. (2013). A New Generalized Lindley Distribution. *Mathematical Theory and Modeling*, **3(13)**, 30-47.

Ghitany, M. E., Atieh, B. and Nadarajah, S. (2008). Lindley Distribution and its application, *Mathematics and computers in simulation*, **78(4)**, 36-38.

Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan, N. and Al-Enezi, L. J. (2013). Power Lindley distribution and associated inference. *Computational Statistics and Data Analysis*, **64**, 20-33.

Ghitany, M. E., Alqallaf, F., Al-Mutairi, D. K. and Hussain, H. A. (2011). A Two Parameter Weighted Lindley Distribution and Its Applications to Survival Data, *Mathematics and Computers in Simulation*, **81(6)**, 1190-1201.

Ghitany, M. E., Al-Mutairi, D. K. and Aboukhamseen, S. M. (2015). Estimation of the Reliability of the Stress-Strength System from Power Lindley Distributions, *Communications in Statistics-Simulation and Computations*, **44(1)**, 118-136.

Hassan, M. K. (2014). On the Convolution of Lindley Distribution. *Columbia International Publishing Contemporary Mathematics and Statistics*, **2(1)**, 47-54.

Mahmoudi, E. and Zakerzadeh, H. (2010). Generalized Poisson Lindley distribution. *Communications in Statistics:Theory and Methods*, **30**, 1785-1798.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society B*, **20**, 102–107.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications*, Academic Press, Boston (USA).

Shanker, R., Sharma, S. and Shanker, R. (2013). A Two-Parameter Lindley Distribution for Modeling Waiting and Survival Times Data. *Applied Mathematics*, **4**, 363-368.

Zakerzadeh, H. and Mahmoudi, E. (2012). A new two parameter lifetime distribution: model and properties. arXiv:12014:4248 VI [Stat.CO].

ANNEXURE

Table 3: Estimate, SE, Bias, MSE and MRE of (θ) based on the Moment and ML estimation methods

Replication	μ	θ	**Moment Estimator**(θ)					**ML Estimator** (θ)				
			$\hat{\theta}$	SE	Bias	MSE	MRE	$\hat{\theta}$	SE	Bias	MSE	MRE
50	0.5	0.3	.3001	.0005	0.00011	1.24E-05	1.00037	0.29992	0.00034	-8.00E-05	5.65E-06	0.99973
	1.5	1.1	1.09760	0.00171	-0.00239	0.00015	0.99782	1.09855	0.00109	-0.00145	6.08E-05	0.99868
	0.5	1.1	1.10084	0.001191	0.000847	7.03E-05	1.00077	1.09918	0.00088	-0.00082	3.83E-05	0.99925
	1.5	.3	0.30036	0.00047	0.00036	1.09E-05	1.0012	0.30014	0.00036	0.00014	6.48E-06	1.00046
100	0.5	0.3	0.29996	0.000291	-3.77E-05	8.34E-06	0.99987	0.30024	0.000234	0.000237	5.47E-06	1.0008
	1.5	1.1	1.09861	0.00131	-0.001391	0.00017	0.99873	1.09871	0.00088	-0.00128	7.98E-05	0.99882
	0.5	1.1	1.09958	0.00133	-0.00041	0.00017	0.99962	1.09997	0.000847	-2.45E-05	7.11E-05	0.99997
	1.5	0.3	0.30003	0.00033	2.82E-05	1.09E-05	1.00009	0.30017	0.000268	0.00017	7.16E-06	1.00058
500	0.5	0.3	0.29992	0.000126	-7.15E-05	8.02E-06	0.99976	0.29995	9.49E-05	-5.04E-05	4.50E-06	0.99983
	1.5	1.1	1.1008	0.00058	0.00079	0.00017	1.00072	1.10043	0.00041	0.00043	8.56E-05	1.00039
	0.5	1.1	1.09996	0.000543	-3.65E-05	0.00015	0.99997	1.09991	0.00041	-9.22E-05	8.02E-05	0.99991
	1.5	0.3	0.30032	0.00013	0.00032	9.13E-05	9.13E-06	0.30006	9.76E-05	6.59E-05	4.76E-06	1.00022

Table 4: Estimate, SE, Bias, MSE and MRE of (μ) based on the Moment and ML estimation methods

Replication	μ	θ	**Moment Estimator**(μ)					**ML Estimator** (μ)				
			$\hat{\mu}$	SE	Bias	MSE	MRE	$\hat{\mu}$	SE	Bias	MSE	MRE
50	0.5	0.3	0.50645	0.01143	0.00645	0.00644	1.01290	0.50128	0.00015	0.00128	2.80E-06	1.00256
	1.5	1.1	1.49876	0.001575	-0.00123	0.000123	0.99917	1.50013	1.46E-05	0.00012	2.73E-08	1.00008
	0.5	1.1	0.50267	0.00156	0.00267	0.00013	1.00534	0.50015	2.33E-05	0.000147	4.82E-08	1.00029
	1.5	0.3	1.50655	0.00911	0.00655	0.00411	1.00436	1.50083	0.00013	0.00083	1.50E-06	1.00055
100	0.5	0.3	0.48804	0.00851	-0.01195	0.00732	0.97608	0.50117	0.000125	0.00117	2.93E-06	1.00234
	1.5	1.1	1.49986	0.00117	-0.00014	0.00014	0.99990	1.50012	1.23E-05	0.00012	3.06E-08	1.00008
	0.5	1.1	0.50103	0.00125	0.00103	0.00016	1.00207	0.50014	1.32E-05	0.00014	3.67E-08	1.00028
	1.5	0.3	1.49556	0.006429	-0.00444	0.00411	0.99704	1.50085	8.22E-05	0.00085	1.39E-06	1.00057
500	0.5	0.3	0.49843	0.003470	-0.00156	0.00601	0.99687	0.50118	5.14E-05	0.00118	2.72E-06	1.00237
	1.5	1.1	1.50048	0.00051	0.00048	0.00013	1.00032	1.50013	5.88E-06	0.00012	3.37E-08	1.00008
	0.5	1.1	0.50009	0.00058	9.00E-05	0.000169	1.00018	0.500123	6.30E-06	0.00014	3.89E-08	1.00027
	1.5	0.3	1.50738	0.002759	0.00738	0.00385	1.00492	1.50074	3.12E-05	0.00074	1.03E-06	1.00049

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 265-277

Bayesian Modeling of VAR Model with Multiple Covariates

Umme Afifa[1], Varun Agiwal[2] and Jitendra Kumar[1]
[1]*Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India*
[2]*Department of Community Medicine, Jawaharlal Nehru Medical College, Ajmer, India*

Received: 20 September 2020; Revised: 26 January 2021; Accepted: 07 May 2021

Abstract

This paper aims to provide a comprehensive overview of the Bayesian estimation methodology for the multiple covariate vector autoregressive (MC-VAR) model, in both methodology and application point of view. In that respect, conditional posterior distributions are derived to obtain the Bayesian estimators and influence based on covariate is analyzed by posterior odds ratio. Due to multiple integrations, the Gibbs sampler method is employed for the estimation of the MC-VAR model. Our approach is applied on both simulation and real data series to show the applicability of the proposed model. The real data result is useful for analyzing the relationship of covariates in economic time series.

Key words: Bayesian inference; Covariate; Vector autoregressive model.

AMS Subject Classification: 62K99; 62J05.

1. Introduction

The VAR model is continuously attracting the researchers to depict the behavior of a variable over time (Al-hajj *et al.* (2017), Sharma *et al.* (2018)). In the VAR model, present value of the variable is expressed as a linear function of past values and a random error (Fuller (1985)). For the analysis of multivariate time series data, VAR is frequently used model see Fuller (1985), Juselius (2006), Tsay (2014) and Lutkepohl (2005). VAR is most preferred and equally popular model for understanding the behavior of financial and economical data in various literatures see Wei (1990), Lutkepohl (2005), Al-hajj *et al.* (2017) and Sharma *et al.* (2018). VAR model is/was also used for structural analysis. In structural analysis, causal impacts of the variables are observed when certain hypotheses are imposed and resultant causal impacts are précised in Granger causality and impulse response function (IRF) in Wei (1990), Hamilton (1994), Lutkepohl (2005), etc. In the VAR model, when one includes exogenous variables, the VAR model extended to a covariate vector autoregressive (CVAR) model and allows those variables in the dataset to be modeled jointly over present and past time periods as considered in Hamilton (1994) and Tsay (2015).

The main motive behind the study of time series model with covariate is to make precise inferences about the impact of covariates on the response series under Bayesian framework. There are so many articles to explore the covariate in various univariate and multivariate time series model. Hansen (1995) developed covariate augmented Dickey-Fuller (CADF) unit root test with some stationary covariates for autoregressive parameter. This CADF test further extended to a point optimal covariate (POC) unit root test by Elliott and Jansson (2003).

Corresponding Author: Jitendra Kumar
E-mail: vjitendrav@gmail.com

Costantini and Lupi (2013) developed panel data model with stationary covariate which is the extension of Hansen (1995) model. Chaturvedi *et al.* (2017) discussed the Bayesian unit root hypothesis for covariate autoregressive model. Chang *et al.* (2017) developed bootstrap unit root tests with covariate method to the CADF test to deal the nuisance parameter dependency and provided a valid basis for inference based on the CADF test. Anggraeni *et al.* (2017) discussed performance of autoregressive integrated moving average with explanatory variable (ARIMAX) with VAR model using Indonesia economic data sets. Based on MAPE results, observed that performance of ARIMAX model is better than VAR model. Kumar *et al.* (2018) discussed Bayesian estimation and testing procedure for panel autoregressive time series model with covariate. Recently, Ji Linying *et al.* (2019) implemented VAR model with non-ignorable missingness in dependent variables and covariates under Bayesian framework. They introduced a Bayesian model which simultaneously represents the time dependency in multivariate and multiple subject time series data via VAR model.

The purpose of present paper is to make inference of multiple covariate-vector autoregressive (MC-VAR) model under Bayesian framework. We use Monte Carlo simulation method to estimate the parameters using conditional posterior distributions and then testing the impact of stationary covariate using posterior odds ratio. A simulation study has been carried out to validate the theoretical results. An empirical study of GDP series with export and import series as covariates is carried out to evaluate the performance of proposed model and obtained the Bayes estimators.

2. Model Description

In this section, we begin with vector autoregressive (VAR) model that captures the complex dynamics behaviour of multiple time series and their interactions and provides multiple series in a systematic manner. The basic form of VAR model represents a vector of N-dimensional time series measured at a particular time period. Let $\{Y_t,\ t = 1,2,\ldots,T\}$ be a VAR process expresses as a linear combination of past observations at lag p. Then, model is defined as

$$Y_t = \mu + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \eta_t. \tag{1}$$

Generally, influence regarding the observation is based not only the study variables but also other associated variables. If these associated variables are also included in the process, then efficiency of the process may be increased. So, we include K stationary covariates ($W_t,\ t = 1,2,\ldots,T$) in the model that having dependence with its own past observations up to lag q then it can be written as

$$Y_t = \mu + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \sum_{l=1}^{K} B_{l1} W_{lt-1} + \ldots + \sum_{l=1}^{K} B_{lq} W_{lt-q} + \eta_t,$$

$$Y_t = \mu + \sum_{i=1}^{p} A_i Y_{t-i} + \sum_{l=1}^{K} \sum_{j=1}^{q} B_{lj} W_{lt-j} + \eta_t,$$

$$Y_t = X_t^{'}\theta + Z_t^{'}\delta + \eta_t. \tag{2}$$

Here,

$$Y_t = \begin{pmatrix} y_{1t} & y_{2t} & \cdots & y_{Nt} \end{pmatrix}', \qquad Y = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_T \end{pmatrix}', \qquad Y = \begin{pmatrix} y_{11} & y_{21} & \cdots & y_{N1} \\ y_{12} & y_{22} & \cdots & y_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1T} & y_{2T} & \cdots & y_{NT} \end{pmatrix},$$

$$\eta_t = \begin{pmatrix} \eta_{1t} & \eta_{2t} & \cdots & \eta_{Nt} \end{pmatrix}', \qquad \eta = \begin{pmatrix} \eta_1 & \eta_2 & \cdots & \eta_T \end{pmatrix}', \qquad \eta = \begin{pmatrix} \eta_{11} & \eta_{21} & \cdots & \eta_{N1} \\ \eta_{12} & \eta_{22} & \cdots & \eta_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{1T} & \eta_{2T} & \cdots & \eta_{NT} \end{pmatrix},$$

$$\theta = \begin{pmatrix} \mu & A_1 & \cdots & A_P \end{pmatrix}', \qquad A_i = \begin{pmatrix} A_{1i} & A_{2i} & \cdots & A_{Ni} \end{pmatrix}, \qquad A_{ij} = \begin{pmatrix} \alpha_{i1}^{(j)} & \alpha_{i2}^{(j)} & \cdots & \alpha_{iN}^{(j)} \end{pmatrix}'$$

$$\gamma_l = \begin{pmatrix} B_{l1} & B_{l2} & \cdots & B_{lq} \end{pmatrix}, \qquad B_{li} = \begin{pmatrix} B_{1i}^l & B_{2i}^l & \cdots & B_{Ni}^l \end{pmatrix}, \qquad B_{ij}^l = \begin{pmatrix} \beta_{i1}^{(lj)} & \beta_{i2}^{(lj)} & \cdots & \beta_{iN}^{(lj)} \end{pmatrix},$$

$$\delta = \begin{pmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_K \end{pmatrix}, \qquad X_t' = \begin{pmatrix} 1 & Y_{t-1} & \cdots & Y_{t-p} \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 & X_2 & \cdots & X_T \end{pmatrix}',$$

$$Z = \begin{pmatrix} Z_1 & Z_2 & \cdots & Z_T \end{pmatrix}', \qquad Z_t' = \begin{pmatrix} Z_{1t}' & Z_{2t}' & \cdots & Z_{Kt}' \end{pmatrix}, \qquad Z_{lt}' = \begin{pmatrix} W_{lt-1} & W_{lt-2} & \cdots & W_{lt-q} \end{pmatrix}.$$

The final model in terms of matrix notation as

$$Y = X\theta + Z\delta + \eta \tag{3}$$

where N is the numbers of variables under study, K is the number of covariates, Y_t and η_t are 1 x N, A_{ij} is N x N. The disturbances η_t are unobservable random variable with $E(\eta_t) = 0$ and $VAR(\eta_t) = \Sigma$. The model is multiple covariate-vector autoregressive of order p time series model.

3. Bayesian Inference

The following prior distributions are considered for Bayesian analysis. We consider a basic prior distribution that enables analytical derivation of the posterior distribution and, thus, fast computations. The matrix variate normal conditional prior distribution is considered for θ and δ. An inverse Wishart marginal prior distribution is assumed for Σ. Let us assume the following prior distributions for the parameters used in the models

$$\theta|\Sigma \sim MN(\theta_0, \Sigma, V_1), \qquad \theta_0 \in \Re;\ 0 \le \Sigma, V_1 \le \infty, \tag{4}$$

$$\delta|\Sigma \sim MN(\delta_0, \Sigma, V_2), \qquad \delta_0 \in \Re;\ 0 \le \Sigma, V_2 \le \infty, \tag{5}$$

$$\Sigma \sim IW_N(S, \nu), \qquad 0 \le S \le \infty;\ \nu > N - 1. \tag{6}$$

Here *MN* and *IW* denote matrix variate normal distribution and inverse Wishart distribution, respectively. The joint prior probability of all parameters $(\Theta = (\theta, \delta, \Sigma))$ for MC-VAR model is determined using the equations (4) to (6)

$$P(\Theta)=\frac{(2\pi)^{-\frac{N(1+Np+NKq)}{2}}|V_1|^{-\frac{(1+NP)}{2}}|V_2|^{-\frac{NKq}{2}}|S|^{\frac{\nu}{2}}}{|\Sigma|^{\frac{\nu+3N+1}{2}}2^{\frac{\nu N}{2}}\Gamma\left(\frac{\nu}{2}\right)} \exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left((\theta-\theta_0)'V_1^{-1}(\theta-\theta_0)+(\delta-\delta_0)'V_2^{-1}(\delta-\delta_0)+S\right)\right\}\right]. \quad (7)$$

The likelihood function of the model is

$$L(Y\mid\Theta)=\frac{1}{(2\pi)^{\frac{NT}{2}}|\Sigma|^{\frac{T}{2}}}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}(Y-X\theta-Z\delta)'(Y-X\theta-Z\delta)\right\}\right]. \quad (8)$$

The posterior distribution is expressed as the product of likelihood function given in equation (8) and joint prior distribution given in equation (7)

$$\begin{aligned}P(\Theta\mid Y)&=KL(Y\mid\Theta)P(\Theta)\\&=K\frac{(2\pi)^{-\frac{N(1+Np+NKq+NT)}{2}}|V_1|^{-\frac{(1+NP)}{2}}|V_2|^{-\frac{NKq}{2}}|S|^{\frac{\nu}{2}}}{|\Sigma|^{\frac{T+\nu+3N+1}{2}}2^{\frac{\nu N}{2}}\Gamma\left(\frac{\nu}{2}\right)}\exp\left[-\frac{1}{2}tr\left\{\Sigma^{-1}\left((Y-X\theta-Z\delta)'(Y-X\theta-Z\delta)\right.\right.\right.\\&\qquad\left.\left.\left.+(\theta-\theta_0)'V_1^{-1}(\theta-\theta_0)+(\delta-\delta_0)'V_2^{-1}(\delta-\delta_0)+S\right)\right\}\right]\end{aligned} \quad (9)$$

where K is the normalizing constant which is given by $K^{-1}=\int_{\Theta}L(Y\mid\Theta)P(\Theta)d\Theta$.

3.1. Bayesian estimation

For Bayesian estimation, the estimator of the parameter is derived by using loss function and the posterior distribution. We consider two loss functions, one is symmetric known as quadratic loss or squared error loss and other one is asymmetric, entropy loss function. The Bayes estimators of any parametric function, say $\phi(\Theta)$ under squared error loss function (SELF), entropy loss function (ELF) and precautionary loss function (PLF) are defined by

$$\phi_{SELF}(\Theta\mid Y)=E_\pi(\phi(\Theta)\mid Y)=K\int_{\Theta}\phi(\Theta)P(\Theta\mid Y)d\Theta, \quad (10)$$

$$\phi_{ELF}(\Theta\mid Y)=\left[E_\pi\left(\phi^{-1}(\Theta)\mid Y\right)\right]^{-1}=\left(K\int_{\Theta}\phi^{-1}(\Theta)P(\Theta\mid Y)d\Theta\right)^{-1}, \quad (11)$$

$$\phi_{PLF}(\Theta\mid Y)=\sqrt{\left[E_\pi\left(\phi^2(\Theta)\mid Y\right)\right]}=\sqrt{K\int_{\Theta}\phi^2(\Theta)P(\Theta\mid Y)d\Theta}. \quad (12)$$

It is to be noticed here that a major difficulty in the implementation of Bayes procedure is the evaluation of the ratio of two integrals as described in equations (10) to (12) for which closed expression is not easy to obtain analytically. Therefore, we use Gibbs sampler algorithm to obtain the posterior samples from posterior distribution. For this, expression of full conditional posterior distribution are obtained

$$\pi(\theta|Y,\delta,\Sigma)\sim MN\left(BA^{-1},\Sigma,A^{-1}\right), \tag{13}$$

$$\pi(\delta|Y,\theta,\Sigma)\sim MN\left(DC^{-1},\Sigma,C^{-1}\right), \tag{14}$$

$$\pi(\Sigma|Y,\theta,\delta)\sim IW\left(S^{*},v^{*}\right), \tag{15}$$

where

$$A=X^{'}X+V_1^{-1}, \qquad B=X^{'}\left(Y-Z\delta\right)+V_1^{-1}\theta_0,$$

$$C=Z^{'}Z+V_2^{-1}, \qquad D=Z^{'}\left(Y-X\theta\right)+V_2^{-1}\delta_0,$$

$$S^{*}=\left(Y-X\theta-Z\delta\right)^{'}\left(Y-X\theta-Z\delta\right)+\left(\theta-\theta_0\right)^{'}V_1^{-1}\left(\theta-\theta_0\right)+\left(\delta-\delta_0\right)^{'}V_2^{-1}\left(\delta-\delta_0\right)+S,$$

$$v^{*}=v+3N+T+1.$$

Using the generated samples from the above conditional posterior distributions, Bayes estimates of the parameters are evaluated under different loss functions. Bayes estimate under SELF, ELF and PLF is the posterior mean, $\left[E\left(\Theta^{-1}|Y\right)\right]^{-1}$ and $\sqrt{E\left(\theta^2|Y\right)}$ respectively.

3.2. Bayesian testing procedure

Under Bayesian perspective, posterior odds ratio (POR)/ Bayes Factor (BF) is used in decision making for hypothesis testing problem/model selection procedure. POR (β_{01}) is product of prior odds ratio with Bayes factor (BF_{01}) of the null (H_0) and alternative (H_1) hypothesis. Here, the null hypothesis considers that covariates do not impact the study series, *i.e.*, H_0: $\delta = 0$ against the alternative hypothesis assumes that there is a significant relationship exists between study variable and covariates, *i.e.*, H_1: $\delta \neq 0$. The model under null and alternative hypothesis is

$$\text{Under } H_0: Y=X\theta+\eta$$

$$\text{Under } H_1: Y=X\theta+Z\delta+\eta$$

Then, β_{01} is expressed as

$$\beta_{01}=\frac{P\left(H_0|y\right)}{P\left(H_1|y\right)}=O\left(H_0\right)\frac{P\left(y|H_0\right)}{P\left(y|H_1\right)}=\frac{p_0}{1-p_0}\frac{P\left(y|H_0\right)}{P\left(y|H_1\right)}$$

where $O(H_o)$ is the prior odds ratio, *i.e.*, it is the ratio of prior probability under null is p_0 and alternative is $(1-p_0)$.

The decision is taken to reject H_0 if POR is less than one, otherwise accept. So, the posterior probability under null and alternative hypothesis is computed as

$$P(y \mid H_0) = \frac{(2\pi)^{-\frac{NT}{2}} |V_1|^{-\frac{(1+NP)}{2}} |S|^{\frac{\nu}{2}} 2^{\frac{NT}{2}} \Gamma\left(\frac{T+\nu}{2}\right)}{\left|A_0^{-1}\right|^{-\frac{(1+NP)}{2}} |S_0|^{\frac{T+\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} \tag{16}$$

$$P(y \mid H_1) = \frac{(2\pi)^{-\frac{NT}{2}} |V_1|^{-\frac{(1+NP)}{2}} |V_2|^{-\frac{NKq}{2}} |S|^{\frac{\nu}{2}} 2^{\frac{NT}{2}} \Gamma\left(\frac{T+\nu}{2}\right)}{\left|A_1^{-1}\right|^{-\frac{(1+NP)}{2}} \left|C_1^{-1}\right|^{-\frac{NKq}{2}} |S_1|^{\frac{T+\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} \tag{17}$$

where

$$A_0 = A_1 = X'X + V_1^{-1}, \quad B_0 = X'Y + V_1^{-1}\theta_0,$$

$$C_1 = Z'Z + V_2^{-1} - Z'XA_1^{-1}X'Z, \quad D_1 = Z'Y + V_2^{-1}\delta_0 - Z'XA_1^{-1}\left(X'Y + V_1^{-1}\theta_0\right),$$

$$S_0 = Y'Y + \theta_0' V_1^{-1}\theta_0 + S - B_0' A_0^{-1} B_0,$$

$$S_1 = Y'Y + \theta_0' V_1^{-1}\theta_0 + \delta_0' V_2^{-1}\delta_0 + S - \left(X'Y + V_1^{-1}\theta_0\right)' A_1^{-1}\left(X'Y + V_1^{-1}\theta_0\right) - D_1' C_1^{-1} C_1.$$

Then, the POR is constructed as

$$\beta_{01} = \frac{p_0}{1-p_0}\frac{P(y \mid H_0)}{P(y \mid H_1)} = \frac{p_0}{1-p_0}\frac{\left|C_1^{-1}\right|^{-\frac{NKq}{2}} |S_1|^{\frac{T+\nu}{2}}}{|V_2|^{-\frac{NKq}{2}} |S_0|^{\frac{T+\nu}{2}}} \tag{18}$$

4. Simulation Study

This section discusses the appropriateness of the testing of hypothesis and record the performance of the estimators in the proposed model using the simulation study. For simulation purpose, a bivariate VAR(2) model with single covariate is generated from equation (2) with starting value of observed series is Y = (4 6) and covariate series is W = (2 3). The results are obtained based on R-language version 3.6.2. We have considered different sizes of the time series $T = c$ (200, 300). For series generation, fixed arbitrarily values are defined for the model parameters in the equation (19).

$$\begin{aligned}\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 0.2 & 0.3 \\ 0.25 & 0.15 \end{pmatrix}\begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} 0.15 & 0.15 \\ 0.2 & 0.1 \end{pmatrix}\begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} \\ + \begin{pmatrix} 0.3 & 0.15 \\ 0.2 & 0.1 \end{pmatrix}\begin{pmatrix} w_{1t-1} \\ w_{2t-1} \end{pmatrix} + \begin{pmatrix} 0.09 & 0.01 \\ 0.01 & 0.25 \end{pmatrix}\begin{pmatrix} \eta_{1t} \\ \eta_{2t} \end{pmatrix}\end{aligned} \tag{19}$$

where $\eta_{.t}$ is a normally distributed random variable. To get more appropriate results, process is repeated 1000 times and each time 5000 posterior samples are generated using the Gibbs sampling procedure. For the different sizes of the series, average estimates (AE) and its standard deviation (SD) are summarized in Tables 1-2 (given in Appendix). For comparison

between the different estimators, average absolute bias (AB) and average mean square error (MSE) of the estimators of the parameters are recorded in Tables 3-4 (given in Appendix).

Tables 1-2 conclude that average estimates are near to the true value of the parameters and standard deviation is also small that shows less variability in the estimation of the parameters. From Tables 3-4, we observed that size of the ser1ies increases, MSE and AB decreases from near to bottom of the series. In comparison of different loss functions, ELF performance better as compared to other estimators because MSE and AB is minimum. The next target for this study is to determine the significant affect of the covariate in the time series. For that, hypothesis testing for the presence or absence of covariate is carried out and records the POR results in Table 5 for different sizes of series and different number of covariates. From Table 5, we reject the null hypothesis as the POR values are less than one for different sizes of simulated series. This concludes that including the covariates in the model, better inference is drawn from the given series. Here, we also notice that number of covariates depends upon the series size since it does not much explain the small series with higher covariate as seen in T = 100 and K = 3.

Table 5: Posterior odds ratio with T and K

T	K=1	K=2	K=3
100	1.47E-05	3.74E-02	2.17E+00
200	1.54E-23	2.82E-21	3.70E-18
300	2.25E-47	2.63E-36	1.69E-31
400	4.07E-66	5.22E-45	3.68E-42
500	1.06E-103	2.50E-65	4.01E-57

5. Real Data Analysis

A macroeconomic data set is taken to illustrate the performance of proposed model. We use yearly series on gross domestic product (GDP), export and import for the period 1962 to 2018 from IMF's International Financial Statistics as well as The World Bank data source (http://datahelp.imf.org/ and https://data.worldbank.org/). For analysis purpose, most developing countries India and China is considered as vector form where GDP is the study variable and export and import are two covariates. The reason behind the selection of these two countries is that most of the Indian market is depended upon the import material of China product so this impacts the GDP of both countries. First, we determine the best suitable order of each variable using the in-built function in R-language and display in Table 6. Based on Table 6, we observe that GDP series have VAR order two (p = 2) model whereas import and export series obtain VAR order one (q = 1) model.

Table 6: Order selection based on various selection criterion

Series	Order (lag)	1	2	3
GDP	AIC	5.0812	4.6713	4.7246
	HQ	5.1670	4.8142	4.9248
	SC	5.3043	5.0430	5.2451
Import	AIC	1.9540	2.0192	2.0602
	HQ	2.0397	2.1621	2.2603
	SC	2.1770	2.3909	2.5806
Export	AIC	1.3716	1.4176	1.4832
	HQ	1.4574	1.5605	1.6834
	SC	1.5947	1.7893	2.0037

Once, we get the order of the series, estimation of model parameters are carried out using the proposed methodology and then obtain the consequence of covariate(s) in the response series. Here, we analyze the inference of the proposed model based on one and two covariates, *i.e*., show the suitable impact of GDP versus import or export or both series. As per simulation study, we recorded that the best estimated values of the parameters of the proposed model is obtained through ELF estimators. So, we only estimate the MC-VAR model parameters under ELF and recorded in equations (20)-(22) for single and bivariate covariates.

$$\begin{pmatrix} GDP_{India,t} \\ GDP_{China,t} \end{pmatrix} = \begin{pmatrix} 4.8678 \\ 4.5303 \end{pmatrix} + \begin{pmatrix} -0.1150 & -0.0963 \\ -0.1810 & -0.0275 \end{pmatrix} \begin{pmatrix} GDP_{India,t-1} \\ GDP_{China,t-1} \end{pmatrix} + \begin{pmatrix} -0.2342 & 0.5426 \\ 0.6116 & -0.4277 \end{pmatrix} \begin{pmatrix} GDP_{India,t-2} \\ GDP_{China,t-2} \end{pmatrix} + \begin{pmatrix} -0.5573 & -0.2204 \\ 0.2431 & 0.2526 \end{pmatrix} \begin{pmatrix} IMPORT_{India,t-1} \\ IMPORT_{China,t-1} \end{pmatrix} with\ \hat{\Sigma} = \begin{pmatrix} 3.1242 & 0.3569 \\ 0.3569 & 4.8835 \end{pmatrix} \quad (20)$$

$$\begin{pmatrix} GDP_{India,t} \\ GDP_{China,t} \end{pmatrix} = \begin{pmatrix} 5.0978 \\ 4.8146 \end{pmatrix} + \begin{pmatrix} -0.0098 & 0.5431 \\ -0.1086 & 0.0042 \end{pmatrix} \begin{pmatrix} GDP_{India,t-1} \\ GDP_{China,t-1} \end{pmatrix} + \begin{pmatrix} -0.0156 & 0.5457 \\ 0.5810 & -0.4243 \end{pmatrix} \begin{pmatrix} GDP_{India,t-2} \\ GDP_{China,t-2} \end{pmatrix} + \begin{pmatrix} -0.2777 & 0.1859 \\ -0.3034 & 0.2396 \end{pmatrix} \begin{pmatrix} EXPORT_{India,t-1} \\ EXPORT_{China,t-1} \end{pmatrix} with\ \hat{\Sigma} = \begin{pmatrix} 3.114 & 0.7960 \\ 0.7960 & 4.8539 \end{pmatrix} \quad (21)$$

$$\begin{pmatrix} GDP_{India,t} \\ GDP_{China,t} \end{pmatrix} = \begin{pmatrix} 4.7290 \\ 4.5312 \end{pmatrix} + \begin{pmatrix} -0.1298 & -0.1083 \\ -0.1928 & -0.0206 \end{pmatrix} \begin{pmatrix} GDP_{India,t-1} \\ GDP_{China,t-1} \end{pmatrix} + \begin{pmatrix} -3.1044 & 0.5108 \\ 0.6121 & -0.4286 \end{pmatrix} \begin{pmatrix} GDP_{India,t-2} \\ GDP_{China,t-2} \end{pmatrix} + \begin{pmatrix} 0.2385 & 0.1588 \\ 0.2838 & 0.2024 \end{pmatrix} \begin{pmatrix} EXPORT_{India,t-1} \\ EXPORT_{China,t-1} \end{pmatrix} + \begin{pmatrix} 0.2948 & 0.1376 \\ 0.3809 & 0.1749 \end{pmatrix} \begin{pmatrix} IMPORT_{India,t-1} \\ IMPORT_{China,t-1} \end{pmatrix} with\ \hat{\Sigma} = \begin{pmatrix} 3.0678 & 0.6221 \\ 0.6221 & 5.0160 \end{pmatrix} \quad (22)$$

The calculated POR for the proposed model under study are recorded in Table 7 that shows the suitable covariate is necessary to analyze the growth of the GDP series. We see that

individual covariate has an impact on GDP series as compared when both covariates are jointly analyzed because it does not reject the null hypothesis.

Table 7: POR value based on real data series

Covariate	Export	Import	Both Export and Import
POR	0.9266	0.4488	1241.5250

6. Conclusion

In this paper, we develop a Bayesian approach for analyzing vector autoregressive (VAR) model with multiple covariates. The model is estimated by deriving the conditional posterior distribution and Bayesian estimators are obtained under different loss functions. We also test the association of covariates in the VAR model using the derived posterior odds ratio. Based on our simulation and empirical results, indicates that Bayesian estimators appropriate estimates the parameter values and import and export variables are related to GDP series individually.

References

Al-hajj, E., Al-Mulali, U. and Solarin, S. A. (2017). The influence of oil price shocks on stock market returns: Fresh evidence from Malaysia. *International Journal of Energy Economics and Policy*, **7**(**5**), 235-244.

Sharma, A., Giri, S. S., Vardhan, H., Surange, S., Shetty, R. and Shetty, V. (2018). Relationship between crude oil prices and stock market: Evidence from India. *International Journal of Energy Economics and Policy*, **8**(**4**), 331-337.

Fuller, W. A. (1985). *Nonstationary Autoregressive Time Series*. Handbook of Statistics. Vol. **5**, Amsterdam: Elsevier Science Publishers. ssp123.

Gourieroux, C. and Monfort, A. (1997). *Time Series and Dynamic Models*. United Kingdom: Cambridge University Press.

Juselius, K. (2006). *The Cointegrated VAR Model: Methodology and Applications*. Oxford: Oxford University Press.

Lutkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.

Tsay, R. S. (2014). *Multivariate Time Series Analysis*. New Jersey: John Wiley and Sons, Inc.

Wei, W. W. S. (1990). *Time Series Analysis: Univariate and Multivariate Methods*. Redwood City, California: Addison-Wesley Publishing Company.

Hamilton, H. (1994). *Time Series Analysis*. Princeton, New Jersey: Princeton University Press.

Ji, L., Chen, M., Oravecz, Z., Cummings, E. M., Lu, Z. H., and Chow, S. M. (2020). A Bayesian vector autoregressive model with nonignorable missingness in dependent variables and covariates: development, evaluation, and application to family processes. *Structural Equation Modeling: A Multidisciplinary Journal*, **27**(**3**), 442-467.

Anggraeni, W., Andri, K. B. and Mahananto, F. (2017). The performance of ARIMAX model and vector autoregressive (VAR) model in forecasting strategic commodity price in Indonesia. *Procedia Computer Science*, **124**, 189-196.

Chang, Y., Sickles, R. C. and Song, W. (2017). Bootstrapping unit root tests with covariates. *Econometric Reviews*, **36**(**1-3**), 136-155.

Chaturvedi, A., Kumar, J. and Kumar, A. (2017). Testing for unit root in the presence of stationary covariate: A Bayesian approach. *International Journal of Intelligent Technologies and Applied Statistics*, **10**(**1**), 1-18.

Costantini, M. and Lupi, C. (2013). A simple panel-cadf test for unit roots. *Oxford Bulletin of Economics and Statistics*, **75**(**2**), 276-296.

Elliott, G. and Jansson, M. (2003). Testing for unit roots with stationary covariates. *Journal of Econometrics*, **115**(**1**), 75-89.

Hansen, B. E. (1995). Rethinking the univariate approach to unit root testing: Using covariates to increase power. *Econometric Theory*, **11**(**5**), 1148-1171.

Kumar, J., Kumar, A. and Agiwal V. (2018). Bayesian unit root test for panel AR(1) model with stationary covariate. *International Journal of Statistics and Economics*, **19**(**2**), 55-65.

APPENDIX

Table 1: AE and SD of C-Var(2) model at T = 200

Estimator	SELF		ELF		PLF	
Parameter (True Value)	AE	SD	AE	SD	AE	SD
μ_1 (1)	0.6718	0.1224	0.6624	0.1358	0.6929	0.1198
μ_2 (2)	1.6965	0.0896	1.5995	0.1067	1.6833	0.0874
$\alpha_{11}^{(1)}$ (0.2)	0.2017	0.0573	0.1835	0.0701	0.2085	0.0554
$\alpha_{12}^{(1)}$ (0.3)	0.3203	0.0424	0.3146	0.0433	0.3230	0.0420
$\alpha_{11}^{(2)}$ (0.15)	0.1468	0.0492	0.1340	0.0634	0.1542	0.0462
$\alpha_{12}^{(2)}$ (0.15)	0.1646	0.0458	0.1495	0.0562	0.1706	0.0425
$\alpha_{21}^{(1)}$ (0.25)	0.2464	0.0965	0.2635	0.0854	0.2629	0.0886
$\alpha_{22}^{(1)}$ (0.15)	0.1523	0.0695	0.1671	0.0932	0.1591	0.0615
$\alpha_{21}^{(2)}$ (0.2)	0.2101	0.0872	0.2100	0.0821	0.2249	0.0794
$\alpha_{22}^{(2)}$ (0.1)	0.1198	0.0830	0.0971	0.0692	0.0973	0.0639
$\beta_{11}^{(11)}$ (0.3)	0.3062	0.0217	0.3050	0.0218	0.3068	0.0217
$\beta_{12}^{(11)}$ (0.15)	0.1524	0.0216	0.1499	0.0220	0.1536	0.0214
$\beta_{21}^{(11)}$ (0.2)	0.2117	0.0391	0.2066	0.0406	0.2141	0.0386
$\beta_{22}^{(11)}$ (0.1)	0.1070	0.0373	0.0947	0.0735	0.1119	0.0355
Σ_{11} (0.09)	0.0739	0.0073	0.0725	0.0072	0.0747	0.0075
Σ_{12} (0.01)	0.0085	0.0670	0.0085	0.0652	0.0086	0.0681
Σ_{21} (0.01)	0.0085	0.0670	0.0085	0.0652	0.0086	0.0681
Σ_{22} (0.25)	0.1972	0.0195	0.1936	0.0191	0.1994	0.0198

Table 2: AE and SD of MC-Var(2) model at T = 300

Estimator	SELF		ELF		PLF	
Parameter (True Value)	AE	SD	AE	SD	AE	SD
μ_1 (1)	0.7122	0.1338	0.7194	0.1530	0.7304	0.1309
μ_2 (2)	1.8086	0.0600	1.7557	0.0760	1.7909	0.0620
$\alpha_{11}^{(1)}$ (0.2)	0.1974	0.0476	0.1843	0.0600	0.2026	0.0462
$\alpha_{12}^{(1)}$ (0.3)	0.3168	0.0387	0.3126	0.0392	0.3189	0.0384
$\alpha_{11}^{(2)}$ (0.15)	0.1529	0.0410	0.1524	0.3054	0.1581	0.0393
$\alpha_{12}^{(2)}$ (0.15)	0.1625	0.0404	0.1506	0.0563	0.1668	0.0391
$\alpha_{21}^{(1)}$ (0.25)	0.2483	0.0773	0.2374	0.0606	0.2596	0.0731
$\alpha_{22}^{(1)}$ (0.15)	0.1507	0.0596	0.1561	0.0756	0.1533	0.0557
$\alpha_{21}^{(2)}$ (0.2)	0.2126	0.0690	0.1893	0.1131	0.2226	0.0653
$\alpha_{22}^{(2)}$ (0.1)	0.1162	0.0650	0.1032	0.0451	0.1136	0.0511
$\beta_{11}^{(11)}$ (0.3)	0.3009	0.0182	0.3000	0.0182	0.3013	0.0181
$\beta_{12}^{(11)}$ (0.15)	0.1536	0.0171	0.1518	0.0173	0.1544	0.0170
$\beta_{21}^{(11)}$ (0.2)	0.2022	0.0289	0.1986	0.0295	0.2040	0.0287
$\beta_{22}^{(11)}$ (0.1)	0.1089	0.0272	0.1010	0.0359	0.1121	0.0263
Σ_{11} (0.09)	0.0810	0.0065	0.0800	0.0063	0.0815	0.0065
Σ_{12} (0.01)	0.0098	0.0634	0.0099	0.0607	0.0099	0.0651
Σ_{21} (0.01)	0.0098	0.0634	0.0099	0.0607	0.0099	0.0651
Σ_{22} (0.25)	0.2128	0.0176	0.2103	0.0173	0.2142	0.0178

Table 3: MSE and AB of MC-Var(2) model at T = 200

Estimator	SELF		ELF		PLF	
Parameter (True value)	MSE	AB	MSE	ABS	MSE	ABS
μ_1 (1)	0.3728	0.5515	0.3405	0.5093	0.3870	0.5674
μ_2 (2)	0.2655	0.2466	0.2683	0.2848	0.2530	0.2722
$\alpha_{11}^{(1)}$ (0.2)	0.2121	0.2193	0.2180	0.2193	0.2104	0.2203
$\alpha_{12}^{(1)}$ (0.3)	0.2246	0.2752	0.2248	0.2715	0.2246	0.2770
$\alpha_{11}^{(2)}$ (0.15)	0.2308	0.2252	0.2445	0.2311	0.2282	0.2227
$\alpha_{12}^{(2)}$ (0.15)	0.2427	0.2236	0.2481	0.2297	0.2408	0.2219
$\alpha_{21}^{(1)}$ (0.25)	0.2066	0.2432	0.2841	0.3032	0.2037	0.2455
$\alpha_{22}^{(1)}$ (0.15)	0.2525	0.2398	0.2642	0.2541	0.2481	0.2375
$\alpha_{21}^{(2)}$ (0.2)	0.2170	0.2202	0.2561	0.2443	0.2119	0.2215
$\alpha_{22}^{(2)}$ (0.1)	0.2789	0.2551	0.2540	0.2393	0.2650	0.2377
$\beta_{11}^{(11)}$ (0.3)	0.1965	0.2558	0.1965	0.2549	0.1965	0.2562
$\beta_{12}^{(11)}$ (0.15)	0.2464	0.2149	0.2472	0.2155	0.2460	0.2146
$\beta_{21}^{(11)}$ (0.2)	0.2022	0.2141	0.2031	0.2131	0.2018	0.2146
$\beta_{22}^{(11)}$ (0.1)	0.2653	0.2342	0.2753	0.2493	0.2631	0.2311
Σ_{11}(0.09)	0.2869	0.2632	0.2876	0.2642	0.2865	0.2626
Σ_{12} (0.01)	0.2462	0.2440	0.2475	0.2462	0.2454	0.2427
Σ_{21} (0.01)	0.2927	0.2734	0.2942	0.2755	0.2918	0.2721
Σ_{22} (0.25)	0.2023	0.1969	0.2029	0.1963	0.2019	0.1974

Table 4: MSE and AB of MC-Var(2) model at T = 300

Estimator	SELF		ELF		PLF	
Parameter (True Value)	MSE	ABS	MSE	ABS	MSE	ABS
μ_1 (1)	0.3874	0.5755	0.4016	0.5442	0.3612	0.5897
μ_2 (2)	0.2441	0.2555	0.2493	0.2821	0.2331	0.2768
$\alpha_{11}^{(1)}$ (0.2)	0.2244	0.2179	0.2231	0.2204	0.2280	0.2181
$\alpha_{12}^{(1)}$ (0.3)	0.2118	0.2686	0.2118	0.2657	0.2118	0.2701
$\alpha_{11}^{(2)}$ (0.15)	0.2342	0.2178	0.2325	0.2368	0.3262	0.2162
$\alpha_{12}^{(2)}$ (0.15)	0.2351	0.2157	0.2337	0.2198	0.2401	0.2149
$\alpha_{21}^{(1)}$ (0.25)	0.2190	0.2421	0.2170	0.2340	0.2201	0.2467
$\alpha_{22}^{(1)}$ (0.15)	0.2352	0.2253	0.2319	0.2775	0.2257	0.2276
$\alpha_{21}^{(2)}$ (0.2)	0.2223	0.2269	0.2200	0.2353	0.2381	0.2274
$\alpha_{22}^{(2)}$ (0.1)	0.2546	0.2431	0.2258	0.2331	0.2425	0.2279
$\beta_{11}^{(11)}$ (0.3)	0.2093	0.2561	0.2092	0.2555	0.2093	0.2564
$\beta_{12}^{(11)}$ (0.15)	0.2380	0.2136	0.2377	0.2139	0.2385	0.2134
$\beta_{21}^{(11)}$ (0.2)	0.2237	0.2139	0.2233	0.2134	0.2246	0.2142
$\beta_{22}^{(11)}$ (0.1)	0.2534	0.2328	0.2520	0.2386	0.2572	0.2308
Σ_{11} (0.09)	0.2661	0.2476	0.2658	0.2483	0.2665	0.2472
Σ_{12} (0.01)	0.2682	0.2558	0.2677	0.2572	0.2691	0.2550
Σ_{21} (0.01)	0.2728	0.2584	0.2723	0.2599	0.2737	0.2576
Σ_{22} (0.25)	0.2197	0.2147	0.2195	0.2140	0.2202	0.2151

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New Series), pp 279–295

On Uniform Truncated Poisson Distribution and its Applications

Krishnarani S. D. and Vidya V. P.
Department of Statistics
Farook College (Autonomous), Kozhikode, Kerala, India

Received: 17 November 2020; Revised: 09 May 2021; Accepted: 13 May 2021

Abstract

The Uniform Truncated Poisson distribution defined on the interval $[0, 1]$ is studied in detail and has shown that this distribution is derivable in three different ways. Analytical properties of this distribution are derived and estimation problems are addressed. Real data sets are modeled using this distribution. Generalization of the distribution on any finite interval is also considered and properties are studied.

Key words: Truncated uniform distribution; Truncated Poisson distribution; Estimation.

AMS Subject Classifications: 62E15, 60E99

1. Introduction

Theoretical probability distribution gives us a law according to which different values of the random variables are distributed with specified probabilities which can be expressed mathematically. Recent studies on probability distributions are mainly concerned with support either in the real line or positive real line. Distributions on finite intervals are less considered by the researchers. But we know that many of our real data sets are lying in finite intervals. Moreover many of the organisms in biology, experimental results in physics, chemistry, *etc.* show a uniform pattern in $[0, 1]$. Some recent distributions defined on $[0, 1]$ are available in the research papers, Altawil (2019) and Hassan *et al.* (2020). Rescaling a data into $[0, 1]$ is useful in machine learning and image processing. The most elegant and common method widely used in these fields are min-max scaling procedure. This is an alternative method to z-score normalization. By the min-max transformation discussed in this paper any random variable with support on a real line can be transformed into $[0, 1]$ and further analysis can be done. Also in neural networks $[0, 1]$ data is required for normalizing pixel intensities. As mentioned in Weigend and Gershenfeld (1993) and Yu *et al.* (2006) normalization has an important role in the data management. By this transformation all the features are kept same, but it results in smaller standard deviations of the observations, which minimizes the outlier effect. So an attempt is made to study a distribution with support on $[0, 1]$ which was mentioned in Hao and Godbole (2014). More recently a new distribution with support on $[0, 1]$ called unifed distribution has been introduced in Quijano Xacur (2019) which can be used as the response distribution for a generalized linear model. When the index parameter is unity this family gives the distribution we study in this

Corresponding Author: Krishnarani S. D.
Email: krishnaranisd@gmail.com

paper. We further explore this distribution by introducing it in another way and bringing together the relevant properties and results concerning it. We used a compounding method for the derivation of this distribution. Derivations of new discrete and continuous distributions compounding two distributions have been discussed by several authors, see for instance uniform-geometric distribution in Akdogan *et al.* (2016), binomial-Poisson distribution in Hu *et al.* (2007), and Weibull-power series distribution in Morais and Barreto-Souza (2011). Similar distributions can also been seen in Adamidis and Loukas (1998), Kus (2007), Tahmasbi and Rezaei (2008), and Chahkandi and Ganjali (2009). We have some well-known distributions like beta distribution and power function distribution with support on $[0, 1]$. These distributions are found to have useful applications in several real life situations in reliability, time series, *etc.* So we have made a comparison of the distribution studied in this paper with these well known distributions. We could also use this new distribution in the modeling of time series data. So an advanced level model diagnosis in non-linear and volatile time series data using this distribution will be quite interesting in the near future.

This paper is organized as follows. In Section 2, uniform truncated Poisson distribution is introduced and its properties are studied. Transformations are considered and corresponding distributions are derived in Section 3. The estimation of the parameter is done in Section 4 and numerical illustrations are given therein. Asymptotic properties of the estimators are also delineated in the same section. A generalization of this new distribution with support on any finite interval is done in Section 5. Application to real data sets is given in Section 6 followed by a concluding Section.

2. Uniform Truncated Poisson Distribution

Distributions defined on [0,1] are not very common in literature and the most widely used distributions belonging to this category are power function distribution and beta distribution. Several applications of the distributions defined on [0,1] have been portrayed in the introduction part. The applications of such distributions in neural networks, pixel intensities, artificial intelligence, physics, engineering, time series *etc.* are the motivation for this present study. Also for the variates in [0,1] like percentages or fractions, we have only few studies on regression/time series models. In this context some of the notable works are Kieschnick and McCullough (2003), Jara *et al.* (2013), Ristic and Popovic (2000), Rocha and Cribari-Neto (2009) and Bayer *et al.* (2018). So the distribution studied in this paper may be applied in the advanced fields of the areas mentioned above even though we have illustrated some of the applications in the last section of this paper. Now we consider the distribution on [0,1] mentioned in Hao and Godbole (2014) and propose its construction in three different ways. These methods are described below.

Method 1: We consider a transformation of the truncated uniform distribution to form a new random variable defined on [0,1]. Let U be a random variable following truncated uniform distribution with probability density function (pdf),

$$g(u) = \frac{1}{e^{\theta} - 1}, \quad 0 \leq u \leq e^{\theta} - 1.$$

Consider the transformation,

$$X = \frac{\log(1 + U)}{\theta}. \tag{1}$$

Then the probability density function of X is

$$f(x) = \frac{1}{e^{\theta}-1}\theta e^{\theta x}, \quad 0 \le x \le 1, \theta \neq 0. \tag{2}$$

It may be noted that when $\theta = 0$ the distribution is uniform in [0,1].

Method 2: The distribution specified in (2) can be derived as a solution of the first order differential equation as given below. We have used this method keeping in mind that the radioactive decay is associated with a differential equation and an exponential random variable is an example for it. So we are trying to construct a distribution with an initial value at time zero as a function of θ but the limit of the initial value function at time point zero is 1. This initial value function $\frac{\theta}{e^{\theta}-1}$ is monotone decreasing in θ. Let

$$\frac{\mathrm{d}y}{\mathrm{d}x} - \theta y = \frac{\theta}{e^{\theta}-1} \tag{3}$$

be the first order differential equation and choose $y = F(x)$.
That is

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} - \theta F(x) = \frac{\theta}{e^{\theta}-1}.$$

Solving we get

$$F(x) = \frac{e^{\theta x}-1}{e^{\theta}-1}$$

$$\bar{F}(x) = 1 - F(x) = \frac{e^{\theta}-e^{\theta x}}{e^{\theta}-1}$$

and hence $f(x)$ is of the form (2). So this distribution is the solution of the first order differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} - \theta y = \frac{\theta}{e^{\theta}-1}.$$

Method 3: Random minimum or maximum of N independent and identically distributed (i.i.d) random variables are studied in Louzada *et al.* (2011), Kus (2007), Cancho *et al.* (2011) and several other papers. It may be noted that Hao and Godbole (2014) has introduced the uniform-Poisson model, deriving it as given below. They have applied the method mentioned above and considered only a few properties in that paper. So using the procedure used there, assuming the random variable N to be truncated Poisson with probability mass function

$$P(N = n) = \frac{e^{-\theta}\theta^{N}}{N!(1-e^{-\theta})}, \quad N = 1, 2, ...$$

and $X_1, X_2, ..., X_N$ i.i.d $U[0,1]$ with distribution function F(.), the distribution of $X = \min_{1 \le i \le N}(X_i)$ is,

$$g(X = x) = f(x) \sum_{n=1}^{\infty} n \left(F(x)\right)^{N-1} P(N = n)$$

which is exactly the same as (2). Hence we call this random variable X with pdf (2) as Uniform Truncated Poisson distribution (θ) denoted as UTPD(θ).

It is quite interesting to note that UTPD(θ) is derived in three different ways. Now we look at the properties of this new distribution.

2.1. Properties

1. The survival function is $\frac{e^{\theta}-e^{\theta x}}{e^{\theta}-1}$.
2. The hazard function, $h(x) = \frac{f(x)}{\bar{F}(x)} = \frac{\theta e^{\theta x}}{e^{\theta}-e^{\theta x}} = \frac{\theta}{e^{\theta(1-x)}-1}$.

   It can be seen that for all θ values, the distribution has an increasing failure rate (IFR).
3. The characteristic function, $\phi_X(t) = \frac{\theta}{\left(e^{\theta}-1\right)} \frac{1}{(\theta+it)} \left(e^{\theta+it}-1\right)$.
4. The r^{th} moment of UTPD is given by

$$\begin{aligned} E(X^r) &= \frac{e^{\theta}}{e^{\theta}-1}\left[1-\frac{r}{\theta}+\frac{r(r-1)}{\theta^2}-\frac{r(r-1)(r-2)}{\theta^3}\right. \\ &\quad \left.+\frac{r(r-1)(r-2)(r-3)}{\theta^4}-...+(-1^r)\frac{r(r-1)(r-2)(r-3)...1}{\theta^r}\right] \\ &\quad +\frac{1}{e^{\theta}-1}(-1)^{r+1}\frac{r(r-1)(r-2)(r-3)...1}{\theta^r}, \text{ for } r=1, 2, ... \end{aligned}$$

5. Mean= $\frac{e^{\theta}(\theta-1)+1}{\theta\left(e^{\theta}-1\right)}$.
6. Variance= $\frac{e^{\theta}}{e^{\theta}-1}\left(1-\frac{2}{\theta}+\frac{2}{\theta^2}\right)-\frac{2}{\theta^2\left(e^{\theta}-1\right)}-\left(\frac{e^{\theta}(\theta-1)+1}{\theta\left(e^{\theta}-1\right)}\right)^2$.
   For $\theta = 1$, Mean=$\frac{1}{(e-1)}$, $Variance = \frac{e^2-3e+1}{(e-1)^2}$.
7. The p^{th} quantile is given by $x_p = \frac{1}{\theta}\log\left\{1+p\left(e^{\theta}-1\right)\right\}$, $0 \le p \le 1$.
8. Entropy, a measure of the uncertainty associated with the random variable is given by $H(X) = \frac{-\theta}{\theta-1}\left\{\frac{\ln\theta e^{\theta}}{\theta}-\ln\theta\left(\frac{1}{\theta}\right)+\frac{e^{\theta}}{\theta}-\frac{1}{\theta^2}\left(e^{\theta}-1\right)-\ln\left(e^{\theta}-1\right)\right\}$.
9. Odds ratio : Odds ratios are often used in the medical literature.
   (a) The odds ratio of surviving beyond time, $\phi^{+} = \frac{\bar{F}(X)}{F(X)} = \frac{e^{\theta}-e^{\theta x}}{e^{\theta x}-1}$.
   (b) The odds ratio of failure by time, $\phi^{-} = \frac{F(X)}{\bar{F}(X)} = \frac{e^{\theta x}-1}{e^{\theta}-e^{\theta x}}$.

The density function, distribution function, and the hazard function for different values of θ are plotted in Figures 1 to 3 respectively. From the density plots, it is clear that the positive value of the parameter θ confirms the left-skewed behavior and a negative value indicates the right-skewed behavior. So it is a distribution on $[0, 1]$, which can be used for modeling left or right skewed data sets. When the value of θ is positive and increases the density function becomes more peaked but less left-skewed. But the behavior is just the opposite when θ is negative. Even though from Figure 3 it is clear that the distribution has IFR for different values of θ, the behavior of the hazard function doesn't vary much. The nature of this distribution is actually very similar to the power function distribution. It means a comparison with power

function distribution will be quite interesting. For illustrating this, the density plots of UTPD and power function distribution are drawn together in Figure 4. As $\theta > 0$ and increases UTPD coincides with power function distribution. A comparison with beta distribution is also interesting since beta distribution is a flexible distribution with wide applications. But we know that the failure rate function of the beta (p, q) distribution is increasing only if $p \geq 1$, and the comparison will be meaningful only under this particular case. So we have not given much importance to this part in this study.

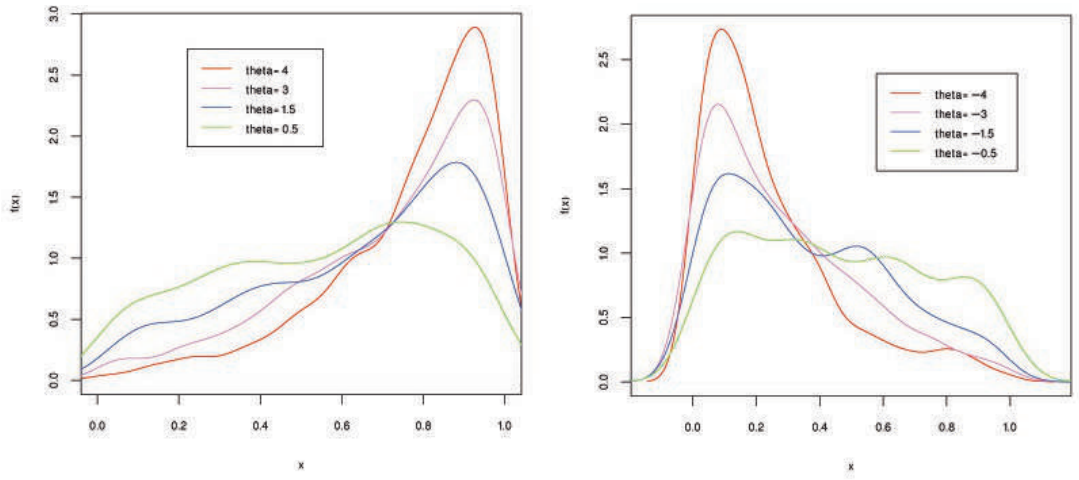


Figure 1: Density plots of UTPD for various values of θ

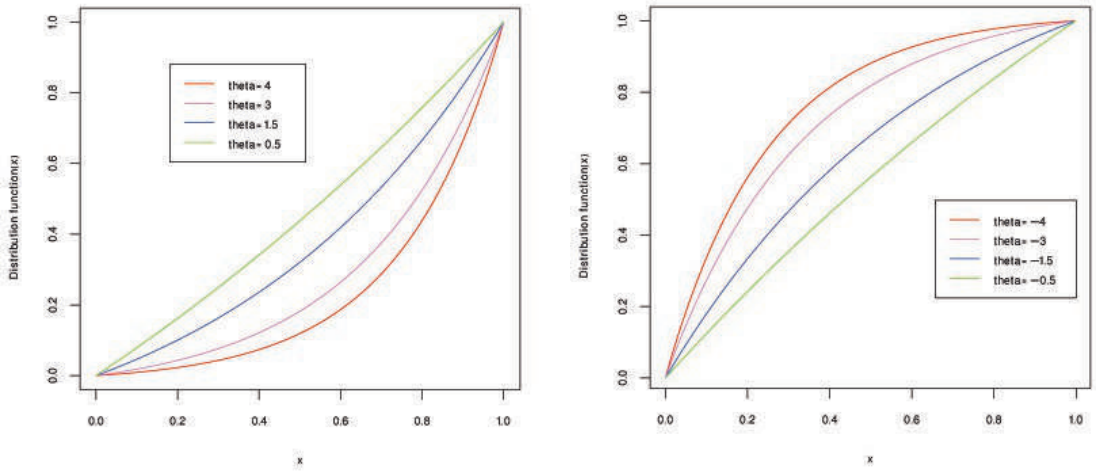


Figure 2: Distribution function of UTPD for various values of θ

10. Skewness and Kurtosis

Using the quantile function given in property 7, the first, second and third quantiles are $x_{0.25}$, $x_{0.50}$, $x_{0.75}$ respectively.

Bowley's measure of skewness,

$$\begin{aligned} S &= \frac{x_{0.75} + x_{0.25} - 2x_{0.50}}{x_{0.75} - x_{0.25}} \\ &= \frac{log[\frac{\frac{1}{16}((3e^{\theta}+1)(e^{\theta}+3)}{(\frac{1}{2}(e^{\theta}+1))^2}]}{log[\frac{3e^{\theta}+1}{e^{\theta}+3}]}. \end{aligned}$$

The kurtosis is measured by the method introduced by Moors (1988). He derived this measure using octiles, where the octiles E_i are defined as,

$$P(X < E_i) \leq \frac{i}{8}$$

and

$$P(X > E_i) \leq 1 - \frac{i}{8}.$$

Using octiles the measure of kurtosis,

$$K = \frac{(E_7 - E_5) + (E_3 - E_1)}{E_6 - E_2}.$$

These two measures are given in Table 1 and the observations we made from the density plots regarding skewness and kurtosis are very well established numerically in this table. It is clear that the distribution is symmetric for θ, and the values of the kurtosis are the same for both the negative and positive values of the parameter.

Table 1: Skewness and Kurtosis

Parameter:θ	Skewness	Kurtosis
0.5	-0.0613	1.0114
2	-0.1953	1.1385
5	-0.2579	1.2925
8	-0.2616	1.3055
-0.5	0.0613	1.0114
-2	0.1953	1.1385
-5	0.2579	1.2925
-8	0.2616	1.3055

Remark 1: This distribution is useful in machine learning specifically for the normalization used for the data representation, further processing and accuracy. The usual transformation used for this purpose is

$$\frac{x_i - min(x_i)}{max(x_i) - min(x_i)}. \tag{4}$$

Later in our real data analysis part described in the last section of this paper, we explain the use of this distribution in such transformations.

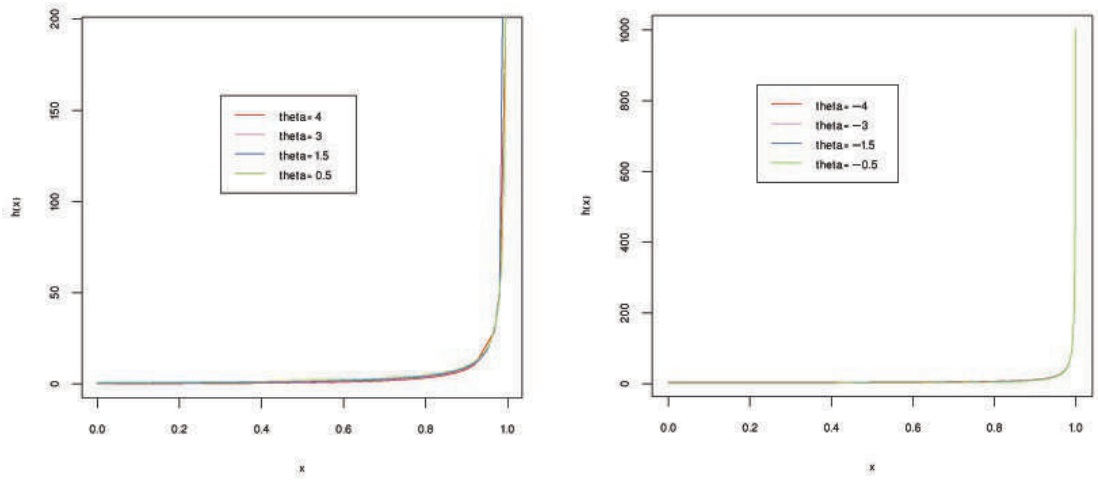


Figure 3: Hazard function of UTPD for various values of θ

2.2. Distribution of order statistics

Assume that $X_1, X_2, X_3, ..., X_n$ are independent random variables following UTPD with parameter θ. The pdf of $\min(X_1, X_2, X_3, ..., X_n)$ is given by

$$f_{X_{(1)}}(x) = \frac{n\theta e^{\theta x}}{(e^{\theta}-1)^n}\left(e^{\theta} - e^{\theta x}\right)^{n-1}$$

and the pdf of $\max(X_1, X_2, X_3..., X_n)$ is given by

$$f_{X_{(n)}}(x) = \frac{n\theta e^{\theta x}}{(e^{\theta}-1)^n}\left(e^{\theta x} - 1\right)^{n-1}.$$

In the next section, we describe some transformed distributions, which seems very similar to some familiar distributions but with different domains.

3. Transformed Distributions

Here we consider some random variables generated through the transformations of (2) and derive their distributions.

Result 1: Considering the transformation

$$U = -\log X,$$

where X follows UTPD with density function given in (2), the pdf of U is

$$g(u) = \frac{\theta}{e^{\theta}-1}e^{-u}e^{\theta e^{-u}}, \quad 0 \le u < \infty, \tag{5}$$

which is the Weibull-Poisson distribution by Morais and Barreto-Souza (2011).

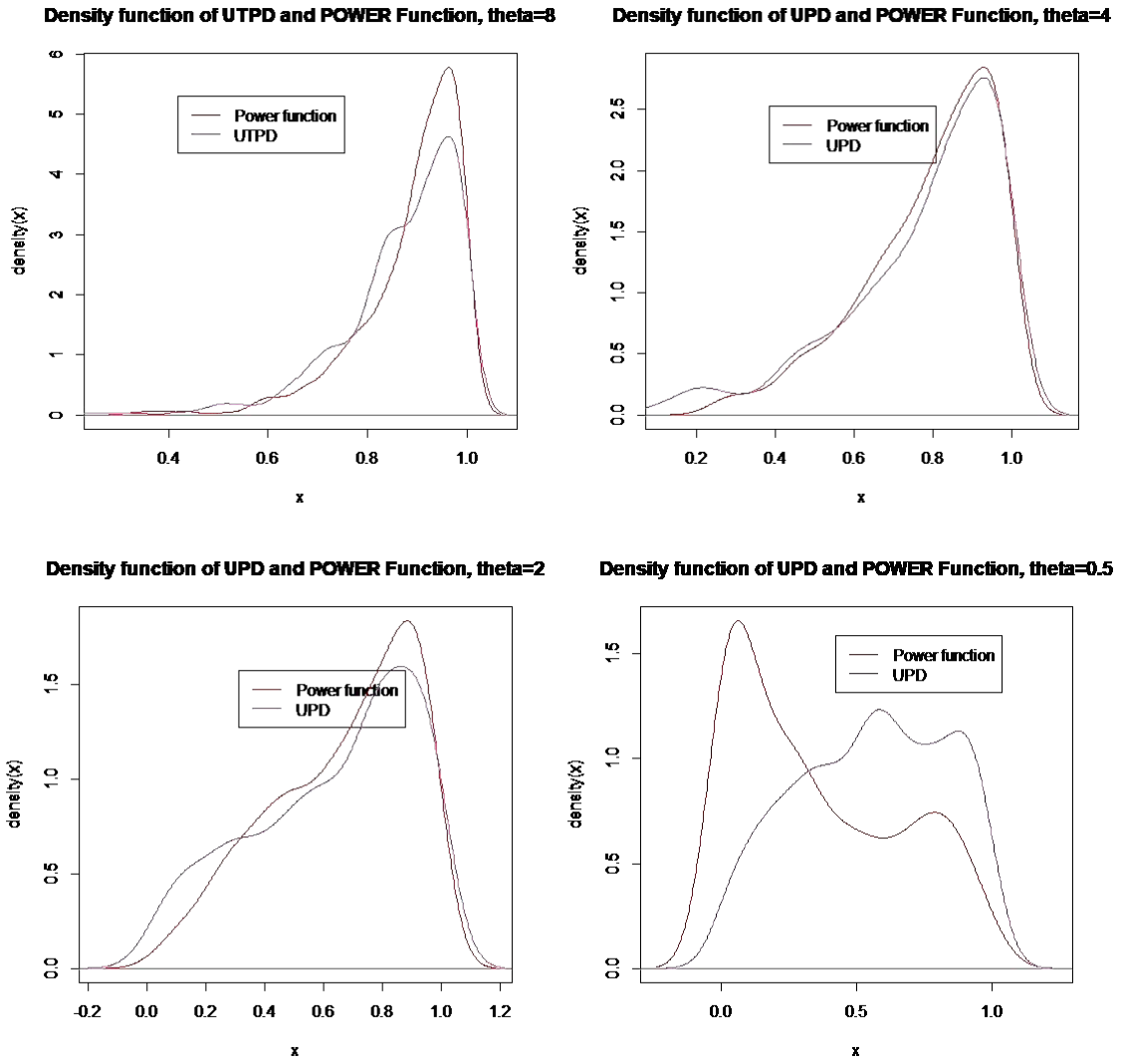


Figure 4: Comparison of UTPD and power function distribution

Result 2: When we take a power transformation

$$V = X^{\frac{1}{\beta}} \tag{6}$$

the density function of V becomes

$$g(v) = \frac{\theta\beta}{e^{\theta} - 1} v^{\beta-1} e^{\theta v^{\beta}}, \quad 0 \leq v \leq 1, \tag{7}$$

which has the form of the Weibull distribution, but domain is quite different.

Result 3: The probability density function of $W = \frac{1}{X}$, where X follows UTPD is

$$h(w) = \frac{\theta}{e^{\theta} - 1} e^{\frac{\theta}{w}} \frac{1}{w^2}, \quad 1 \leq w < \infty. \tag{8}$$

Estimation of the parameter of the UTPD is done in the next section.

4. Estimation of the Parameter

For the estimation of the parameter, we employ the maximum likelihood (ML) method and the method of moments (MM), and comparisons are made with numerical illustrations.

4.1. Maximum likelihood estimation (MLE)

Suppose a sample of size n is taken from UTPD with pdf (2). By taking logarithm of the likelihood function and finding the derivative with respect to θ, we have a nonlinear equation

$$\frac{\partial \log L}{\partial \theta} = 0 \Rightarrow \frac{n}{\theta} - \frac{ne^{\theta}}{e^{\theta}-1} + \sum_{i=1}^{n} x_i = 0,$$

which can be solved numerically to estimate the parameter.

4.2. Method of moments

Another method used for the estimation of the parameter is the method of moments. Equating the first raw moment to the corresponding sample moment, the following equation is obtained, and solving the same for θ results in the estimate.

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{e^{\theta}(\theta-1)+1}{\theta(e^{\theta}-1)}.$$

4.3. Large sample properties

The asymptotic properties of the ML estimators, assuming the usual regularity conditions are provided in this section.

Property 1: The ML estimator $\hat{\theta}$ is asymptotically normally distributed with mean θ and variance $\frac{1}{nI(\theta)}$ where $I(\theta)$ is the well known information matrix.

Proof: We have the log likelihood function

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

Then

$$\frac{\partial^2 logL}{\partial\theta^2} = \frac{-n}{\theta^2} + \frac{ne^{\theta}}{(e^{\theta}-1)^2}.$$

If we denote the gradient of $\log L$, the score statistic as $S(\theta)$, and $-\frac{\partial^2 logL}{\partial\theta^2}$ as $K(\theta)$, then the above equation can be written as,

$$K(\theta) = -S'(\theta) = \frac{n}{\theta^2} - \frac{ne^{\theta}}{(e^{\theta}-1)^2}.$$

Also we know that,

$$S(\theta) = \frac{\partial logL}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial logf(X_i,\theta)}{\partial\theta} \quad and$$

$$K(\theta) = \sum_{i=1}^{n} K(X_i, \theta).$$

Then, $E(K(\theta)) = nI(\theta)$ where $I(\theta) = E[\frac{\partial log f(X_i,\theta)}{\partial \theta}]^2$, the information matrix.

Using Taylor's formula,

$0 = S(\hat{\theta}) = S(\theta) - K(\theta)(\hat{\theta} - \theta) + R$, where R tends to zero.

And finally after adjusting the terms, $\sqrt{n}(\hat{\theta} - \theta) = \frac{S(\theta)/\sqrt{n}}{K(\theta)/n}$.

By Slutsky's theorem $\hat{\theta}$ converges in distribution to $N(\theta, \frac{1}{nI(\theta)})$. □

Now the consistency property of $\hat{\theta}$ is stated below, the proof of which readily follows as in Kale (2007).

Property 2: The likelihood equation admits a consistent solution and the consistent estimator is essentially unique.

4.4. Numerical examples

Simulated samples of sizes 20, 60 and 100 from the population following UTPD for selecting the better method of estimation. For the comparison purpose of the two methods discussed above, each sample is generated 1000 times. The estimate of θ, standard error (SE), mean square error (MSE), 95% confidence intervals (CI) for the parameters and the coverage probabilities (CP) are shown in Table 2. All the simulation works and other computations are done using R-programming and the R codes are presented in the Annexure. The SE and MSE are decreasing with an increase in sample size. The coverage probabilities are increasing when the sample size is increasing. But for smaller sample sizes, the coverage probabilities of the parameters estimated using the ML method are lesser than that generated by MM. From the table, it is clear that both the ML method and MM are equally good for estimation purposes based on the MSE. Both the methods give us approximately equal values as parameter estimates.

In the next section, an attempt is made to generalize the UTPD into a general finite interval (a,b).

5. Generalized UTPD

In this section, we construct a generalization of UTPD. As we have seen in the definition of UTPD, the domain is $[0, 1]$. This can be generalized to a distribution defined on a finite interval (a, b). Let X be a continuous random variable defined on (a, b). The probability density function of X is given by

$$f(x) = \frac{\theta}{e^{\theta b} - e^{\theta a}} e^{\theta x}, \quad a < x < b, \theta \neq 0.$$

When $\theta = 0$, it becomes the uniform distribution defined on (a,b).

Properties

1. The k^{th} raw moment is given by

$$E(X^k) = C\{\frac{b^k e^{\theta b} - a^k e^{\theta a}}{\theta} - \frac{k}{\theta^2}(b^{k-1}e^{\theta b} - a^{k-1}e^{\theta a}) + \frac{(k-1)k}{\theta^3}(b^{k-2}e^{\theta b} - a^{k-2}e^{\theta a}) + ... + \frac{(-1)^k(1.2.3...k)}{\theta^{k+1}}(e^{\theta b} - e^{\theta a})\},$$

   where $C = \frac{\theta}{(e^{\theta b} - e^{\theta a})}$.

2. Mean= $\frac{be^{\theta b} - ae^{\theta a}}{\left(e^{\theta b} - e^{\theta a}\right)} - \frac{1}{\theta}$.

3. Variance= $\frac{\theta}{\left(e^{\theta b} - e^{\theta a}\right)}\left\{\frac{b^2e^{\theta b} - a^2e^{\theta a}}{\theta} - \frac{2}{\theta^2}\left(be^{\theta b} - ae^{\theta a}\right) + \frac{2}{\theta^3}\left(e^{\theta b} - e^{\theta a}\right)\right\} - \left(\frac{be^{\theta b} - ae^{\theta a}}{\left(e^{\theta b} - e^{\theta a}\right)} - \frac{1}{\theta}\right)^2$.

4. The hazard rate function, $h(x) = \frac{\theta e^{\theta x}}{e^{\theta b} - \left\{e^{\theta x}\right\}}$.

5. The mean residual life function (MRL), $\mu(t) = \frac{1}{e^{\theta b} - e^{\theta a} - e^{\theta t} - e^{\theta a}}\left\{e^{\theta b}(b - t - 1) + e^{tb}\right\}$.

6. The quantile function is $x = \frac{1}{\theta}\ln\left[e^{\theta a}(1-u) + ue^{\theta b}\right]$.

More interesting features are the topics for further studies.

6. Applications

In this section, the application of the distribution is illustrated by fitting the UTPD to four data sets. As mentioned in Section 2, we are comparing the UTPD and power function distributions for all these data sets.

Data Set 1: This data set is obtained from the Los Angeles Department of Water and Power (LADWP) solar incentive program, which offers incentives to offset the cost of installing a solar rooftop system in the homes/business of the people of Los Angeles. This metric measures the Net Energy Metering (NEM) installed capacity (Kilowatts), which is available in https://catalog.data.gov/dataset. The data consists of the observations from 2016 to 2018, which describes the application of UTPD in time series as well as physics. As mentioned in Remark 1, the data can be transformed using (4) to bring the data into the range [0,1]. Now we try to fit the power function and UTPD to this transformed data. By the Kolmogorov-Smirnov (K-S) distance measure and p-value given in Table 3, it is clear that both these distributions are good fit for this data set. The p value is greater than 0.05, confirming that the UTPD and power function distribution are good approximations. But we may conclude that UTPD is a better fit for this data since K-S distance is lesser but p-value is greater when compared with power function distribution.

Data Set 2: Data set 2 is the total tax and non-tax revenue of Egypt from 2002 to 2018 available in https://stats.oecd.org and these are time series observations from financial sector. Transform the data using (4) and here also we obtain the power function and UTPD as

suitable fit for this data based on K-S distance and p-value given in Table 3. In the light of these two values we could identify that UTPD is a better fit for this data than power function distribution.

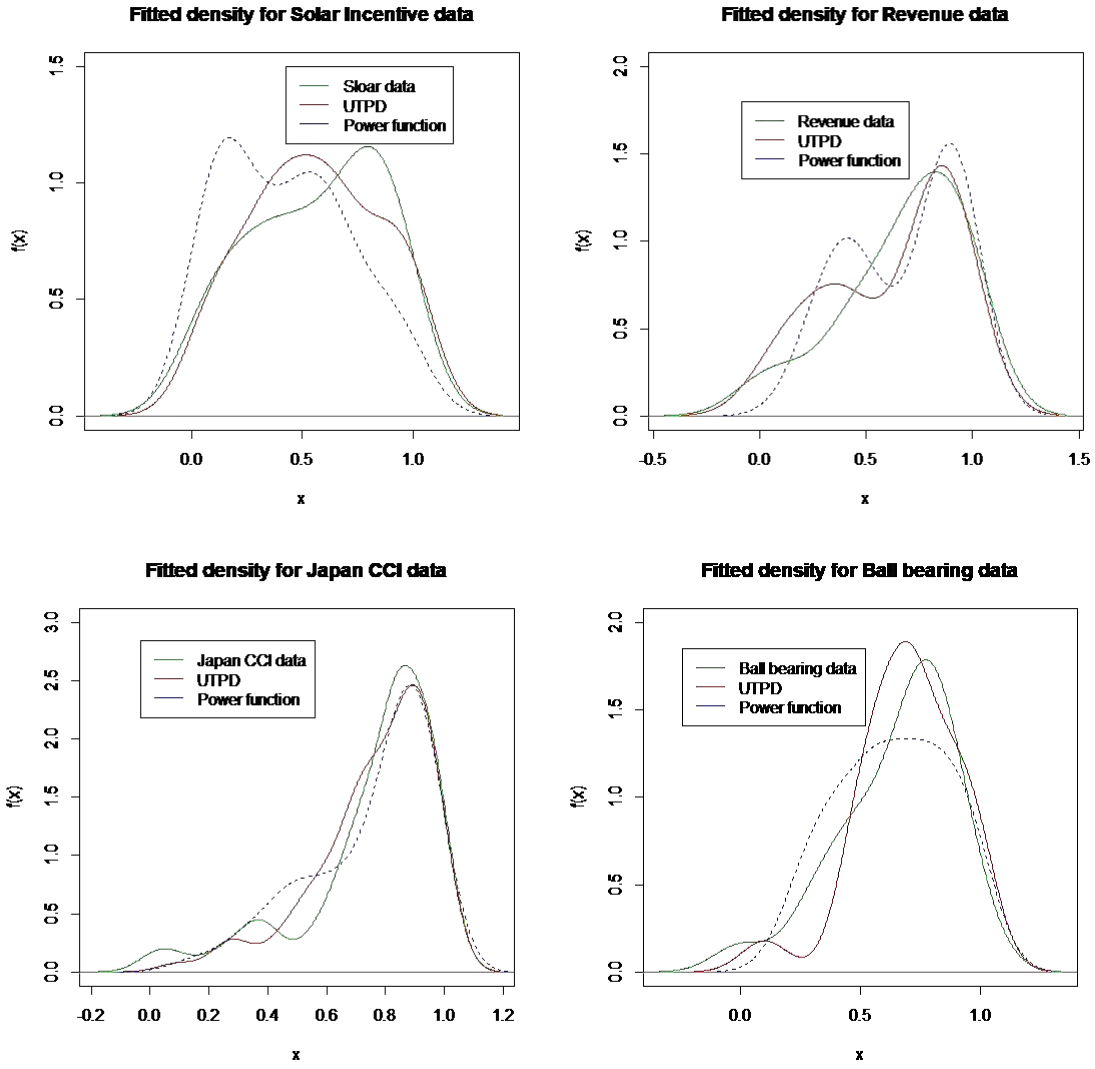


Figure 5: Density plots of fitted data sets

Data Set 3: Now we consider another time series data for describing the applications of UTPD. This is a set of observations of Japan consumer confidence index from January 2014 to March 2021. (Ref: https://stats.oecd.org). Again after suitable transformation we get the power function and UTPD as suitable fit for this transformed data from the values in Table 3. But since the p-value is higher and K-S distance is lesser for UTPD than power function, it is clear that UTPD is a better fit.

Data Set 4: The fourth data set we consider is the ball bearing data taken from Lawless (2003) to employ it in the engineering field. The data are the number of million revolutions before failure for each of the 23 ball bearings in the life test and they are 17.88, 28.92, 33.00,

41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, and 173.40. As mentioned in Remark 1, the data can be transformed using (4) to $[0, 1]$. The K-S distance and p-value given in Table 3, reveal that UTPD is a better fit for this data than power function distribution.

The densities of the original data sets together with the fitted densities plotted in Figure 5 reveal that UTPD is a good fit for all the data sets considered.

Table 2: Parameter Estimates

Sample size (n)	Parameter (θ)	Method	Estimate ($\hat{\theta}$)	SE	MSE	CI	CP
20	4	MLE	4.12	1.110	1.307	(3.15,5.09)	0.627
		MM	4.10	0.252	1.288	(1.88,6.37)	0.957
60		MLE	4.07	0.625	0.385	(3.12,5.02)	0.880
		MM	4.03	0.080	0.394	(2.80,5.26)	0.948
100		MLE	4.01	0.482	0.234	(3.07,4.96)	0.955
		MM	4.02	0.048	0.246	(3.06,4.98)	0.954
20	3	MLE	3.06	0.968	1.020	(2.21,3.91)	0.620
		MM	3.09	0.233	1.098	(1.05,5.14)	0.938
60		MLE	3.06	0.553	0.303	(2.22,3.90)	0.876
		MM	3.04	0.072	0.312	(1.95,4.13)	0.950
100		MLE	3.01	0.424	0.183	(2.18,3.84)	0.950
		MM	3.01	0.043	0.190	(2.15,3.86)	0.947
20	1.5	MLE	1.56	0.835	0.740	(0.83,2.29)	0.622
		MM	1.54	0.196	0.770	(-0.17,3.26)	0.950
60		MLE	1.50	0.475	0.231	(0.78,2.22)	0.861
		MM	1.55	0.063	0.147	(0.79,2.23)	0.942
100		MLE	1.50	0.367	0.147	(0.78,2.22)	0.942
		MM	1.49	0.036	0.130	(0.78,2.20)	0.947
20	0.5	MLE	0.52	0.792	0.653	(-0.16,1.22)	0.610
		MM	0.51	0.180	0.653	(-1.07,2.09)	0.950
60		MLE	0.50	0.452	0.209	(-0.18,1.19)	0.868
		MM	0.50	0.058	0.208	(-0.38,1.40)	0.957
100		MLE	0.49	0.349	0.125	(-0.20,1.16)	0.949
		MM	0.48	0.034	0.119	(-0.18,1.16)	0.950
20	-2	MLE	-2.06	0.870	0.780	(-2.83,-1.31)	0.631
		MM	-2.10	0.203	0.840	(-3.89,-0.31)	0.954
60		MLE	-2.04	0.495	0.263	(-2.79,-1.28)	0.865
		MM	-2.04	0.062	0.234	(-2.98,-1.09)	0.948
100		MLE	-2.01	0.382	0.148	(-2.78,-1.26)	0.941
		MM	-2.03	0.036	0.132	(-2.74,-1.33)	0.952

Table 3: Fitting of real data sets

	Distribution	Parameter	K-S distance	p-value
Dataset 1	UTPD	0.75	0.0967	0.9991
	Power function	1.28	0.1290	0.9634
Dataset 2	UTPD	2.29	0.1176	0.9999
	Power function	2.08	0.1764	0.7631
Dataset 3	UTPD	3.64	0.0919	0.8585
	Power function	3.04	0.1149	0.6164
Dataset 4	UTPD	1.91	0.1421	0.7657
	Power function	1.86	0.2173	0.6487

7. Conclusion

In this paper, we have studied in detail the uniform truncated Poisson distribution as the solution of a first order differential equation and derived the same from the truncated uniform distribution. Comparisons with some well known distributions are done. The expressions for moments, distributions of the order statistics, *etc.* are further derived. Some transformed distributions are also studied. Some of the estimation procedures of the parameter are discussed. The newly constructed distribution is applied on real data. Characterizations and further applications of UTPD in time series, regression and reliability are the topics for further studies.

Acknowledgements

The authors sincerely thank the anonymous referees and editor for their valuable comments and suggestions.

References

Adamidis, K. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics and Probability Letters*, **39**, 35–42.

Akdogan, Y., Kus, C., Asgharzadeh, A., Kinaci, I. and Sharafi, F. (2016). Uniform-geometric distribution. *Journal of Statistical Computation and Simulation*, **86(9)**, 1754–1770.

Altawil, J. A. (2019). $[0, 1]$ Truncated Lomax – uniform distribution with properties. *Journal of Interdisciplinary Mathematics*, **22(8)**, 1415–1431.

Bayer, F. M., Cintra, R. J. and Cribari-Neto, F. (2018). Beta seasonal autoregressive moving average models. *Journal of Statistical Computation and Simulation*, **88(15)**, 2961–2981.

Cancho, V. G., Louzada-Neto, F. and Barriga, G. D. C. (2011). The Poisson-exponential lifetime distribution. *Computational Statistics and Data analysis*, **55**, 677–686.

Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics and Data Analysis*, **53**, 4433–4440.

Hao, J. and Godbole, A. (2014). Distribution of the maximum and minimum of a random number of bounded random variables. *Open Journal of Statistics*, **06(02)**.

Hassan, A. S., Sabry, M. A. H. and Elsehetry, A. M. (2020). A new family of upper-truncated distributions: properties and estimation. *Thailand Statistician*, **18(2)**, 196–214.

Hu, Y., Peng, X., Li, T. and Guo, H. (2007). On the Poisson approximation to photon distribution for faint lasers. *Physics Letters A*, **367**, 173–176.

Jara, A., Nieto-Barajas, L. E. and Quintana, F. (2013). A time series model for responses on the unit interval. *Bayesian Analysis*, **8(3)**, 723–740.

Kale, B. K. (2007). *A first course on parametric inference*. Narosa, New Delhi.

Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0,1); percentages, proportions and fractions. *Statistical Modelling*, **3**, 193–213.

Kus, C. (2007). A new lifetime distribution. *Computational Statistics and Data Analysis*, **51**, 4497–4509.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley and Sons, New York, USA, 2nd edition.

Louzada, F., Roman, M. and Cancho, V. G. (2011). The complementary exponential geometric distribution: model, properties, and a comparison with its counterpart. *Computational Statistics and Data analysis*, **55(8)**, 2516–2524.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *The Statistician*, **37**, 25–32.

Morais, A. L. and Barreto-Souza, W. (2011). A compound class of Weibull and power series distributions. *Computational Statistics and Data Analysis*, Volume **55(3)**, 1410–1425.

Quijano Xacur, O. (2019). The unifed distribution. *Journal of Statistical Distributions and Applications*, **6(13)**.

Ristic, M. M. and Popovic, B. (2000). A new uniform AR(1) time series model (NUAR(1)). *Publications de l'Institut Mathématique*, **68(82)**, 145–152.

Rocha, A. V. and Cribari-Neto, F. (2009). Beta autoregressive moving average models. *TEST*, **18**, 529.

Tahmasbi, R. and Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Computational Statistics and Data Analysis*, **52**, 3889–3901.

Weigend, A. S. and Gershenfeld, N. A. (1993). *Time series prediction: Forecasting the future and understanding the past*. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XV. Reading, MA:Addison-Wesley.

Yu, L., Wang, S. and Lai, K. K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, **18**, 217–230.

ANNEXURE

A1: R Code for simulation studies using MLE method

```
m=1000
n=20
para=4
x<-list(mode = "vector", length = m)
z<-list(mode = "vector", length = m)
z<-list(mode = "vector", length = m)
out<-list(mode = "vector", length = m)
est<-list(mode = "vector", length = m)
dut<-function(x,a=4)((a/(exp(a)-1))*exp(a*x))
put<-function(x,a=4)(1-(1/(exp(a)-1)*(exp(a)-exp(a*x))))
qut<-function(u,a=4)((1/a)*log(1+u*(exp(a)-1)))
rut<-function(n,a=4)qut(runif(n),a)
for(i in 1:m)
{
x[[i]]<-runif(n)
z[[i]]<-qut(x[[i]])
fn <- function(theta)
sum(log((exp(theta)-1)/theta)-theta*z[[i]])
out[[i]]<-nlm(fn, theta<-0.1, hessian=TRUE)
out
}
w<-vector(mode = "numeric", length = m)
setheta=vector(mode="numeric", length = m) #standard error of theta
for(j in 1:m)
{
w[j]=out[[j]]$estimate
setheta[j]=sqrt(diag(solve(out[[j]]$hessian)))
}
#Confidence intervals
lcla<-vector(mode = "numeric", length = m)
ucla<-vector(mode = "numeric", length = m)
for( i in 1:m)
{
lcla[i] <- (w[i] - 1.96*setheta[i])
ucla[i] <- (w[i] + 1.96*setheta[i])
}
#Coverage Probability
covera<-vector(mode = "numeric", length = m)
for(i in 1:m)
{
covera[i] <- (lcla[i]<=4)&(ucla[i]>=4)
}
```

A2: R Code for simulation studies using the method of moments

```
for(i in 1:m)
{
x[[i]]<-runif(n)
z[[i]]<-qut(x[[i]])
z[[i]]<-na.omit(z[[i]])
func=function(theta){
(sum(z[[i]])/n)- ((exp(theta)*(theta-1)+1)/(theta*(exp(theta)-1)))
}
 Result[[i]]=BFfzero_2(func,-10,100)
}
# w1<-vector(mode = "numeric", length = m)
sew1<-vector(mode = "numeric", length = m)
for(j in 1:m)
{
w1[j]=Result[[j]]$root
}
#Confidence intervals
lclaw1<-vector(mode = "numeric", length = m)
uclaw1<-vector(mode = "numeric", length = m)
for( i in 1:m)
{
lclaw1[i] <- (w1[i] - 1.96*sqrt(var(w1)))
uclaw1[i] <- (w1[i] + 1.96*sqrt(var(w1)))
}
#Coverage Probability
coveragew1<-vector(mode = "numeric", length = m)
for(i in 1:m)
{
coveragew1[i]<-(lclaw1[i]<=4)&(uclaw1[i]>=4)
}
```

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 1, 2022 (New series), pp 297-309

Fitting Model for Self-Similar Traffic - Time Dependent Markovian Process and Second Order Statistics

Abhilash Vollala and Malla Reddy Perati
Department of Mathematics, Kakatiya University, Warangal, India.

Received: 18 March 2021; Revised: 28 April 2021; Accepted: 20 May 2021

Abstract

Various self-similar traffic models proposed earlier are asymptotic. In this connection, many Markovian models are proposed, but the performance analysis was possible only when the system is in steady state, and such traffic models are not realistic. In this paper, a procedure is proposed to fit Markov Modulated Poisson Process (MMPP) with time dependent sinusoidal arrival rates that emulates self-similar traffic. This is done by matching the variance of the both during prescribed time scales. Numerical results represent what extent MMPP could reproduce self-similar traffic in specified time scales.

Key words: Self-similar traffic; Variance-time; Markovain Modulated Poisson process (MMPP); Interrupted Poisson process (IPP).

1. Introduction

The pivotal studies at AT&T Bell labs (Leland, *et al.* 1994; Paxson and Flyod, 1995) revealed that IP packet traffic over LAN and WWW internet traffic (Crovella and Bestavros, 1997) are self-similar (fractal like behavior), and this behavior effects efficiency of network nodes such as routers or switches (Misra, *et al.* 2012). Many traffic models such as Chaotic maps, FARIMA, and FBM are proposed to emulate the self-similar behavior. These models are parsimonious, but are asymptotic, hence they are not practically useful (Erramilli, *et al.* 1996; Norros, 1994). Andersen and Nielsen (1998) used Markovian Arrival Process (MAP) in particular Switched Poisson Process (SPP) to model the Long Range Dependent (LRD) characteristics (statistical definition of self-similarity) of traffic over different time scales, and proposed a fitting procedure wherein covariance of second-order self-similar process, and that of resultant MAP (superposition of several SPPs) are equated. Kasahara, *et al.* (2001) proposed a method based on variance of second-order self-similar traffic using Interrupted Poisson Process (IPP). Later, Reddy, *et al.* (2005) extended the work by making modulating parameters of each IPP unequal, and investigated the relation between traffic parameters, time scale, and parameters of fitting. In all the above, resultant MAP is homogeneous and the arrival rates were not functions of time, and queueing behavior of traffic nodes was investigated in the steady state. Steady state outcomes never give actual queueing behavior, since it relies on a prolonged performance of system to nullify the initial conditions (Kelton and Law, 1985). In real time, network traffic is not homogeneous over all time scales, and the modeled system by no means attains steady state. Because of these reasons, necessity of time dependent (transient) analysis is warranted. Abate and Whitt (1988) analyzed *M/M/1* queue in transient with Poisson arrivals using Laplace transforms. Eick, *et al.* (1993) studied the $M_t/G/\infty$ queue with time

Corresponding Author:Abhilash Vollala
Email : ashwesa.abhi@gmail.com

dependent arrival rates, and determined the number of busy servers at time *t*. Different methods were developed by Jennings and Massey (1997) for analysis blocking in circuit switched networks with transient arrival traffic. Massey (2002) proposed canonical queueing models with time-varying rates, and derived necessary mathematical tools for analysis. Qian and Tipper (2004) proposed a framework for adaptively determining optimal channel allocation scheme, and evaluated performance of the scheme under time varying loads. Liu and Whitt (2014) developed an algorithm to find number of servers required to preserve the prerformance in a multi server queue with time varying arrival rates for extension of feed-forward netwoks. Pant and Ghimire (2016) determined the expected queue length and waiting time of the customers at time t using transient arrival rates. In the papers cited above, authors worked with various queueing systems using time based sinusoidal arrival rates to address issues in various domains. These models are based on the fact that many real time arrival phenomenons are almost periodic in nature. In this paper, procedure to fit Markovian Arrival Process (MAP) with sinusoidal arrival rates that emulates self-similar traffic over prescribed time scales is proposed. Variance of number of arrivals due to self-similar traffic, and that of resultant MAP (superposition of IPPs with sinusoidal arrival rates) are equated at certain time points in order to compute the MAP parameters.

The remaining part is arranged in the following way. In section 2, outline of second-order self-similar process and sinusoidal IPP are presented. In section 3, fitting procedure with time dependent sinusoidal arrival rates is given. Numerical results are demonstrated in section 4. Finally, some conclusions are given in section 5.

2. Second-Order Self-Similar and Sinusoidal IPP Processes

Self-similarity is a property, wherein a certain feature of the object is maintained with respect to scaling in space and time. It is statistically defined as follows. Let X be a second order process with variance σ^2, and the time axis is splitted into disjoint sub intervals of unit length. Let $X = \{X_t / t = 1,2,3,.....\}$ be the points (packet arrivals) in t^{th} interval. Let $X^{(r)} = \{X_t^{(r)}\}$ be a new sequence obtained by averaging the original sequence over non-overlapping blocks of size r . *i.e.*,

$$X_t^{(r)} = \frac{1}{r}\sum_{i=1}^{r} X_{(t-1)r+i} \ , \quad t = 1,2,3,..... \tag{1}$$

The obtained sequence is a second order process, and is called exactly second order self-similar with Hurst Parameter, $H = 1-\beta/2,$ if

$$Var\,(X^{(r)}) = \sigma^2 r^{-\beta} \ , \ \forall r \geq 1 \tag{2}$$

This feature can be emulated by Markov Modulated Poisson Process (MMPP) over the desired time scale (Reddy, *et al.* 2005). For the reason mentioned in the introduction, a special type of two-state IPP is proposed using time dependent arrival rates. The proposed IPP is given as follows:

$$Q = \begin{bmatrix} -d_1 & d_1 \\ d_2 & -d_2 \end{bmatrix}, \quad R(t) = \begin{bmatrix} a + b\sin(t) & 0 \\ 0 & 0 \end{bmatrix}. \tag{3}$$

In the above, Q is transition rate matrix with two distinct parameters and $R(t)$ is arrival rate matrix, which says that, when Markov process is in state 1, arrival process is with sinusoidal arrival rate $\lambda(t) = a + b\sin(t)$, where a, b are the constants, and $b = a\gamma, 0 < \gamma < 1$ (Eick, *et al.* 1993), and when the Markov process in state 2, there are no arrivals. The number of arrivals in $(0, t]$ of the said IPP is denoted by N_t, and J_t be the state of Markov process at time *t*. The Generating function $P^*(z,t)$ obtained by forward Chapman-Kolmogorov equations with time dependent Markovian process (Fischer and Hellstern, 1993) is given by

$$P^*(z,t) = Exp\left(Qt - (1-z)\int_0^t R(t)dt \right). \tag{4}$$

The mean value of N_t is given by (Heffes and Lucantoni, 1986)

$$E[N_t] = \Pi' \left[\frac{\partial}{\partial z} P^*(z,t) \right]_{z=1} e,$$

Differentiating with respect to *z*, and solving above equation (Neuts, 1979; Coddington and Levinson, 1987), one can obtain as

$$\begin{aligned} & E[N_t] = \Pi' R^{\circ}(t)\, e, \\ \Rightarrow\ & E[N_t] = \frac{d_2 \int_0^t \lambda(t)\, dt}{d_1 + d_2}, \\ \Rightarrow\ & E[N_t] = \frac{d_2\, (at + b(1-\cos(t)))}{d_1 + d_2}. \end{aligned} \tag{5}$$

The variance of N_t is given by (Heffes and Lucantoni, 1986)

$$Var[N_t] = \Pi' \left[\frac{\partial^2}{\partial z^2} P^*(z,t) \right]_{z=1} e + E[N_t] - \{E[N_t]\}^2$$

After differentiating with respect to *z*, and by applying some algebraic manipulation (Neuts, 1979; Coddington and Levinson, 1987), one can obtain as

$$\begin{aligned} Var[N_t] = E[N_t] + \frac{2}{t}[(E[N_t])^2 - \Pi' R^{\circ}(t)(Q + e\Pi')^{-1} R^{\circ}(t)e] \\ + \frac{2}{t^2}[\Pi' R^{\circ}(t)(e^{Qt} - I) R^{\circ}(t)\, e], \end{aligned}$$

$$\Rightarrow Var[N_t] = \frac{d_2 \int_0^t \lambda(t)\, dt}{d_1 + d_2} + \frac{2 d_1 d_2 \left(\int_0^t \lambda(t)\, dt \right)^2}{t\, (d_1 + d_2)^3} - \frac{2 d_1 d_2 \left(\int_0^t \lambda(t)\, dt \right)^2}{t^2\ (d_1 + d_2)^4} [1 - e^{-(d_1 + d_2)t}],$$

$$\Rightarrow Var[N_t] = \frac{d_2\ (at + b(1-\cos(t)))}{d_1 + d_2} + \frac{2\,d_1 d_2\ (at + b(1-\cos(t)))^2}{t\ (d_1 + d_2)^3} - \frac{2\,d_1 d_2\ (at + b(1-\cos(t)))^2}{t^2\ (d_1 + d_2)^4}[1 - e^{-(d_1+d_2)t}]. \tag{6}$$

where Π' is steady state vector of Markov chain, e is the vector of appropriate dimension with each entry as 1, and $R^{\circ}(t) = \int_0^t R(t)dt.$

The Index of Dispersion for Counts (IDC) is

$$IDC(t) = \frac{Var[N_t]}{E[N_t]}$$

By using Eqs. (4) and (5), one can have

$$IDC(t) = 1 + \frac{2\,d_1\left(\int_0^t \lambda(t)\,dt\right)}{t\ (d_1 + d_2)^2} - \frac{2\,d_1\left(\int_0^t \lambda(t)\,dt\right)}{t^2\ (d_1 + d_2)^3}[1 - e^{-(d_1+d_2)t}]. \tag{7}$$

One can derive the following results:

i. $IDC(t) \to 1,$ as $t \to 0$ (i.e., MMPP move towards to a Poisson process).

ii. $IDC(t) \to 1 + \frac{2\,d_1\,a}{(d_1 + d_2)^2}$, a constant, as $t \to \infty$.

3. Fitting Procedure

Variance based fitting procedure is to obtain the MMPP parameters using time dependent arrival rates. From earlier, (Reddy, *et al.* 2005; Kasahara, *et al.* 2001) it is known that modelling of self-similar traffic involves superposition of number of two-state MMPPs (in particular IPPs), and the fundamental requirements for fitting process are

i) $[r_{\min}, r_{\max}]$: Minimum, maximum limits of the time scale range.

ii) $\lambda_w(t)$: Arrival rate of whole process at time t.

iii) n : Number of superposed two-state IPPs.

iv) H : Hurst parameter.

v) σ^2 : Variance.

The j^{th} IPP of the process is given by

$$Q_j = \begin{bmatrix} -d_{1j} & d_{1j} \\ d_{2j} & -d_{2j} \end{bmatrix}, \quad R_j(t) = \begin{bmatrix} a + b_j \sin(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad 1 \le j \le n \tag{8}$$

where a, b_j are the constants and $b_j = a\gamma_j$, $0 < \gamma_j < 1$. In this context, a is assumed that, it is equal to the whole arrival rate. The superposition of n IPPs, and a Poisson process is stated as

$$Q = Q_1 \oplus Q_2 \oplus \cdots \oplus Q_n,$$

$$R(t) = R_1(t) \oplus R_2(t) \oplus R_3(t) \cdots \oplus R_n(t) \oplus \lambda_p(t). \tag{9}$$

Here, time dependent arrival rate of classical Poisson Process is denoted by $\lambda_p(t)$, and $\oplus$ indicates the kronecker's sum, where the resultant of above sum is also an MMPP. The arrival rate of whole process at time t is given by

$$\lambda_w(t) = \lambda_p(t) + \sum_{j=1}^{n} \frac{d_{2j}}{d_{1j} + d_{2j}} \lambda_{1j}(t). \tag{10}$$

Let Y_t be the number of arrivals in whole Markovian process, and the arrivals from j^{th} IPP and Poisson process are denoted by $N_{<j,t>}$, $N_{<p,t>}$ respectively, during t^{th} time slot. The count of arrivals after the averaging process are $Y_t^{(r)}$, $N_{<j,t>}^{(r)}$ and $N_{<p,t>}^{(r)}$ respectively.

Therefore, $Y_t^{(r)} = \sum_{i=1}^{n} N_{<j,t>}^{(r)} + N_{<p,t>}^{(r)}$. (11)

The variance of j^{th} IPP and Poisson process are given as

$$Var\,[N_{<j,t>}^{(r)}] = \frac{d_{2j}\left(\int_o^r \lambda_{1j}(t)\,dt\right)}{r^2(d_{1j}+d_{2j})} + \frac{2d_{1j}d_{2j}\left(\int_o^r \lambda_{1j}(t)\,dt\right)^2}{r^3(d_{1j}+d_{2j})^3} - \frac{2d_{1j}d_{2j}\left(\int_o^r \lambda_{1j}(t)\,dt\right)^2}{r^4(d_{1j}+d_{2j})^4}[1-e^{-r(d_{1j}+d_{2j})}], \tag{12}$$

$$Var\left[N_{<p,t>}^{(r)}\right] = \frac{\int_0^r \lambda_p(t)\,dt}{r^2}. \tag{13}$$

From Eqs. (11), (12), (13), and using the fact that variance of a resultant process is preserved by the superposition of distinct sub-processes, the following relation is obtained.

$$Var\left[Y_t^{(r)}\right] = \frac{\lambda_w(t)}{r} + \sum_{i=1}^{n} \xi_j \left(\int_0^r \lambda_{1j}(t)\,dt\right)^2, \tag{14}$$

where $\xi_j = \frac{2d_{1j}d_{2j}}{r^3(d_{1j}+d_{2j})^3} - \frac{2d_{1j}d_{2j}}{r^4(d_{1j}+d_{2j})^4}[1-e^{-r(d_{1j}+d_{2j})}]$. (15)

Using (1) and (14), at n different points r_j, $j = 1,2,3,..,n$ variance of both processes are equated. The time scale over which self-similarity of traffic exhibits is taken as $[r_{\min}, r_{\max}]$ (*i.e.*, $r_{\min} \le r \le r_{\max}$), then r_j is given by

$$r_j = r_{\min}\alpha^{j-1}, \quad j = 1,2,...,n. \tag{16}$$

where $\alpha = \left[\frac{r_{\max}}{r_{\min}}\right]^{\frac{1}{n-1}}, \quad n > 1.$ (17)

Consider that $r\,d_{1j} = l1,\ r\,d_{2j} = l2$ and make use of inequality given below in Eq. (15),

$$1-(l1+l2) < e^{-(l1+l2)} < 1-(l1+l2)+\frac{(l1+l2)^2}{2!},\ \forall\, l1,\ l2 > 0,\ l1+l2 << 1,$$

One can find that the ξ_j lies in between $(0,\ 1),\ \forall j$. Since ξ_j is bounded $\forall j$, $Var[Y_t^{(r)}]$ is bounded for $t > 0$. By virtue of self-similarity, the following relations can be adopted: $r_j\,d_{1j} = d_1$ (a constant), $r_j\,d_{2j} = d_2$ (a constant), $j = 1,\ 2,\ ...\ ,n,$ where d_1 and d_2 are independent,

$$i.e.,\ d_{1j} = \frac{r_1}{r_i}d_{11}\ ,\ d_{2j} = \frac{r_1}{r_i}d_{21}. \tag{18}$$

The above assumptions are based on the fact that self-similar process is identical for all time scales. Now the parameters to be obtained are d_{11}, d_{21}. Once their values are obtained, then the values of d_{1j} and d_{2j} can be generated using (18). By using these values, $\lambda_p(t)$ can be obtained from (10). Finally, the required transition parameters d_{11}, d_{21} can be determined, such that the value of the integral $\int_{r_{\min}}^{r_{\max}} \left[Var(X^{(r)}) - Var(Y_t^{(r)})\right] dr$ is minimum.

4. Numerical Results

Accuracy of fitting (self similar traffic as of MMPP) is presented using different samples given in Table 1. The samples are pertaining to seminal studies at AT& T Bell labs (Leland, *et al.* 1994). The number of superposed IPPs n is taken to be 4. The sinusoidal arrival rates are given in Table 2. The variance versus time curves of resultant MMPPs and self-similar traffic are shown in Figs.1-12 for $n = 4$ in the time scale ranges $[10, 10^4]$, $[10^2, 10^5]$, $[10^2, 10^6]$, $[10^2, 10^7]$. The n is taken to be 3 in typical time scale range $[10^2, 10^5]$ to represent the effect of the number of superposed components, and the pertinent results are presented in Figs.13-15. The results exhibit good agreement with that of self-similar traffic.

Table 1: Fitting data of samples in time scale range [10^2, 10^5]

Sample Number	Parameter Values	*n*=4		*n*=3	
		d_{11}	d_{21}	d_{11}	d_{21}
Sample 1	$H = 0.7,\ \lambda_w(t) = 1$, and $\sigma^2 = 0.6$	0.22	0.001	0.09	0.0924
Sample 2	$H = 0.8,\ \lambda_w(t) = 1$, and $\sigma^2 = 0.6$	0.5	0.0359	0.385	0.025
Sample 3	$H = 0.9,\ \lambda_w(t) = 1$, and $\sigma^2 = 0.6$	0.040015	0.005	0.202	0.001

Source: Leland, et al. 1994 *(for samples)*

Table 2: Sinusoidal arrival rates

Arrival rate	Value
$\lambda_{11}(t)$	$1+0.1\times\sin(t)$
$\lambda_{12}(t)$	$1+0.3\times\sin(t)$
$\lambda_{13}(t)$	$1+0.5\times\sin(t)$
$\lambda_{14}(t)$	$1+0.7\times\sin(t)$

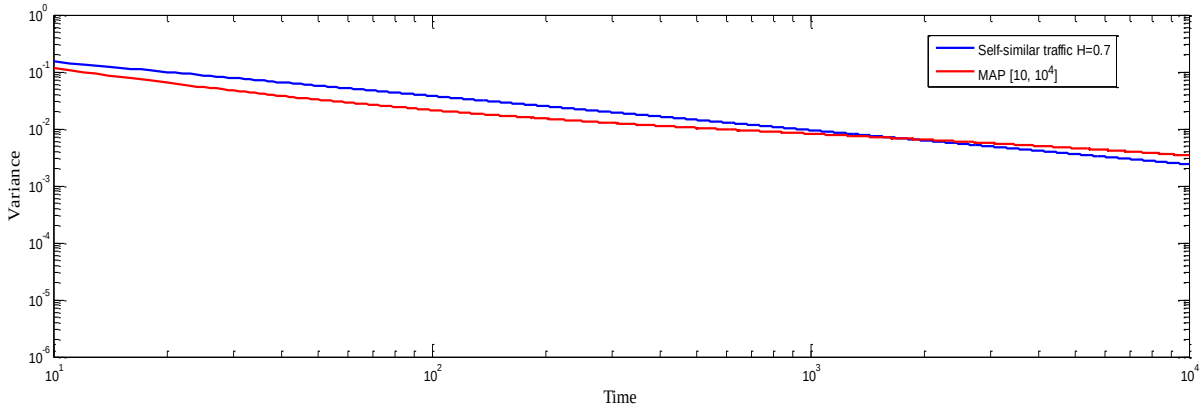


Figure 1: Variance versus time curves with $n = 4$ for sample 1during time scale $[10, 10^4]$

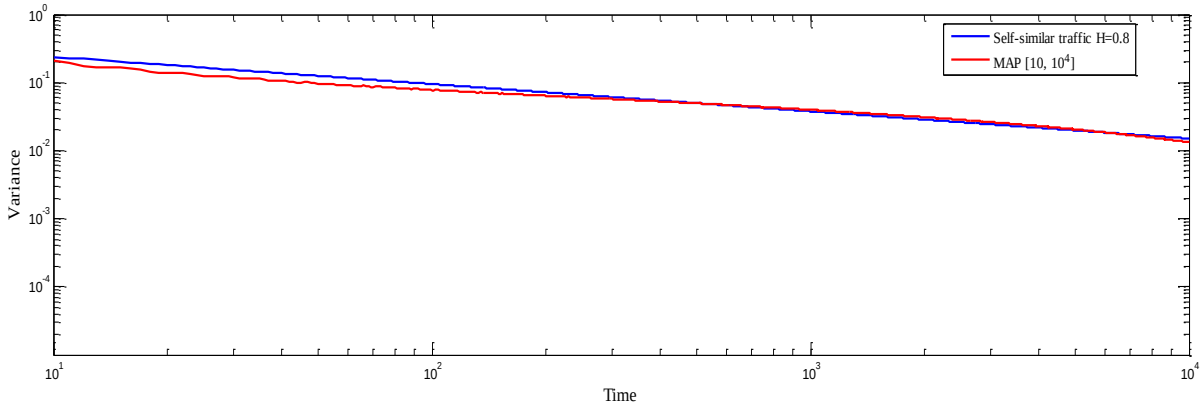


Figure 2: Variance versus time curves with $n = 4$ for sample2 during time scale $[10, 10^4]$

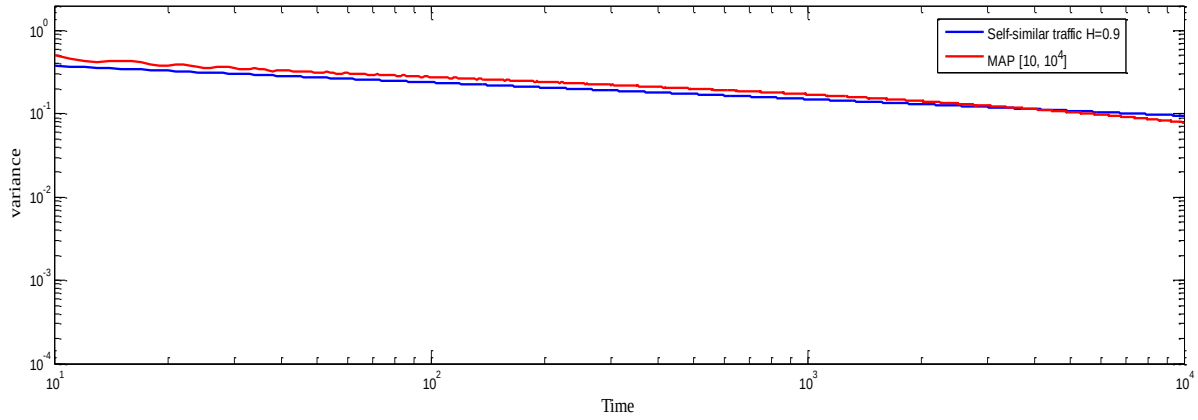


Figure 3: Variance versus time curves with $n = 4$ for sample3 during time scale $[10, 10^4]$

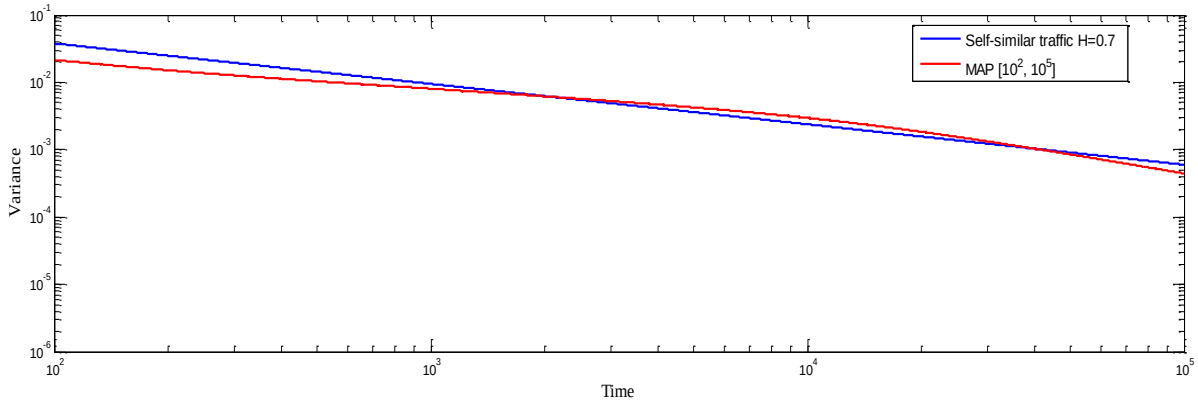


Figure 4: Variance versus time curves with $n = 4$ for sample1 during time scale $[10^2, 10^5]$

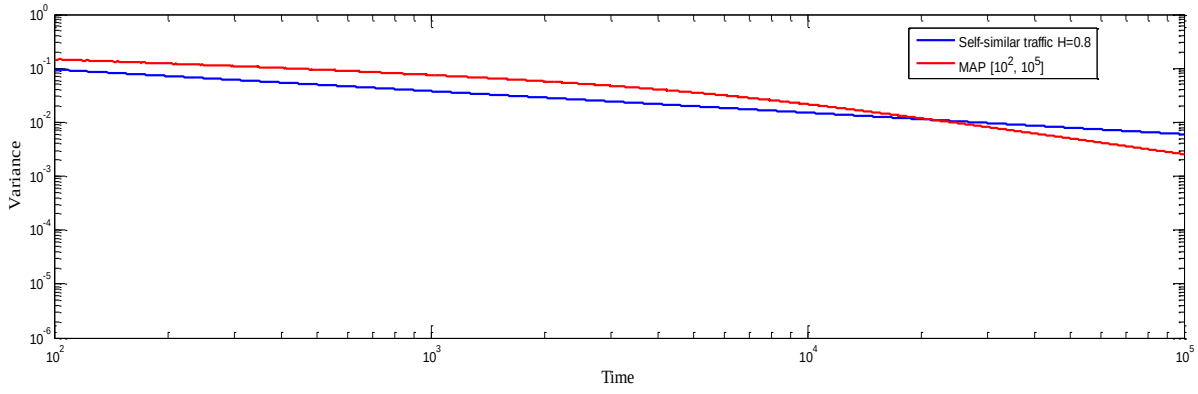


Figure 5: Variance versus time curves with $n = 4$ for sample2 during time scale $[10^2, 10^5]$

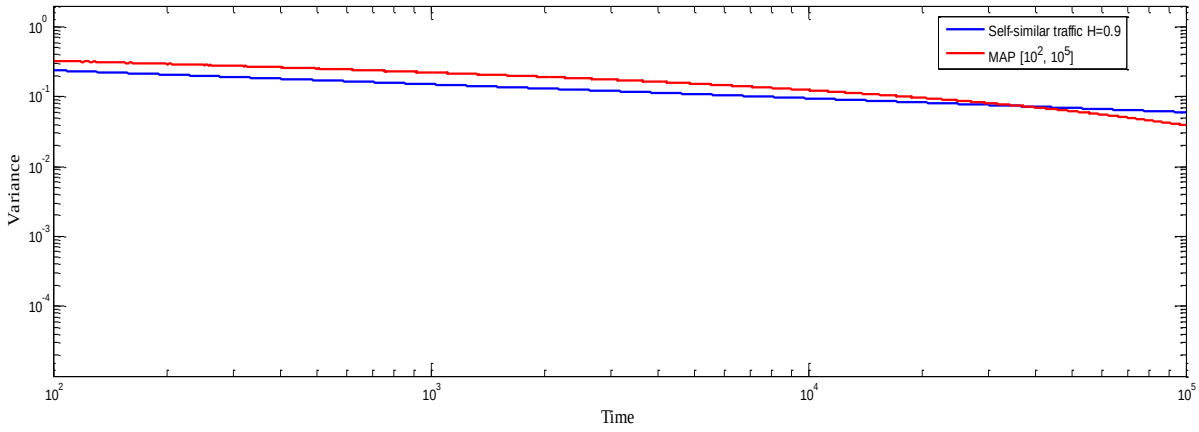


Figure 6: Variance versus time curves with $n = 4$ for sample3 during time scale $[10^2, 10^5]$

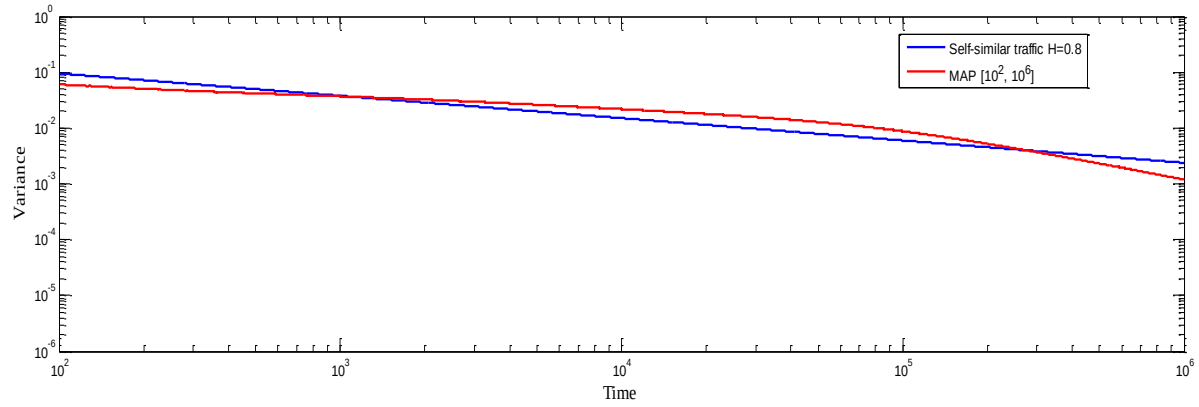


Figure 7: Variance versus time curves with $n = 4$ for sample1 during time scale $[10^2, 10^6]$

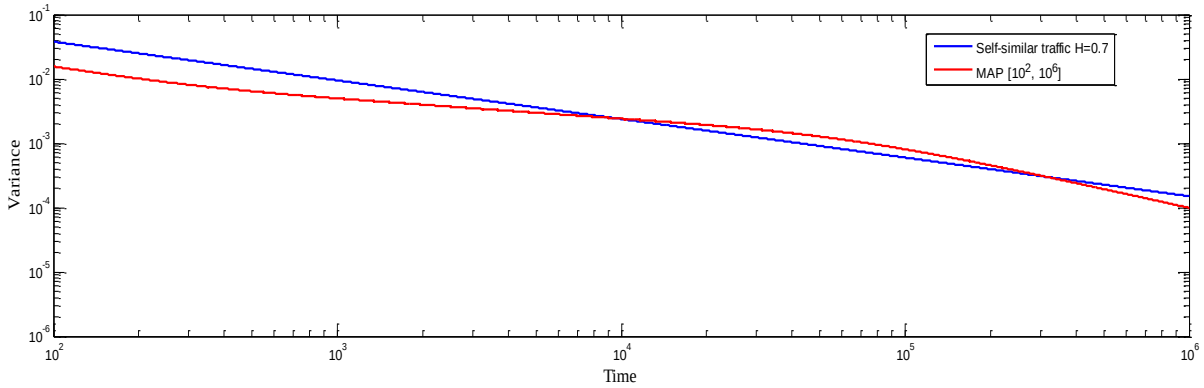


Figure 8: Variance versus time curves with $n = 4$ for sample2 during time scale $[10^2, 10^6]$

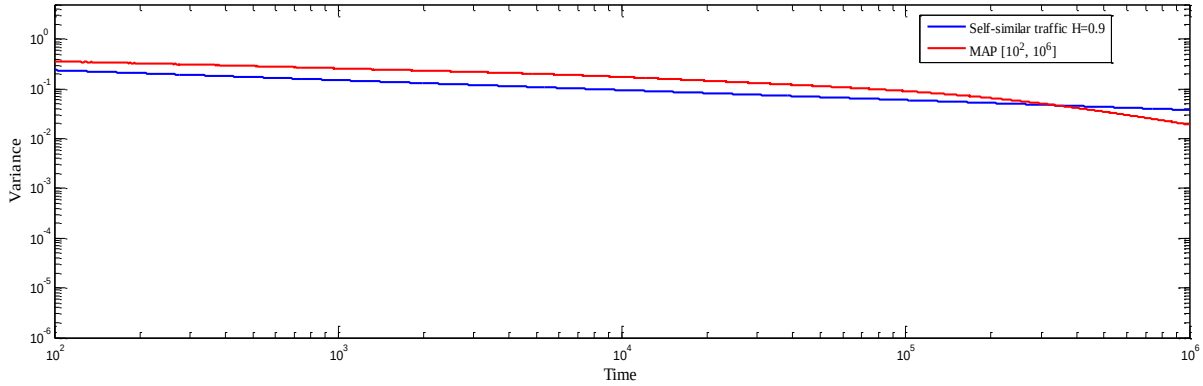


Figure 9: Variance versus time curves with $n = 4$ for sample3 during time scale $[10^2, 10^6]$

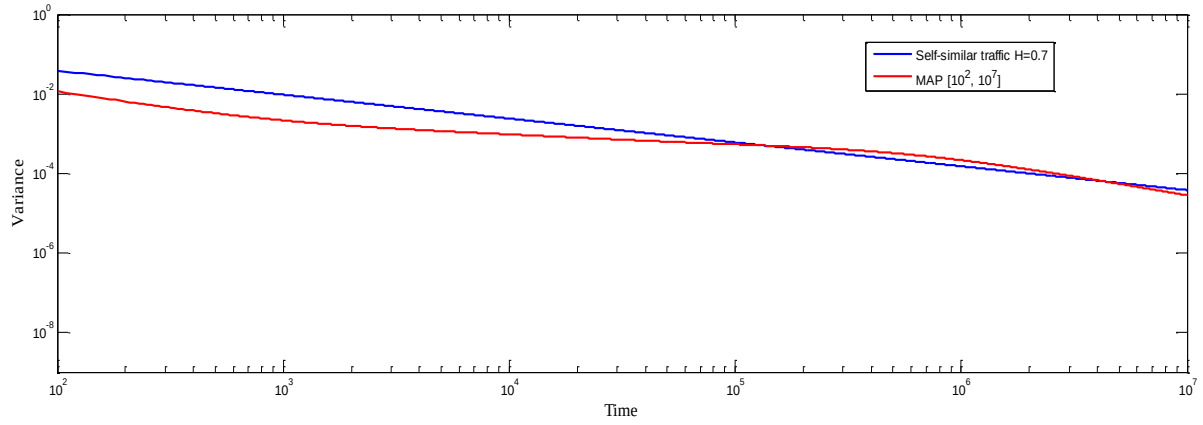


Figure 10: Variance versus time curves with $n = 4$ for sample1 during time $[10^2, 10^7]$

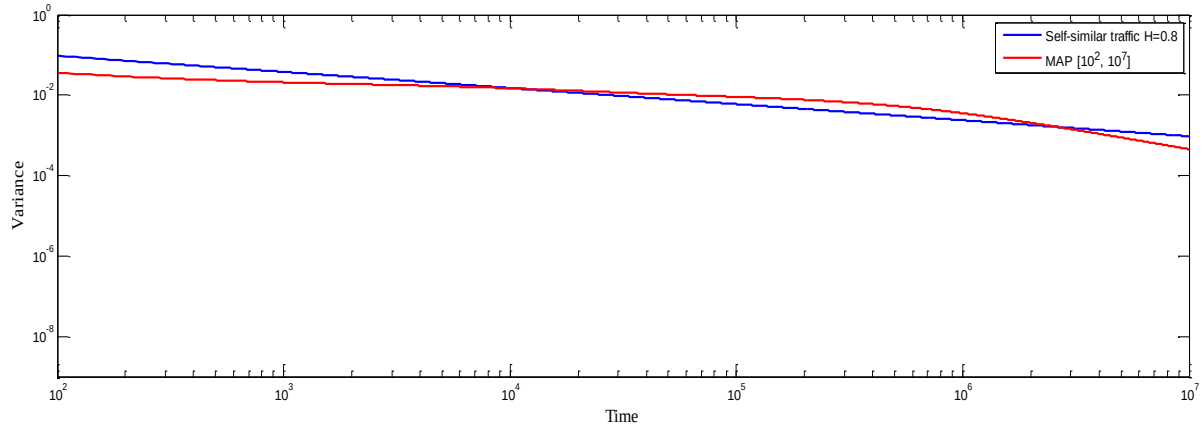


Figure 11: Variance versus time curves with $n = 4$ for sample2 during time $[10^2, 10^7]$

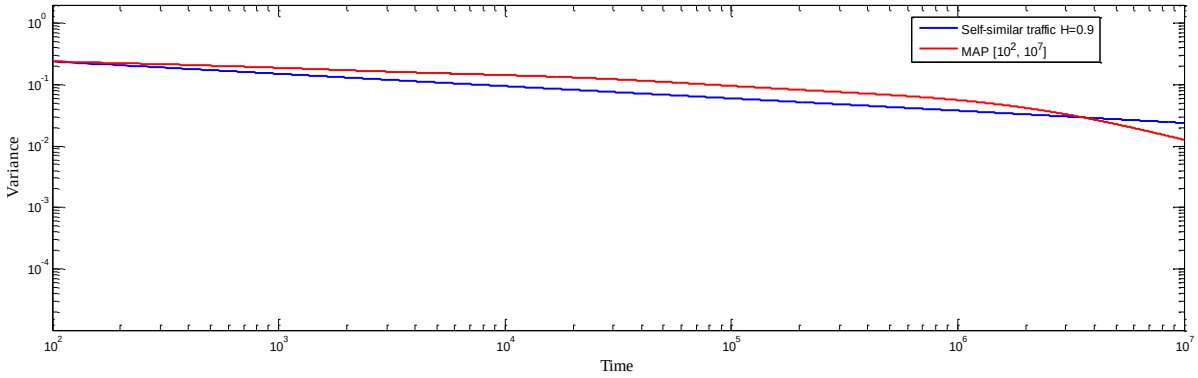


Figure 12: Variance versus time curves with $n = 4$ for sample3 during time $[10^2, 10^7]$

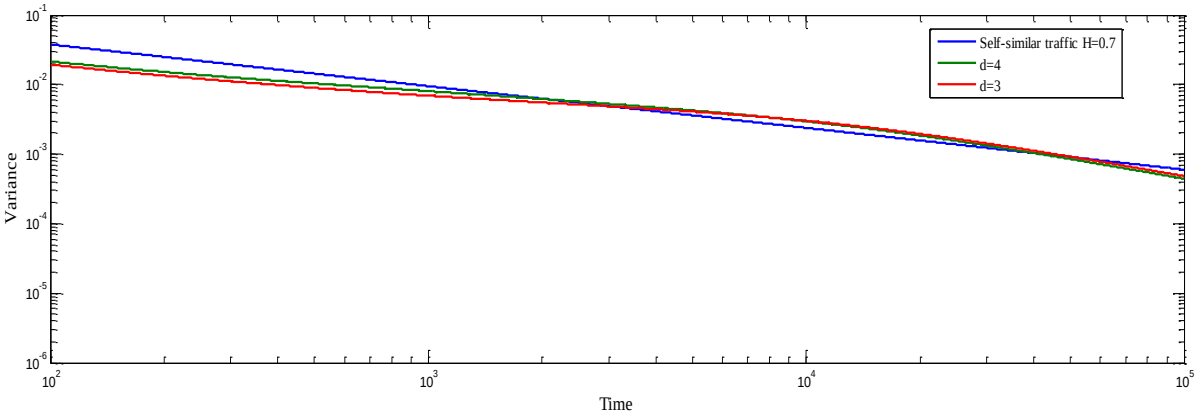


Figure 13: Variance vs time curves with $n = 3, 4$ over time-scale $[10^2, 10^5]$ for sample1

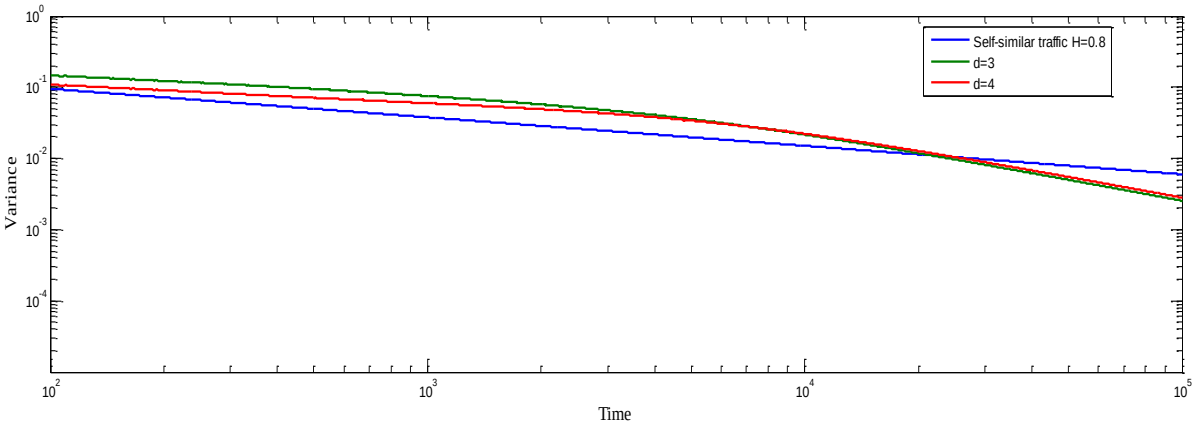


Figure 14: Variance vs time curves with $n = 3, 4$ over time-scale $[10^2, 10^5]$ for sample2

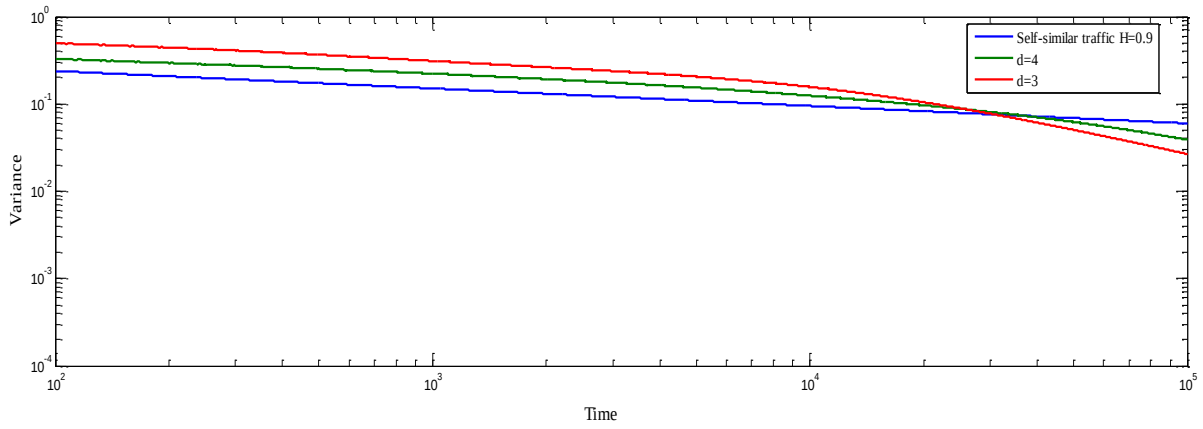


Figure 15: Variance vs time curves with $n = 3, 4$ over time-scale $[10^2, 10^5]$ for sample3

5. Conclusion

Self-similar traffic models proposed earlier are independent of time (homogeneous), and they do not work for time dependent queueing analysis. Here, variance based Markovian fitting procedure is presented using time dependent arrival rates. For the validation of fitting variance–time curves are presented, which show how the resultant MMPPs exhibit legitimate agreement with that of self-similar traffic in specified time scales. In addition, it is seen that the accuracy improved as number of MMPPs in superposition increases. This model is useful for time dependent queuing based performance analysis.

Acknowledgement

The authors wish to acknowledge Council of Scientific and Industrial Research (CSIR), Government of India, for their funding under the Major Research Project (MRP) scheme (File. No: 25(0301)/19/EMR-II).

References

Abate, J. and Whitt, W. (1988). Transient behavior of the M/M/1 queue via Laplace transforms. *Advances in Applied Probability*, **20**(**1**), 145-178.

Andersen, A. T. and Nielsen, B. F. (1998). A Markovian approach for modeling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications*, **16**(**5**), 719-732.

Coddington, A. E. and Levinson, N. (1987). *Theory of Ordinary Differential Equations*. 9th Reprint, Tata McGraw-Hill.

Crovella, M. E. and Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, **5**(**6**), 835-846.

Eick, S. G., Massey, W. A. and Whitt, W. (1993). $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Management Science*, **39**(**2**), 241-252.

Erramilli, A., Narayan, O. and Willinger, W. (1996). Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, **4**(**2**), 209-223.

Fisher, W. and Hellstern, M. K. (1993). The Markov-modulated Poisson process (MMPP) cook book. *Performance Evolution*, **18**(**2**), 149-171.

Heffes, H. and Lucantoni, D. (1986). A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, **4**(**6**), 856-868.

Jennings, O. B. and Massey, W. A. (1997). A modified offered load approximation for nonstationary circuit switched networks. *Telecommunication Systems,* **7**, 229-251.

Kasahara, S., Yoshihara, T. and Takahasi, Y. (2001). Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommunication Systems,* **17,** 185-211.

Kelton, W. D. and Law, A. M. (1985). The transient behavior of M/M/s queue with implications for steady state simulation. *Operation Research*, **33**(**2**), 378-396.

Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. (1994). On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transanctions on Networking*, **2**(**1**), 1-15.

Liu, Y. and Whitt, W. (2014). Stabilizing performance in networks of queue with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, **28**(**4**), 419-449.

Massey, W. A. (2002). The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems,* **21**, 173-204.

Misra, S., Banerjee, B. and Wolfinger, B. E. (2012). A learning automata-based uplink scheduler for supporting real-time multimedia interactive traffic in IEEE 802.16 WiMAX networks. *Computer Communications,* **35**(**15**), 1871-1881.

Neuts, M. (1979). A versatile Markovian point process. *Journal of Applied Probability*, **16**(**4**), 764-779.

Norros, I. (1994). A storage model with self-similar input. *Queueing Systems*, **16**, 387-396.

Pant, A. and Ghimire, R. (2016). M(t)/M/1 queueing system with sinusoidal arrival rate. *Journal of the Institute of Engineering*, **11**(**1**), 120-127.

Paxson, V. and Floyd, S. (1995). Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, **3**(**3**), 226-244.

Qian, Y. and Tipper, D. (2004). Adaptive channel allocation scheme for next generation wireless networks. *IEEE 60th Vehicular Technology Conference, VTC 2004-Fall*, Los Angels, CA, USA, **7,** 4918-4922.

Reddy, P. M., Shao, S. K., Tsai, M. G., Tsao, H. W. and Wu, J. (2005). Generalized variance-based Markovian fitting for self-similar traffic modeling. *IEICE Transactions on Communications*, **E88-B**(**4**), 1493-1502.

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
2022

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203