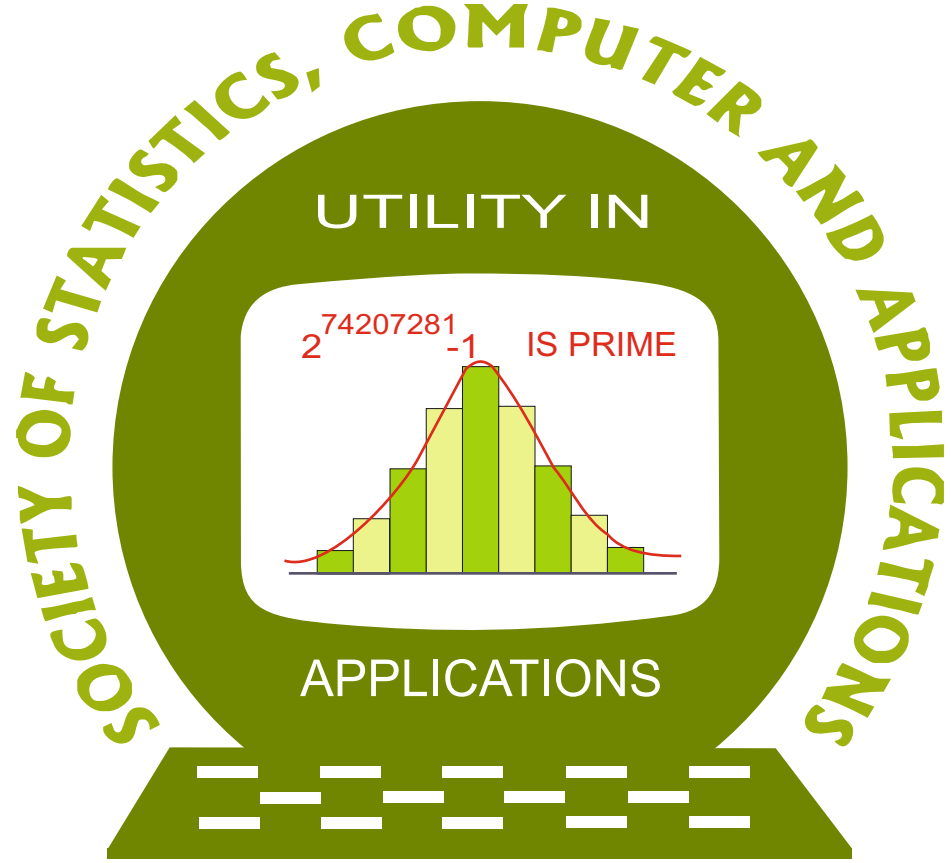
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Vol. 19, No. 2, 2021 (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
Sudhir Gupta

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri
Shibani Roy Choudhury

Council Members

Alka Arora
Anil Kumar Yadav
Anshu Bhardwaj
B. Re. Victor Babu
Manish Sharma
Manisha Pal
Piyush Kant Rai
Praggya Das
Ranjit Kumar Paul
Raosaheb V. Latpate
S.A. Mir
Sanjeev Panwar
V. Srinivasa Rao
V.M. Chacko
Vishal Deo

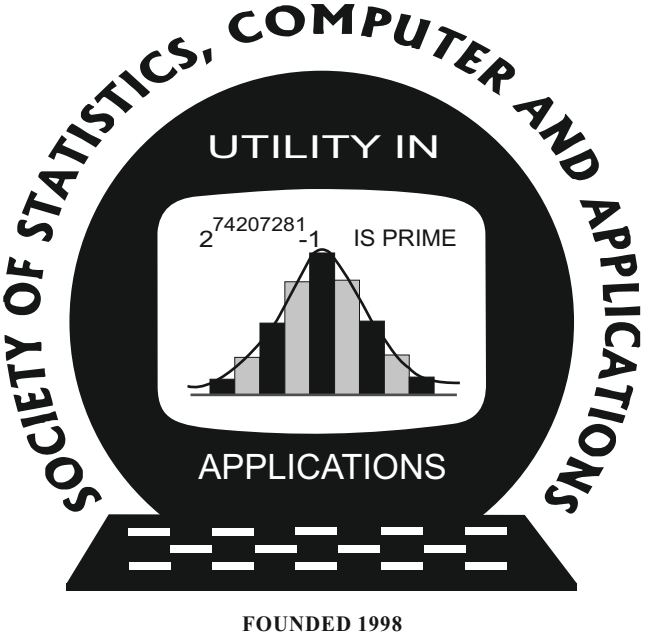
Ex-Officio Members (By Designation)

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, B - 133, Ground Floor, C.R. Park, New Delhi - 110019

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

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

Volume 19, No. 2, 2021 (New Series)

Statistics and Applications

Volume 19, No. 2, 2021 (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

CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 19, No. 2 (New Series) 2021

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 1-12

On DUS Transformed Weibull Distribution and its Properties

Kalsoom Akhtar Chaudhry and Javaria Shareef
Kinnaird College for Women, Lahore, Pakistan

Received: 27 March 2020; Revised: 27 June 2020; Accepted: 01 July 2020

Abstract

In Statistics literature, there are a number of methods to develop the new distributions. In this paper, a new distribution is developed using DUS transformation. A number of structural properties of this distribution such as moments, moment generating function, mean, median, mode, hazard rate and its shape, survival function and reverse hazard rate among others are derived. Further, the parameters of the newly developed distribution are estimated using method of moments, MLE and through simulation. The newly derived distribution was applied to two real data for the real life applications. The distribution will be a viable model for life-length of components and systems.

Key words: DUS Transformation; Survival Analysis; Hazard Rate; Cumulative distribution function; Maximum likelihood estimation.

1. Introduction

There are several methods to propose a new distribution using some baseline distribution. For example, Gupta *et al*. (1998) have proposed the cumulative distribution function (cdf) $G_1(x)$ of new distribution corresponding to the cdf, $F(x)$ of baseline distribution as,

$$G_1(x) = \{F(x)\}^a$$

where, $a > 0$ is the shape parameter.

Shaw and Buckley (2009) have developed a stimulating method called the quadratic rank transmutation map (QRTM) to develop the new distribution. It was used in order to form flexible distribution families by adding a new parameter to an existing distribution. Such family is called the transmuted extended distribution that holds the parental distribution as a special case and offers additional suppleness in order to model the numerous types of data sets.

If $G_2(x)$ is the cumulative function of transmuted distribution consistent to the baseline distribution having $F(x)$, then

$$G_2(x) = (1+ \lambda)F(x)- \lambda\{F(x)\}^2$$

where $|\lambda| \leq 1$.

Recently, various generalizations have been introduced based on QRTM such as transmuted extreme value distribution [see, Aryal and Tsokos (2011)], transmuted inverse

Corresponding Author: Kalsoom Akhtar Chaudhry
E-mail: Kalsoom.akhtar@kinnaird.edu.pk

Weibull distribution [see, Khan *et al*. (2014)], transmuted modified Weibull distribution [see, Khan and King (2013)], transmuted log-logistic distribution [see, Aryal (2013)], transmuted exponential distribution (Kumar *et al*. (2015)) and many more.

$$g(x)=\frac{1}{e-1} f(x)\, e^{F(x)} \tag{1}$$

The transformation (1) is known as DUS transformation and is used for generating the new distribution. The cumulative function and hazard rate consistent to the g(x) are specified in (2) and (3) respectively.

$$G(x)=\frac{1}{e-1}\left[e^{F(x)}-1\right] \tag{2}$$

and

$$h(x)=\frac{1}{e-e^{F(x)}} f(x) e^{F(x)} \tag{3}$$

2. DUS Transformation of Weibull Distribution

In this section, we have proposed a probability density function of a newly formed distribution obtained using DUS transformation technique for Weibull distribution as a baseline distribution. The distribution will be useful for lifetime modeling.

Using equation (1) the probability density function of DUS$_W$(k, λ)-distribution is given by

$$g(x) =\frac{1}{e-1} f(x)\, e^{F(x)} \tag{4}$$

The probability density function of the two parameter Weibull distribution is

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} \tag{5}$$

and

$$F(x) = 1- e^{-\left(\frac{x}{\lambda}\right)^{k}} \tag{6}$$

Now, putting f(x) and F(x) in equation (1), we get

$$g(x) = \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} e^{\left(1- e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}, x>0, \lambda>0, k>0 \tag{7}$$

Equation (7) represents the probability density function of DUS$_W$(k, λ)-distribution (DUS transformed Weibull distribution) with k as a shape parameter and λ as a scale parameter. The shape of newly developed distribution for various values of parameters is given.

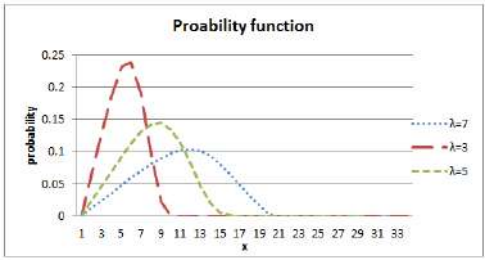


Figure 1: Probability density function of DUS$_W$(k, λ)-distribution when k = 2 is fixed and λ is varied (λ = 3, 5, 7)

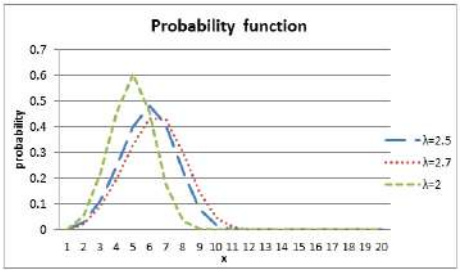


Figure 2: Probability density function of DUS$_W$(k, λ)-distribution when k = 3 is fixed and scale parameter λ is varied (λ = 2.5, 2.7, 2)

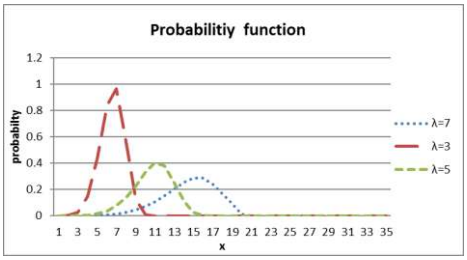


Figure 3: Probability density function of DUS$_W$(k, λ)-distribution when k = 5 is fixed and scale parameter λ is varied (λ = 3, 5, 7)

The shape of DUS transformed is pretty flexible, including moderately positively skewed, approximately symmetric and moderately negatively skewed shapes for different values of parameters, the DUS$_W$(k,λ)-distribution seems to be a viable model for life-length of components and systems as well as non-negative variables. The cdf of DUS$_W$(k,λ)-distribution can be written as

$$G(x)=\frac{1}{e-1}\left(e^{\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}-1\right) \tag{8}$$

whereas the survival function of the distribution is obtained as

$$S(x)=1-\frac{1}{e-1}\left(e^{\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}-1\right) \tag{9}$$

Using equations (8) and (9), the hazard function is obtained

$$h(x)=\left(\frac{\left(\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)e^{\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}}{\left(e-e^{\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}\right)}\right) \tag{10}$$

Figure 4: Survival function of DUS$_W$(k, λ)-distribution when k = 3 is fixed and λ is varied (λ = 2.5, 2.7, 2)

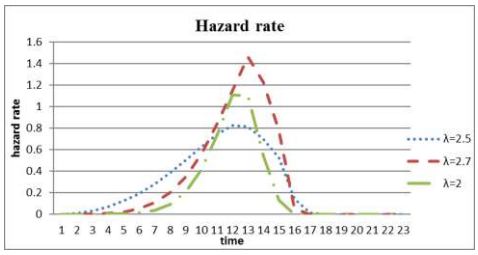


Figure 5: Hazard rate of DUS$_W$(k, λ)-distribution when k = 3 is fixed and λ is varied (λ = 2.5, 2.7, 2)

The graph of hazard rate shows that at the starting of time, hazard rate has an increasing trend whereas after completing its median time (approximately), it drastically goes down to zero.

3. Statistical Properties of DUS$_W$(k, λ) Distribution

The mean of DUS$_W$(k, λ) distribution is obtained as

$$E(x)=\int_0^{\infty} x\, g(x)dx$$

$$E(x)=\lambda\frac{e}{e-1}\sum_{i=0}^{\infty}\frac{(-1)^i}{i!}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}\Gamma\left(i+\frac{1}{k}+1\right)\ \frac{1}{j^{\left(i+\frac{1}{k}+1\right)}} \tag{11}$$

whereas the median of $DUS_W(k, \lambda)$ distribution is the solution of the following,

$$G(M)=\frac{1}{2}$$

for M and the same is obtained as follows,

$$\text{Median} = m = \lambda\left(\ln\left(\frac{1}{(1-\ln(1+(e-1)0.5))}\right)\right)^{\frac{1}{k}} \tag{12}$$

In order to obtain the mode of the distribution, differentiating equation (7) with respect to x, we get

$$g'(x)=\left(\frac{k}{\lambda}\right)\left(\frac{1}{e-1}\right)\left(\frac{(k\text{-}1)}{\lambda}\left(\frac{x}{\lambda}\right)^{k-2}\left(-\frac{1}{\lambda}\right)^{k}\ e^{-\left(\frac{x}{\lambda}\right)^{k}}\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)\left(-e^{\left(1-e^{-\left(\frac{x}{\lambda}\right)^{k}}\right)}\right)\right) \tag{13}$$

It is easy to show that $g''(x)$ is a decreasing function hence the expression for the mode may be obtained by putting equation (13) equal to zero.

The harmonic mean of $DUS_W(k, \lambda)$ distribution is obtained by solving the following expression and is obtained as

$$\frac{1}{H}=\int_0^{\infty}\frac{1}{x}g(x)dx$$

$$H=\left[\frac{1}{\lambda}\frac{e}{e\text{-}1}\sum_{i=0}^{\infty}\frac{(-1)^i}{i!}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}\Gamma\left(i-\frac{1}{k}+1\right)\ \frac{1}{j^{\left(i-\frac{1}{k}+1\right)}}\right]^{-1} \tag{14}$$

The variance of $DUS_W(k, \lambda)$ distribution can be obtained as

$$\text{Var}(x)=\lambda^2\frac{e}{e\text{-}1}\sum_{i=0}^{\infty}\frac{(-1)^i}{i!}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}\Gamma\left(i+\frac{2}{k}+1\right)\ \frac{1}{j^{\left(i+\frac{2}{k}+1\right)}}-\left(\lambda\frac{e}{e\text{-}1}\sum_{i=0}^{\infty}\frac{(-1)^i}{i!}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}\Gamma\left(i+\frac{1}{k}+1\right)\frac{1}{j^{\left(i+\frac{1}{k}+1\right)}}\right)^2 \tag{15}$$

The moment generating function of $DUS_W(k, \lambda)$ distribution is obtained as

$$M_x(t)=E(e^{tx})$$

$$E(e^{tx})=\frac{e}{e\text{-}1}\ \sum_{i=0}^{\infty}\frac{(-1)^i}{i!}\left[i-\frac{t\lambda}{k}+1\right]^{-1} \tag{16}$$

The characteristic function of $DUS_W(k, \lambda)$-distribution for the variable X is obtained as

$$\phi_x(t) = \mathrm{E}(e^{itx})$$

$$\mathrm{E}(\mathrm{e}^{itx}) = \frac{e}{e\text{-}1} \sum_{j=0}^{\infty} \frac{(\text{-}1)^j}{j!} \left[\frac{j\text{-}it\lambda+1}{k}\right]^{-1} \tag{17}$$

The raw moments of $\mathrm{DUS}_W(k, \lambda)$ distribution are obtained as follow

$$\mu'_r = \mathrm{E}(x^r) = \lambda^r \frac{e}{e\text{-}1} \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \Gamma\left(i + \frac{r}{k} + 1\right) \frac{1}{j^{\left(i+\frac{r}{k}+1\right)}} \tag{18}$$

The quantile function for $\mathrm{DUS}_W(k, \lambda)$-distribution is obtained as

$$x_q = \lambda \left[\ln\left[\frac{1}{1-\ln(1+q(e-1))}\right]\right]^{\frac{1}{k}} \tag{19}$$

4. Estimation of the Parameters of $\mathrm{DUS}_W(k, \lambda)$ Distribution

In order to assess the real life application of the $\mathrm{DUS}_W(k, \lambda)$ distribution, the parameters of the distribution are estimated. We estimate the parameters 'k' and 'λ' of $\mathrm{DUS}_W(k,\lambda)$ distribution using the maximum likelihood estimation method. By definition

$$\frac{\partial \ln L(k; x_1, x_2, \ldots, x_n)}{\partial k} = 0 \tag{20}$$

and

$$\frac{\partial \ln L(\lambda; x_1, x_2, \ldots, x_n)}{\partial \lambda} = 0 \tag{21}$$

So solving the equations simultaneously, we have

$$\frac{n}{k} + \ln\left[\prod_{i=1}^{n} \left(\frac{x_i}{\lambda}\right)^{k-1}\right] - \sum_{i=1}^{n} \left(\frac{x_i}{\lambda}\right)^k \ln\left(\frac{k}{\lambda}\right) - \sum_{i=1}^{n} \left(\mathrm{e}^{1-\left(\frac{x_i}{\lambda}\right)^k}\right) \left(\sum_{i=1}^{n} \left(\frac{x_i}{\lambda}\right)^k \ln\left(\frac{k}{\lambda}\right)\right) = 0 \tag{22}$$

$$-\frac{n}{\lambda^2} + \frac{1-k}{\lambda} + \frac{k}{\lambda} \sum_{i=1}^{n} \left(\frac{x_i}{\lambda}\right)^k - \frac{k}{\lambda} \sum_{i=1}^{n} \left(e^{1-\left(\frac{x_i}{\lambda}\right)^k}\right) \left(\frac{x_i}{\lambda}\right)^k = 0 \tag{23}$$

Equations (22) and (23) were difficult to solve analytically, some numerical methods may be used to solve the equations simultaneously for k and λ respectively. In order to estimate the parameters analytically, we have estimated the parameters using method of moments. Following equations will be solved to estimate the parameters k and λ,

$$\acute{\mu}_1 = m'_1$$

and

$$\acute{\mu}_2 = m_2'$$

$$\hat{\lambda} = \frac{\bar{x}}{\frac{e}{(e\text{-}1)}\sum_{i=1}^{\infty}\frac{(-1)^i}{i!}\sum_{j=1}^{\infty}\frac{(-1)^j}{j!}\Gamma\frac{(k(i+1)+1)}{k}\frac{1}{j^{\left(\frac{k(i+1)+1}{k}\right)}}} \tag{24}$$

for parameter k,

$$\frac{1}{n}\left(A\left(\frac{e-1}{e}\right)BC\right) = 0$$

where $A = n\sum x_i^2 - (\sum x_i)^2;\ B = \left(\frac{1}{\sum_{i=1}^{\infty}\frac{(-1)^i}{i!}\sum_{j=1}^{\infty}\frac{(-1)^j}{j!}\Gamma\frac{(k(i+1)+1)}{k}\frac{1}{j^{\left(\frac{k(i+1)+1}{k}\right)}}}\right)^2$ and

$$C = \sum_{i=1}^{\infty}\frac{(-1)^i}{i!}\sum_{j=1}^{\infty}\frac{(-1)^j}{j!}\ \Gamma\left(\frac{(k(i+1)+2)}{k}\right)\ \frac{1}{j^{\left(\frac{k(i+1)+2}{k}\right)}} = 0. \tag{25}$$

As the expression for k may not be solved analytically therefore numerical method may be used to estimate the parameters k and λ. A simulation study is carried out taking 1000 samples of various sizes n drawn from the $DUS_W(k, \lambda)$ distribution for different values of the parameters k and λ. For inversion theorem the relation $X = F^{-1}(u)$ is used to generate the random values for the variable X with the given distribution function. By definition

$$F(x) = u$$

$$x = F^{-1}(u)$$

$$\hat{x}\ = \lambda\left[\ln\left[\frac{1}{1-\ln(1+u(e-1))}\right]\right]^{1/k} \tag{26}$$

Hence the above expression is used to generate random samples from the $DUS_W(k, \lambda)$ distribution for the given values of the parameters. A computer program is developed to obtain the mean values of the $DUS_W(k, \lambda)$-distribution using R language. For each pair of values (k, λ), various values of the mean of means are obtained. For a given data, the mean will be calculated and the parameters will be estimated for the given mean using the Tables generated for DUS transformed Weibull distribution. The values of the mean of transformed data of the DUS Weibull distribution are presented in the Tables 1 to 9 in the Appendix.

5. Real Life Application

To assess the applicability of $DUS_W(k, \lambda)$ distribution, we have considered a real data of 63 observations related to the strengths of 1.5 cm glass fibers. This set was obtained by workers at the UK National Physical Laboratory and was used by Smith and Naylor (1987) whereas the second data set was about the hole diameter (Dasgupta, 2011). The first data set is related to the strengths of 1.5 cm glass fibers, a total of 63 observations were obtained and are given as follows

0.55, 0.74, 0.77, 0.81, 0.84, 1.24, 0.93, 1.04, 1.11, 1.13, 1.30, 1.25, 1.27, 1.28, 1.29, 1.48, 1.36, 1.39, 1.42, 1.48, 1.51, 1.49, 1.49, 1.50, 1.50, 1.55, 1.52, 1.53, 1.54, 1.55, 1.61, 1.58, 1.59, 1.60, 1.61, 1.63, 1.61, 1.61, 1.62, 1.62, 1.67, 1.64, 1.66, 1.66, 1.66, 1.70, 1.68, 1.68, 1.69, 1.70, 1.78, 1.73, 1.76, 1.76, 1.77, 1.89, 1.81, 1.82, 1.84, 1.84, 2.00, 2.01, 2.24.

After arranging the above data, arithmetic mean of the transformed data is calculated which comes out to be 1.506. Now searching this value in table 5, we find that the value of the mentioned mean is 1.506 for $k = 3$, $\lambda = 1.5$ and $n = 63$. The $DUS_W(k, \lambda)$ distribution is fitted on the data using $\lambda = 1.5$ and $k = 3$. The chi-square goodness of fit test ($\chi^2 = 3.9168$, $p = 0.86$) revealed that the $DUS_W(k, \lambda)$ model is a good fit model on the data of strengths of glass fibers. Further, the Weibull distribution is fitted on the data for the same choice of the parameters $\lambda = 1.5$ and $k = 3$. The chi-square goodness of fit revealed that the $DUS_W(k, \lambda)$ distribution is a better fit model compared to two-parameter Weibull distribution**.**

The second data set of 50 observations (in the unit of millimeter) is related to different machines under comparison for the similar operations in the same site of a factory and was used by Dasgupta (2011). The observations are given below

0.04, 0.02, 0.06, 0.12, 0.14, 0.08, 0.22, 0.12, 0.08, 0.26, 0.24, 0.04, 0.14, 0.16, 0.08, 0.26, 0.32, 0.28, 0.14, 0.16, 0.24, 0.22, 0.12, 0.18, 0.24, 0.32, 0.16, 0.14, 0.08, 0.16, 0.24, 0.16, 0.32, 0.18, 0.24, 0.22, 0.16, 0.12, 0.24, 0.06, 0.02, 0.18, 0.22, 0.14, 0.06, 0.04, 0.14, 0.26, 0.18, 0.16.

After arranging the above mentioned data, its mean is calculated as 0.1632. Now searching this value in Table 4, we have found that the value of the mentioned mean is 0.1632 for $k = 1.5$, $\lambda = 0.15$ and $n = 50$. The chi-square goodness of fit test ($\chi^2 = 24.8039$, $p = 0.81$) concluded that the proposed model is a good fit for the given data set.

6. Conclusion

From the simulation study, it is evident that the proposed DUS transformed Weibull distribution is a flexible model for application. The distribution may be used as a lifetime model and can be fitted on the life length of various components.

Acknowledgements

The authors are grateful to the editor and anonymous referees for the valuable comments to improve the paper.

References

Aryal, G. R. (2013). Transmuted log-logistic distribution. *Journal of Statistics Applications and Probability,* **2** (**1**), 11-20.

Aryal, G. R. and Tsokos, C. P. (2011). Transmuted Weibull distribution: A generalization of the Weibull probability distribution. *European Journal of Pure and Applied Mathematics,* **4**(**1**), 89-102.

Dasgupta, R. (2011). On the distribution of Burr with applications. *Sankhya,* **B73(1)**, 1-19.

Gupta, R. C. Gupta, R. D. and Gupta, P. L. (1998). Modeling failure time data by Lehman alternatives. *Communication in Statistics-Theory and Methods*, **27**(**4**), 887-904.

Khan, M. S. and King, R. (2013). Transmuted modified Weibull distribution: A generalization of the modified Weibull probability distribution. *European Journal of Pure and Applied Mathematics*, **6**(**1**), 66–88.

Khan, M. S., King, R. and Hudson, I. L. (2014). Characterizations of the transmuted inverse Weibull distribution. *Journal of Statistical Theory and Applications* **55**(**3**), 197–217.

Kumar, D., Singh, U., and Singh, S. K. (2015). A method of proposing new distribution and its application to bladder cancer patients data. *Journal of Statistics Applications and Probability Letters*, **2**(**3**), 235-245.

Shaw, W. and Buckley, I. (2009). The alchemy of probability distributions: beyond Gram-Charlier expansions and a skew-kurtotic-normal distribution from a rank transmutation map. *Research Report, King's College, London, U.K.*

Smith, R. L. and Naylor, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **36**(**3**), 358–369.

APPENDIX

Table 1: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 10

k/λ	2	2.2	2.3	2.4	2.5	2.6	2.7	3	3.4
1	2.517245	2.811399	2.885155	2.986473	3.137331	3.283466	3.399174	3.787712	4.336345
1.15	2.376857	2.566277	2.698917	2.861838	2.956662	3.07399	3.183724	3.470086	3.948485
1.25	2.23171	2.469243	2.583506	2.736134	2.837344	2.939732	3.041505	3.372101	3.849027
1.5	2.155965	2.33097	2.477656	2.59744	2.696651	2.774533	2.891594	3.204253	3.608331
1.75	2.076177	2.275707	2.379886	2.484644	2.620812	2.696242	2.805161	3.126713	3.504668
2	2.023193	2.234959	2.353963	2.440517	2.529771	2.645726	2.742195	3.043873	3.435698
2.25	2.00609	2.214886	2.310757	2.399094	2.512973	2.591598	2.722587	3.031537	3.392291
2.5	2.000772	2.182101	2.290734	2.376628	2.49335	2.581597	2.691004	2.97438	3.375509
3	1.969566	2.158932	2.272332	2.362112	2.460654	2.562589	2.644976	2.961288	3.343580

Table 2: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 20

k/λ	0.2	0.4	0.45	2	2.3	2.5	3	3.3	3.5
1	0.251216	0.569388	0.569388	2.504554	2.904909	3.167895	3.755406	4.124498	4.377891
1.15	0.235446	0.530137	0.530137	2.347480	2.714062	2.929789	3.530666	3.880406	4.109790
1.25	0.225846	0.511246	0.511246	2.281141	2.614768	2.839802	3.405611	3.718937	3.980987
1.5	0.214381	0.481564	0.481564	2.149555	2.466677	2.684022	3.214789	3.561088	3.778796
1.75	0.206709	0.468092	0.468092	2.083877	2.386881	2.599988	3.102101	3.449115	3.610533
2	0.204347	0.456636	0.456636	2.028692	2.342961	2.523202	3.049356	3.34684	3.555917
2.5	0.199243	0.447806	0.447806	1.991692	2.284143	2.486346	2.985243	3.272693	3.481387
3	0.195725	0.443646	0.443646	1.970628	2.262800	2.45306	2.960755	3.261811	3.443872
3.2	0.189226	0.390668	0.390668	1.894560	2.241700	2.42293	2.46892	3.241709	3.422876

Table 3: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 30

k/λ	2	2.3	2.5	2.7	2.8	3	3.3	3.5
1	2.504009	2.912170	3.171027	3.404401	3.499660	3.746924	4.135979	4.470881
1.15	2.347347	2.887324	2.949099	3.162511	3.252916	3.554996	3.875747	4.093062
1.25	2.273302	2.615983	2.832352	3.051734	3.186968	3.396030	3.746377	3.968723
1.5	2.138868	2.470648	2.685668	2.898625	3.009160	3.225825	3.527038	3.756814
1.75	2.071820	2.385379	2.599343	2.804341	2.895190	3.112046	3.427742	3.641127
2	2.038699	2.329662	2.554939	2.746986	2.842897	3.056338	3.365053	3.554760
2.25	2.003136	2.305109	2.500910	2.707960	2.803177	3.007277	3.308174	3.509270
2.5	1.986168	2.281053	2.486458	2.674299	2.794896	2.997515	3.268091	3.482483
3	1.971597	2.267932	2.454803	2.661883	2.754530	2.956542	3.243436	3.436502

Table 4: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 50

k/λ	0.1	0.15	0.2	0.5	1	1.5	2	2.5	3
1	0.126208	0.189445	0.252161	0.629009	1.254004	1.898039	2.531597	3.138138	3.777200
1.15	0.116912	0.176062	0.234553	0.584946	1.175404	1.755803	2.339846	2.929436	3.504400
1.25	0.113295	0.170105	0.226849	0.565130	1.134217	1.698214	2.281033	2.83516	3.390600
1.5	0.107107	0.164074	0.214397	0.534685	1.071657	1.609600	2.148722	2.666563	3.208000
1.75	0.103833	0.160124	0.206910	0.518443	1.040654	1.558920	2.074412	2.594035	3.101780
2	0.101413	0.158331	0.203733	0.507773	1.015262	1.519897	2.032245	2.540005	3.050230
2.5	0.099707	0.155434	0.198530	0.498642	0.992869	1.490626	1.993589	2.499018	2.977090
3	0.098321	0.149604	0.196730	0.491703	0.984630	1.475060	1.970222	2.462718	2.955900
3.5	0.087644	0.142777	0.192780	0.468930	0.945520	1.439970	1.935520	2.390330	2.909700

Table 5: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 63

k/λ	0.5	0.75	1	1.5	1.75	2	2.5	3	3.5
0.5	1.399557	2.136622	2.865456	4.554282	4.869946	5.709310	7.132205	8.457595	9.895400
0.75	0.784859	1.177411	1.561745	2.469169	2.763686	3.141139	3.915627	4.710288	5.481340
1	0.631977	0.943295	1.256300	1.987784	2.210679	2.528557	3.140937	3.748432	4.414500
1.5	0.536525	0.803973	1.070778	1.872589	1.890160	2.147369	2.681373	3.225480	3.755600
1.75	0.519181	0.776352	1.034615	1.685650	1.814858	2.074194	2.594843	3.118343	3.628500
2	0.507381	0.763716	1.015113	1.605890	1.773439	2.035301	2.542794	3.040332	3.558100
2.5	0.497888	0.746088	0.995065	1.591450	1.746025	1.987331	2.483504	2.989640	3.486800
2.75	0.493780	0.741295	0.989270	1.556450	1.729200	1.974721	2.471013	2.965545	3.455800
3	0.492343	0.739055	0.982879	1.510709	1.728782	1.965731	2.456757	2.955809	3.443000
3.5	0.489345	0.734759	0.979046	1.480709	1.714510	1.955997	2.44798	2.934148	3.429600

Table 6: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size *n* = 100

k/λ	2	2.3	2.5	2.7	2.8	3	3.3	3.5
1	2.531597	2.894259	3.138138	3.422545	3.540568	3.777278	4.169518	4.433289
1.15	2.339846	2.696134	2.929436	3.159015	3.289008	3.504461	3.879313	4.098301
1.25	2.281033	2.603607	2.83516	3.058677	3.170763	3.390670	3.723769	3.979242
1.5	2.148722	2.475804	2.666563	2.906334	3.005207	3.208049	3.528592	3.767585
1.75	2.074412	2.377109	2.594035	2.795645	2.900345	3.101789	3.436162	3.634684
2	2.032245	2.332237	2.540005	2.735711	2.76723	3.050239	3.34302	3.556535
2.5	1.993589	2.283756	2.499018	2.687975	2.785568	2.977095	3.277005	3.479074
3	1.970222	2.267965	2.462718	2.658942	2.749937	2.955899	3.248275	3.438107

Table 7: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size *n* = 300

k/λ	2	2.3	2.5	2.7	2.8	3	3.3	3.5
1	2.524565	2.898132	3.142753	3.403772	3.528445	3.774485	4.156479	4.411083
1.15	2.346215	2.699438	2.937056	3.174084	3.285149	3.517790	3.878156	4.102618
1.25	2.272043	2.605824	2.832281	3.063584	3.174093	3.407877	3.748495	3.969933
1.5	2.148019	2.466352	2.677368	2.896174	2.996944	3.222330	3.540961	3.750759
1.75	2.076095	2.386813	2.598522	2.802113	2.905126	3.114686	3.425741	3.625521
2	2.034038	2.336548	2.543534	2.743206	2.845385	3.049329	3.357493	3.556055
2.5	1.988411	2.288006	2.482313	2.683379	2.782118	2.984053	3.277599	3.482176
3	1.967270	2.264435	2.460363	2.656226	2.757550	2.949514	3.248816	3.442170

Table 8: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size *n* = 500

k/λ	2	2.3	2.5	2.7	2.8	3	3.3	3.5
1	2.518699	2.898791	3.154413	3.399437	3.531351	3.777867	4.154576	4.404194
1.15	2.345587	2.699761	2.939658	3.166351	3.287146	3.523665	3.877953	4.105032
1.25	2.268411	2.609544	2.836316	3.063409	3.178322	3.406563	3.743683	3.971175
1.5	2.143511	2.463945	2.684016	2.892896	3.004482	3.218393	3.536348	3.754384
1.75	2.075971	2.388087	2.596298	2.801879	2.90648	3.110323	3.423013	3.632914
2	2.033545	2.33581	2.540464	2.742854	2.846879	3.051042	3.355247	3.556760
2.5	1.988111	2.286722	2.484683	2.684510	2.782593	2.982218	3.278046	3.480877
3	1.969891	2.263312	2.461614	2.657846	2.757536	2.950255	3.252892	3.443803

Table 9: Table of the means of ($\bar{X}_{\text{trans weibull}}$) when sample size n = 1000

k/λ	2	2.3	2.5	2.7	2.8	3	3.3	3.5
1	2.519694	2.898938	3.150607	3.398360	3.528852	3.781955	4.159814	4.407745
1.15	2.344847	2.699903	2.935890	3.168151	3.285946	3.523283	3.872619	4.106633
1.25	2.271496	2.609546	2.836777	3.063625	3.180062	3.403251	3.744170	3.971481
1.5	2.146576	2.468929	2.683170	2.896734	3.007135	3.215928	3.538355	3.753207
1.75	2.076046	2.388072	2.595777	2.801017	2.904149	3.113252	3.424946	3.629920
2	2.033407	2.340196	2.542027	2.745568	2.845711	3.049193	3.353305	3.556922
2.5	1.988675	2.285758	2.484805	2.684776	2.782991	2.983072	3.283154	3.479466
3	1.968194	2.263994	2.460396	2.656330	2.754646	2.952897	3.247433	3.446004

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 13-26

Mixture Designs Generated Using Orthogonal Arrays From Mutually Orthogonal Latin Squares

Poonam Singh[1], **Vandana Sarin**[2] **and Neha Midha**[1]
[1]*Department of Statistics, University of Delhi, Delhi 110007, India*
[2]*Department of Statistics, Kirori Mal College, University of Delhi, Delhi 110007, India*

Received: 20 March, 2020; Revised: 30 June, 2020; Accepted: 06 July, 2020

Abstract

This paper proposes an algorithm for constructing mixture designs based on orthogonal arrays of index unity containing the smallest number of runs for a given number of levels and a given strength using mutually orthogonal Latin squares. The algorithm allows the generation of cost effective and efficient mixture designs for Scheffé's canonical polynomials.

Key words: Mixture experiments; Mutually orthogonal Latin squares; Restricted region; G-efficiency.

1. Introduction

In experiments with mixtures, the response is a function only of the proportions of the q components present in the mixture and not of the total amount of the mixture. If x_i is the proportion of ith component, $i = 1, 2, \ldots, q$, then

$$0 \leq x_i \leq 1, \quad \sum_{i=1}^{q} x_i = 1 \tag{1}$$

These restrictions force the factor space of the q components to take form of a $(q-1)$ dimensional simplex. The general purpose of mixture experimentation is to estimate the properties of an entire multicomponent system from only a limited number of observations. These observations are taken at preselected combinations of the components to determine which of the combinations in some sense optimize the response.

In many practical situations, one can encounter certain additional constraints that are placed on some or all component proportions besides (1). These are of the form

$$0 \leq L_i \leq x_i \leq U_i \leq 1;\ i = 1, 2, \ldots, q \tag{2}$$

where, L_i and U_i denote the lower bound and upper bound for the component proportion x_i; $i = 1, 2, \ldots, q$. These supplementary restrictions limit the experimentation to some sub-region of the simplex, thereby altering the shape of the experimental region from a simplex to an irregularly shaped convex polyhedron inside the simplex. In such situations, directing the design and modelling only to the sub-region can help in lowering the experimentation cost and time and increasing the precision of model estimates.

Corresponding author: Neha Midha
E-mail: nehamidha1189@gmail.com

Mixture designs have a variety of applications in several industries. Amongst many others, Cafaggi *et al.* (2003) illustrated the application of a constrained mixture design to a pharmaceutical formulation. Mirabedini *et al.* (2012) discussed the application of mixture designs for the formulation of thermoplastic road markings. Schrevens and Cornell (1993) analysed the mixture designs for plant nutrition research. Buruk *et al.* (2016) reviewed the recent applications of mixture designs in the food industry.

Scheffé (1958, 1963) was the first to develop simplex lattice and simplex centroid designs for fitting the canonical polynomial models:

Linear model: $Y=\sum_{i=1}^{q}\beta_i x_i+\varepsilon$ (3)

Quadratic model: $Y=\sum_{i=1}^{q}\beta_i x_i+\sum\sum_{i<j}^{q}\beta_{ij}x_i x_j+\varepsilon$ (4)

Special cubic model: $Y=\sum_{i=1}^{q}\beta_i x_i+\sum\sum_{i<j}^{q}\beta_{ij}x_i x_j+\sum\sum\sum_{i<j<k}^{q}\beta_{ijk}x_i x_j x_k+\varepsilon$ (5)

McLean and Anderson (1966) developed extreme vertices designs (EVD) which satisfy both the constraints (1) and (2). A partial solution to the restricted exploration problem is the work of Thompson and Myers (1968) who considered an ellipsoidal region centred about a point of maximum interest. Snee and Marquardt (1974) obtained subsets of the extreme vertices which provide precise estimates of the parameters of a linear model. Snee (1975) used the computer to develop designs in constrained mixture spaces for the quadratic model. Saxena and Nigam (1977) explored the restricted mixture region using symmetric simplex design. Murthy and Murty (1983) discussed a method of construction of mixture designs for the exploration of the restricted region using factorials.

Much of the work on Latin squares has been done by various authors, for example, Bose (1938), Mann (1942) Parker (1959 a, b), Bose, Shrikhande and Parker (1960), Menon (1961) and Wallis (1984), who gave the methods of construction of mutually orthogonal Latin squares in various ways.

In this paper, we present an algorithm for constructing orthogonal arrays based mixture designs. The orthogonal arrays used in the proposed algorithm are constructed using a complete set of mutually orthogonal Latin squares. These orthogonal arrays have index unity and contain the smallest number of runs for a given number of levels and a given strength. This algorithm, therefore, leads to designs with small number of distinct runs.

We have examined and compared the designs constructed through this algorithm with the existing designs based on G-efficiency. The manageable number of design points help in reducing the cost and time in statistical experiments.

2. Orthogonal Arrays Based on Mutually Orthogonal Latin Squares

Hypercubes of strength *'d'* were defined by Rao (1946). Later, Rao (1947) extended the definition of hypercubes of strength d to cover a wider class of arrays called orthogonal arrays. An $N \times k$ array $\boldsymbol{A}$ with entries from S is said to be an orthogonal array $OA(N, k, s, t)$ with s levels, strength t and index λ (for some t in the range $0 \le t \le k$) if every $N \times t$ sub-array of $\boldsymbol{A}$ contains each t-tuple based on S exactly λ times as a row. If $\lambda = 1$, then such arrays are referred to as orthogonal arrays of index unity. (Bush 1952). Orthogonal arrays can be constructed using mutually orthogonal Latin squares.

A Latin square arrangement is an arrangement of s symbols in s rows and s columns, such that every symbol occurs once in each row and each column. When two Latin squares of same order are superimposed on one another, in the resultant array if every ordered pair of symbols occurs exactly once, then the two Latin squares are said to be orthogonal. A collection of ω Latin squares of order s, every pair of which is orthogonal, is called a set of mutually orthogonal Latin squares, and is denoted by $MOLS(s, \omega)$. Such a collection constitutes a complete set of mutually orthogonal Latin squares when $\omega = s - 1$.

2.1. Design criteria

If $\boldsymbol{X}$ denotes the $N \times k$ design matrix, then a useful criterion for evaluating the design is the minimum-maximum variance criterion. This refers to minimizing the maximum variance of prediction over the experimental region, where the prediction variance at the point $\boldsymbol{x}$ ($1 \times k$ row vector) is given by $\sigma^2 \upsilon$ and $\upsilon = \boldsymbol{x(X'X)^{-1}x'}$. Computation of the maximum variance provides a criterion of how close is a design to optimality. The G-efficiency or Global efficiency of the design is given by,

$$G\text{-}efficiency\ (percent\ of\ optimum) = 100p/Nd$$

where, p is the number of model parameters, N is the number of design points and d is the maximum value of υ over the experimental region.

Wheeler (1972) stated as a practical rule of thumb that any design with a $G\text{-}efficiency \geq 50\%$ could be called good for practical purposes.

3. OABMD Algorithm

Let s be a prime or a power of a prime, then there are $(s-1)$ mutually orthogonal Latin squares of order s. Superimpose these $(s-1)$ mutually orthogonal Latin squares on one another. Label the rows and columns of this array with $0, 1, \ldots, s-1$. Prefix labels of the rows and post fix the labels of the columns to the entries of the superimposed mutually orthogonal Latin squares. The elements of the resultant give an $OA(s^2, s+1, s, 2)$ with the maximum number of factors. We now present Orthogonal Array Based Mixture Design (OABMD) algorithm for constructing q component mixture designs using mutually orthogonal Latin squares.

Step 1: Construct an orthogonal array $OA(s^2, s+1, s, 2)$ using the set of $(s-1)$ mutually orthogonal Latin squares, with $q = (s+1)$ factors. Denote this matrix by $\boldsymbol{A}$.

Step 2: Create a matrix $\boldsymbol{M}$ of order $q \times q$ which is symmetrical but not orthogonal, having all elements as integers with sum of elements in each row and each column being zero. The choice of $\boldsymbol{M}$ is arbitrary and is useful in getting more vertices of the experimental region in the design as mentioned by Murthy and Murty (1983).

Step 3: Identify the minimum value in each column of $\boldsymbol{A} \times \boldsymbol{M}$ and subtract it from all the entries of that corresponding column to create a new matrix $\boldsymbol{T}$.

Step 4: Obtain the row totals for matrix $\boldsymbol{T}$. Divide the entries of each row of $\boldsymbol{T}$ by its corresponding row total to obtain a new matrix $\boldsymbol{Z}$. The resultant matrix is a mixture design satisfying

$$0 \leq z_i \leq 1 \text{ and } \sum_{i=1}^{q} z_i = 1,$$

z_i being the proportion for ith component.

Further, if the mixture experiment has to be performed in the restricted region, where each component is bounded by lower or upper bounds, or both, then proceed as follows:

Step 5: Rank the components in order of their increasing ranges $R_i = (U_i - L_i)$ such that X_1 has the smallest range and X_q has the largest range, assuming range to be inversely proportional to the importance (in terms of cost, effectiveness, *etc.*) of the components in the experiment.

Step 6: Using the transformation given by Saxena and Nigam (1977), $X_i = L_i + (R_i \times z_i)$, compute the entries for the first $(q-1)$ components of the design matrix $\boldsymbol{X}$. The levels of X_q are obtained by $X_q = 1 - \sum_{i=1}^{q-1} X_i$.

Step 7: In case X_q lies beyond the specified bounds, generate candidate design points. There may be multiple candidate points corresponding to a given design point. The candidate points are generated by adjusting the level of one of the components by a quantity equal to the difference between the substituted upper or lower bound and the computed value for X_q. Additional points are produced only from those components whose adjusted levels remain within the limits of the components.

We have illustrated the OABMD algorithm for generating designs for three, four and five components. These designs have been found to be efficient designs.

4. Mixture Designs For Three, Four And Five Components

4.1. Three component example

Consider the three-component mixture experiment, where all the components satisfy (1). Construct an orthogonal array $OA(4, 3, 2, 2)$ with three factors. Denote it by $\boldsymbol{A}$.

$$\boldsymbol{A}^{\boldsymbol{T}} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Multiply it with a symmetric and non-orthogonal matrix, $\boldsymbol{M}$,

$$\boldsymbol{M} = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

having row sums and column sums as zero, to obtain $\boldsymbol{T}$,

$$\boldsymbol{T} = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 0 & 0 \\ 0 & 0 & 3 \\ 0 & 3 & 0 \end{pmatrix}$$

Using step 4 of the OABMD algorithm, we obtain the design matrix $\boldsymbol{Z}$ for unrestricted region as follows:

$$\boldsymbol{Z}=\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{6}$$

We observe that $0 \le z_i \le 1$ and sum of the elements for each row is unity. This design has a G-efficiency of 81.8% for fitting (3).

Next, using step 6 of the OABMD algorithm, we compute the design matrix for the constrained region, as discussed below.

Example 1: Consider the following three component mixture experiment, discussed by Murthy and Murty (1983), in order of increasing ranges:

$$\begin{aligned} 0.2 \le X_1 \le 0.3 \\ 0.3 \le X_2 \le 0.5 \\ \text{and} \quad 0.2 \le X_3 \le 0.5 \end{aligned}$$

The first $(q-1)$ columns of the design matrix $\boldsymbol{X}$ are constructed using $X_i = L_i + (R_i \times z_i)$, where z_i is the proportion of the ith component of $\boldsymbol{Z}$ in (6). The levels of X_q are obtained by $X_q = 1 - \sum_{i=1}^{q-1} X_i$. The four design points of the resulting design matrix $\boldsymbol{X}$ are given in (7).

$$\boldsymbol{X}=\begin{pmatrix} 0.233 & 0.367 & 0.4 \\ 0.3 & 0.3 & 0.4 \\ 0.2 & 0.3 & 0.5 \\ 0.2 & 0.5 & 0.3 \end{pmatrix} \tag{7}$$

This design has a G-efficiency of 81.8% for fitting (3).

Other choices of $\boldsymbol{M}$ and resultant unconstrained and constrained design matrix are listed in Table 1.

Table 1: Unconstrained and constrained mixture designs corresponding to different choices of $\boldsymbol{M}$

Choice of $\boldsymbol{M}$	Resultant $\boldsymbol{Z}$	Resultant $\boldsymbol{X}$
$\boldsymbol{M_1}=\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$	$\boldsymbol{Z_1}=\begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 0 & 3/4 & 1/4 \\ 1/4 & 3/4 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}$	$\boldsymbol{X_1}=\begin{pmatrix} 0.225 & 0.4 & 0.375 \\ 0.2 & 0.45 & 0.35 \\ 0.225 & 0.45 & 0.325 \\ 0.25 & 0.3 & 0.45 \end{pmatrix}$
$\boldsymbol{M_2}=\begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix}$	$\boldsymbol{Z_2}=\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 \\ 0 & 1 & 0 \end{pmatrix}$	$\boldsymbol{X_2}=\begin{pmatrix} 0.233 & 0.367 & 0.4 \\ 0.267 & 0.3 & 0.433 \\ 0.233 & 0.3 & 0.467 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}$

Choice of $\boldsymbol{M}$	Resultant $\boldsymbol{Z}$	Resultant $\boldsymbol{X}$
$\boldsymbol{M_3}=\begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}$	$\boldsymbol{Z_3}=\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 2/3 & 1/3 & 0 \end{pmatrix}$	$\boldsymbol{X_3}=\begin{pmatrix} 0.233 & 0.367 & 0.4 \\ 0.2 & 0.433 & 0.367 \\ 0.233 & 0.3 & 0.467 \\ 0.267 & 0.367 & 0.367 \end{pmatrix}$
$\boldsymbol{M_4}=\begin{pmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 1 \end{pmatrix}$	$\boldsymbol{Z_4}=\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \\ 0 & 1/3 & 2/3 \end{pmatrix}$	$\boldsymbol{X_4}=\begin{pmatrix} 0.233 & 0.367 & 0.4 \\ 0.267 & 0.3 & 0.433 \\ 0.233 & 0.433 & 0.333 \\ 0.2 & 0.367 & 0.433 \end{pmatrix}$
$\boldsymbol{M_5}=\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$	$\boldsymbol{Z_5}=\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}$	$\boldsymbol{X_5}=\begin{pmatrix} 0.233 & 0.367 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.25 & 0.4 & 0.35 \\ 0.25 & 0.3 & 0.45 \end{pmatrix}$

The unconstrained and constrained design matrices obtained using different choices of $\boldsymbol{M}$ listed in the table above also yield a G-efficiency of 81.8% for fitting (3).

Example 2: Consider the following three component mixture experiment, discussed by Snee and Marquardt (1974) and by Saxena and Nigam (1977), in order of increasing ranges:

$$0.1 \le X_1 \le 0.6$$
$$0.1 \le X_2 \le 0.7$$
$$\text{and} \quad 0 \le X_3 \le 0.7$$

Using $\boldsymbol{M}$,

$$\boldsymbol{M}=\begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}$$

the four design points of the resulting design matrix $\boldsymbol{X}$ are given in (8).

$$\boldsymbol{X}=\begin{pmatrix} 0.267 & 0.3 & 0.433 \\ 0.6 & 0.1 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.1 & 0.7 & 0.2 \end{pmatrix} \tag{8}$$

We observe that the limits for the third component of the third design point of (8) lies outside the specified bounds, so we shall adjust the run (0.1, 0.1, 0.8), using step 7 of the OABMD algorithm, to create two candidate sub-points (0.1, 0.2, 0.7) and (0.2, 0.1, 0.7).

Case (a): Four-point design with candidate point (0.1, 0.2, 0.7)

The design matrix in (8) is modified to incorporate the candidate sub-point (0.1, 0.2, 0.7) to yield the following four design runs.

$$\boldsymbol{X_a}=\begin{pmatrix}0.267 & 0.3 & 0.433\\ 0.6 & 0.1 & 0.3\\ 0.1 & 0.2 & 0.7\\ 0.1 & 0.7 & 0.2\end{pmatrix}$$

The G-efficiency for the above design matrix for fitting Scheffé's linear model given in (3) is 79.19%.

Case (b): Four-point design with candidate point (0.2, 0.1, 0.7)

The design matrix in (8) is adjusted to include the candidate sub-point (0.2, 0.1, 0.7) to give the following four design runs.

$$\boldsymbol{X_b}=\begin{pmatrix}0.267 & 0.3 & 0.433\\ 0.6 & 0.1 & 0.3\\ 0.2 & 0.1 & 0.7\\ 0.1 & 0.7 & 0.2\end{pmatrix}$$

The G-efficiency for the above design matrix for fitting Scheffé's linear model given in (3) is 78.65%.

4.2. Four component example

To construct a mixture design in four components satisfying (1), construct an orthogonal array with four factors, say, $\boldsymbol{A} = OA(9, 4, 3, 2)$ and a symmetric and non-orthogonal matrix, $\boldsymbol{M}$ as

$$\boldsymbol{A^T}=\begin{pmatrix}0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2\\ 0 & 1 & 2 & 1 & 2 & 0 & 2 & 0 & 1\\ 0 & 1 & 2 & 2 & 0 & 1 & 1 & 2 & 0\\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2\end{pmatrix} \text{ and } \boldsymbol{M}=\begin{pmatrix}-3 & 1 & 1 & 1\\ 1 & -3 & 1 & 1\\ 1 & 1 & -3 & 1\\ 1 & 1 & 1 & -3\end{pmatrix}$$

The unconstrained mixture design $\boldsymbol{Z}$, is

$$\boldsymbol{Z}=\begin{pmatrix}0.2 & 0.27 & 0.27 & 0.26\\ 0.4 & 0.2 & 0.2 & 0.2\\ 0.6 & 0.13 & 0.13 & 0.14\\ 0.2 & 0.27 & 0 & 0.53\\ 0.2 & 0 & 0.53 & 0.27\\ 0.2 & 0.53 & 0.27 & 0\\ 0 & 0.07 & 0.33 & 0.6\\ 0 & 0.6 & 0.07 & 0.33\\ 0 & 0.33 & 0.6 & 0.07\end{pmatrix} \tag{9}$$

Consider the constrained four component mixture experiment as discussed by McLean and Anderson (1966). In manufacturing one particular type of flare the chemical constituents are magnesium (x_1), sodium nitrate (x_2), strontium nitrate (x_3), and binder (x_4). Engineering experience has indicated that the following constraints (in order of increasing ranges) on a proportion by weight basis should be utilized:

$$\begin{gathered} 0.03 \le X_1 \le 0.08 \\ 0.40 \le X_2 \le 0.60 \\ 0.10 \le X_3 \le 0.50 \\ \text{and } 0.10 \le X_4 \le 0.50 \end{gathered}$$

Using the OABMD algorithm, the constrained mixture design, $\boldsymbol{X}$, is

$$\boldsymbol{X} = \begin{pmatrix} 0.04 & 0.453 & 0.207 & 0.3 \\ 0.05 & 0.44 & 0.18 & 0.33 \\ 0.06 & 0.427 & 0.153 & 0.36 \\ 0.04 & 0.453 & 0.1 & 0.407 \\ 0.04 & 0.4 & 0.313 & 0.247 \\ 0.04 & 0.506 & 0.207 & 0.247 \\ 0.03 & 0.413 & 0.233 & 0.324 \\ 0.03 & 0.52 & 0.127 & 0.323 \\ 0.03 & 0.467 & 0.34 & 0.163 \end{pmatrix} \tag{10}$$

The design matrix $\boldsymbol{Z}$ in (9) as well as matrix $\boldsymbol{X}$ in (10) have a G-efficiency of 72.72% for fitting (3). The design matrix achieved using OABMD algorithm proves to be more economical in terms of number of design points and exploration of the constrained region than the XVERT algorithm.

If limitation of resources demands a reduction in the number of design points, we may use the method of normalization, as discussed by McLean and Anderson (1966). The operative idea is to compute a normalized distance d_{ij} between points of the design and randomly omit points that are less than a certain minimum distance from other design points.

$$d_{ij} = \left(\sum_{r=1}^{q} \left(\frac{x_{ir} - x_{jr}}{b_r - a_r} \right)^2 \right)^{\frac{1}{2}}$$

The design matrices, $\boldsymbol{Z_{1*}}$ and $\boldsymbol{X_{1*}}$, corresponding to the unrestricted and restricted region, consisting of just four design points, obtained using the above-mentioned technique are:

$$\boldsymbol{Z_{1*}} = \begin{pmatrix} 0.6 & 0.13 & 0.13 & 0.13 \\ 0.2 & 0.27 & 0 & 0.53 \\ 0 & 0.07 & 0.33 & 0.6 \\ 0 & 0.6 & 0.07 & 0.33 \end{pmatrix} \tag{11}$$

and

$$\boldsymbol{X_{1*}} = \begin{pmatrix} 0.06 & 0.427 & 0.153 & 0.36 \\ 0.04 & 0.453 & 0.1 & 0.407 \\ 0.04 & 0.4 & 0.313 & 0.247 \\ 0.03 & 0.52 & 0.127 & 0.323 \end{pmatrix} \tag{12}$$

The design matrix $\boldsymbol{Z_{1*}}$ in (11) as well as matrix $\boldsymbol{X_{1*}}$ in (12) have a G-efficiency of 100% for fitting (3). To allow estimation of error variance, we may add another design point, (0, 0.33, 0.6, 0.07) to the design matrix $\boldsymbol{Z_{1*}}$in (11), then the resultant design consisting of five runs has a G-efficiency of 84.89% for fitting (3). Similarly, adding one point of the restricted region, say, (0.03, 0.467, 0.34, 0.163) to the design matrix $\boldsymbol{X_{1*}}$ in (12) yields a G-efficiency of 80% for fitting (3).

In practical situations, fitting linear model is not always suitable. A higher model may provide a better fit to the given design. Consider the design matrices, $\boldsymbol{Z}$ and $\boldsymbol{X}$, given in (6) and (7). Further, in addition to these points, we may add the boundary points, centroids or the extreme vertices to facilitate the computation of G-efficiency for fitting higher order models as stated in (4) and (5). Addition of three centroid points to the unconstrained design matrix $\boldsymbol{Z}$ of (6) and corresponding design points in $\boldsymbol{X}$ of (7) yield two 7-point designs, both of which have a G-efficiency of 85.7% for fitting Scheffé quadratic model specified in (4). Similarly, adding one more point to the 7-point design of $\boldsymbol{Z}$ and $\boldsymbol{X}$ gives an 8-point design, both of which have a G-efficiency of 87.5% for fitting Scheffé cubic model mentioned in (5). Likewise, adding boundary points to the unconstrained and constrained design matrices, for any choice of $\boldsymbol{M}$, as stated in Table 1, yield a G-efficiency of 85.7% and 87.5% for fitting (4) and (5) respectively.

Similarly, for the four-component example, 11-point designs obtained by adding two boundary points, (0, 1, 0, 0) and (0, 0, 1, 0) to (9) and adding the points (0.03, 0.4, 0.1, 0.47) and (0.08, 0.6, 0.1, 0.22) to (10) yield a G-efficiency of 90.90% for fitting (4). This value of G-efficiency is computed using $\boldsymbol{X_E}$ in place of $\boldsymbol{X}$, where $\boldsymbol{X_E}$ given below is the extended design matrix for model (4).

$$\boldsymbol{X_E} = \begin{pmatrix}
0.04 & 0.453 & 0.207 & 0.3 & 0.018 & 0.008 & 0.012 & 0.093 & 0.136 & 0.062 \\
0.05 & 0.44 & 0.18 & 0.33 & 0.022 & 0.009 & 0.016 & 0.079 & 0.145 & 0.059 \\
0.06 & 0.427 & 0.153 & 0.36 & 0.025 & 0.009 & 0.021 & 0.065 & 0.153 & 0.055 \\
0.04 & 0.453 & 0.1 & 0.407 & 0.018 & 0.004 & 0.016 & 0.045 & 0.184 & 0.041 \\
0.04 & 0.4 & 0.313 & 0.247 & 0.016 & 0.012 & 0.009 & 0.125 & 0.098 & 0.077 \\
0.04 & 0.506 & 0.207 & 0.247 & 0.020 & 0.008 & 0.009 & 0.104 & 0.125 & 0.051 \\
0.03 & 0.413 & 0.233 & 0.324 & 0.012 & 0.007 & 0.009 & 0.096 & 0.134 & 0.075 \\
0.03 & 0.52 & 0.127 & 0.323 & 0.015 & 0.003 & 0.009 & 0.066 & 0.168 & 0.041 \\
0.03 & 0.467 & 0.34 & 0.163 & 0.014 & 0.010 & 0.005 & 0.158 & 0.076 & 0.055 \\
0.03 & 0.4 & 0.1 & 0.47 & 0.012 & 0.003 & 0.014 & 0.04 & 0.188 & 0.047 \\
0.08 & 0.6 & 0.1 & 0.22 & 0.048 & 0.008 & 0.017 & 0.06 & 0.132 & 0.022
\end{pmatrix}$$

The G-efficiency values of $\boldsymbol{Z}$ and $\boldsymbol{X}$ for fitting (3) and (4) for different choices of $\boldsymbol{M}$ are listed in Table 2.

Table 2: G-efficiency of unconstrained and constrained mixture designs corresponding to different choices of *M*

Choice of ***M***	G-efficiency for fitting (3)		G-efficiency for fitting (3)		G-efficiency for fitting (4)	
	9-point design		5-point design		11-point design	
	Z	***X***	***Z***	***X***	***Z***	***X***
$\boldsymbol{M_1}=\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$	72.72%	72.72%	80%	80%	90.90%	90.90%
$\boldsymbol{M_2} = -\boldsymbol{M_1}$	72.72%	72.72%	80%	80%	90.90%	90.90%
$\boldsymbol{M_3}=\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$	72.72%	72.72%	84.89%	80%	90.90%	90.90%
$\boldsymbol{M_4} = -\boldsymbol{M_3}$	72.72%	72.72%	84.89%	80%	90.90%	90.90%
$\boldsymbol{M_5} = -\boldsymbol{M}$	72.72%	72.72%	84.89%	80%	90.90%	90.90%

4.3. Five component example

We may extend the application of our OABMD algorithm to five component constraints. To construct a mixture design, ***Z***, in five components satisfying (1), construct an orthogonal array with five factors, ***A*** = *OA*(16, 5, 4, 2).

$$\boldsymbol{A^T} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 0 & 1 & 2 & 3 & 1 & 0 & 3 & 2 & 2 & 3 & 0 & 1 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 2 & 3 & 0 & 1 & 3 & 2 & 1 & 0 & 1 & 0 & 3 & 2 \\ 0 & 1 & 2 & 3 & 3 & 2 & 1 & 0 & 1 & 0 & 3 & 2 & 2 & 3 & 0 & 1 \\ 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \end{pmatrix}$$

and a symmetric and non-orthogonal matrix, ***M*** as

$$\boldsymbol{M} = \begin{pmatrix} -4 & 1 & 1 & 1 & 1 \\ 1 & -4 & 1 & 1 & 1 \\ 1 & 1 & -4 & 1 & 1 \\ 1 & 1 & 1 & -4 & 1 \\ 1 & 1 & 1 & 1 & -4 \end{pmatrix}$$

to give the unconstrained mixture design ***Z***,

$$\boldsymbol{Z} = \begin{pmatrix} 0.158 & 0.210 & 0.210 & 0.211 & 0.211 \\ 0.264 & 0.184 & 0.184 & 0.184 & 0.184 \\ 0.368 & 0.158 & 0.158 & 0.158 & 0.158 \\ 0.474 & 0.132 & 0.132 & 0.131 & 0.131 \\ 0.211 & 0.263 & 0.131 & 0 & 0.395 \\ 0.211 & 0.395 & 0 & 0.131 & 0.263 \\ 0.211 & 0 & 0.395 & 0.263 & 0.131 \\ 0.211 & 0.131 & 0.263 & 0.395 & 0 \\ 0.105 & 0.158 & 0.026 & 0.289 & 0.422 \\ 0.105 & 0.026 & 0.158 & 0.422 & 0.289 \\ 0.105 & 0.422 & 0.289 & 0.026 & 0.158 \\ 0.105 & 0.289 & 0.422 & 0.158 & 0.026 \\ 0 & 0.053 & 0.316 & 0.184 & 0.447 \\ 0 & 0.184 & 0.447 & 0.053 & 0.316 \\ 0 & 0.316 & 0.053 & 0.447 & 0.184 \\ 0 & 0.447 & 0.184 & 0.316 & 0.053 \end{pmatrix} \tag{13}$$

Using step 6 of the OABMD algorithm, we compute the design matrix of the constrained region. Consider the five-component example, discussed by Snee and Marquardt (1974). The gasoline blending model for a five-component system, namely, Butane (X_1), Alkylate (X_2), Lt. St. Run (X_3), Reformate (X_4) and Cat Cracked (X_5), with the following component ranges:

$$\begin{gathered} 0.00 \le X_1 \le 0.10 \\ 0.00 \le X_2 \le 0.10 \\ 0.05 \le X_3 \le 0.15 \\ 0.20 \le X_4 \le 0.40 \\ \text{and } 0.40 \le X_5 \le 0.60 \end{gathered}$$

Using the OABMD algorithm, the design matrix, $\boldsymbol{X}$ is as follows:

$$\boldsymbol{X} = \begin{pmatrix} 0.0158 & 0.0211 & 0.0710 & 0.2421 & 0.6500 \\ 0.0263 & 0.0184 & 0.0684 & 0.2369 & 0.6500 \\ 0.0368 & 0.0158 & 0.0658 & 0.2316 & 0.6500 \\ 0.0473 & 0.0132 & 0.0632 & 0.2263 & 0.6500 \\ 0.0211 & 0.0263 & 0.0632 & 0.2000 & 0.6894 \\ 0.0211 & 0.0394 & 0.0500 & 0.2263 & 0.6632 \\ 0.0211 & 0.0000 & 0.0895 & 0.2526 & 0.6368 \\ 0.0211 & 0.0132 & 0.0763 & 0.2789 & 0.6105 \\ 0.0105 & 0.0158 & 0.0526 & 0.2579 & 0.6632 \\ 0.0105 & 0.0026 & 0.0658 & 0.2842 & 0.6369 \\ 0.0105 & 0.0421 & 0.0789 & 0.2053 & 0.6632 \\ 0.0105 & 0.0289 & 0.0921 & 0.2316 & 0.6369 \\ 0.0000 & 0.0053 & 0.0816 & 0.2368 & 0.6763 \\ 0.0000 & 0.0184 & 0.0948 & 0.2105 & 0.6763 \\ 0.0000 & 0.0316 & 0.0552 & 0.2895 & 0.6237 \\ 0.0000 & 0.0447 & 0.0684 & 0.2632 & 0.6237 \end{pmatrix} \tag{14}$$

The design matrix $\boldsymbol{X}$ in (14) has many points, particularly of the fifth component, which lie beyond the specified limits of the component. Using step 7 of the OABMD algorithm, we

adjust the matrix $\boldsymbol{X}$ in (14) to obtain the design matrix $\boldsymbol{X}^*$, given in (15), which has all the design points within the permissible limits of the components involved in the five-component example.

$$\boldsymbol{X}^* = \begin{pmatrix} 0.0158 & 0.0211 & 0.0710 & 0.2921 & 0.6000 \\ 0.0263 & 0.0184 & 0.0684 & 0.2869 & 0.6000 \\ 0.0368 & 0.0158 & 0.0658 & 0.2816 & 0.6000 \\ 0.0473 & 0.0132 & 0.0632 & 0.2763 & 0.6000 \\ 0.0211 & 0.0263 & 0.0632 & 0.2894 & 0.6000 \\ 0.0211 & 0.0394 & 0.0500 & 0.2895 & 0.6000 \\ 0.0211 & 0.0000 & 0.0895 & 0.2894 & 0.6000 \\ 0.0211 & 0.0132 & 0.0763 & 0.2894 & 0.6000 \\ 0.0105 & 0.0158 & 0.0526 & 0.3211 & 0.6000 \\ 0.0105 & 0.0026 & 0.0658 & 0.3211 & 0.6000 \\ 0.0105 & 0.0421 & 0.0789 & 0.2685 & 0.6000 \\ 0.0105 & 0.0289 & 0.0921 & 0.2685 & 0.6000 \\ 0.0000 & 0.0053 & 0.0816 & 0.3131 & 0.6000 \\ 0.0000 & 0.0184 & 0.0948 & 0.2868 & 0.6000 \\ 0.0000 & 0.0316 & 0.0552 & 0.3132 & 0.6000 \\ 0.0000 & 0.0447 & 0.0684 & 0.2869 & 0.6000 \end{pmatrix} \tag{15}$$

The design matrix $\boldsymbol{X}^*$ in (15), computed using the OABMD algorithm, has a G-efficiency of 58.10% for fitting (3). The design matrix $\boldsymbol{X}^*$ in (15) is space filling and allows for the greater exploration of the interior of the restricted region in contrast to the only extreme vertices generated by the XVERT algorithm.

Other choices of $\boldsymbol{M}$ with corresponding G-efficiency values of $\boldsymbol{Z}$ and $\boldsymbol{X}$ for fitting (3) and (4) are listed in Table 3.

Table 3: G-efficiency of unconstrained and constrained mixture designs corresponding to different choices of $\boldsymbol{M}$

Choice of $\boldsymbol{M}$	G-efficiency for fitting (3)		G-efficiency for fitting (4)	
	$\boldsymbol{Z}$	$\boldsymbol{X}^*$	$\boldsymbol{Z}$	$\boldsymbol{X}^*$
$\boldsymbol{M_1} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$	73.96%	88.57%	93.75%	92.20%
$\boldsymbol{M_2} = -\boldsymbol{M_1}$	73.96%	74.55%	93.75%	39.70%*
$\boldsymbol{M_3} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$	73.96%	76.34%	93.75%	87.93%
$\boldsymbol{M_4} = -\boldsymbol{M_3}$	73.96%	72.88%	93.75%	49.98%**

Choice of $\boldsymbol{M}$	G-efficiency for fitting (3)		G-efficiency for fitting (4)	
	$\boldsymbol{Z}$	$\boldsymbol{X}^*$	$\boldsymbol{Z}$	$\boldsymbol{X}^*$
$\boldsymbol{M_5} = -\boldsymbol{M}$	73.96%	70.42%	93.75%	94.26%
$\boldsymbol{M_6}=\begin{pmatrix} 0 & 1 & -1 & -1 & 1 \\ 1 & 0 & 1 & -1 & -1 \\ -1 & 1 & 0 & 1 & -1 \\ -1 & -1 & 1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 \end{pmatrix}$	73.96%	81.76%	93.75%	81.39%
$\boldsymbol{M_7} = -\boldsymbol{M_6}$	73.96%	80.64%	93.75%	86.95%

*indicates non G-efficient designs as per the thumb rule stated by Wheeler (1972) and should not be used for fitting the quadratic model

**the value is almost equal to 50% and hence the design can be used for practical purposes, as suggested by Wheeler (1972)

5. Conclusion

The orthogonal arrays with index unity have been considered in our proposed OABMD algorithm. The designs, hence, constructed have the smallest number of runs for a given number of levels, thereby allowing higher cost efficiency. Furthermore, the flexibility in choice of matrix $\boldsymbol{M}$ allows for enhanced variety of design points. The manageable number of distinct design points help in reducing the cost and time in statistical experiments.

When the region of interest is pre-defined, the proposed OABMD algorithm can be customized to explore the restricted space. The constructed designs have a sufficiently high G-efficiency that make them suitable for practical purposes.

Acknowledgements

The authors are thankful to the editor and the referees for their valuable suggestions that helped in the improvement of the paper.

References

Bose, R. C. (1938). On application of the properties of Galois Fields to the problem of construction of Hyper-Graeco-Latin squares. *Sankhya*, **3**, 323-338.

Bose, R. C., Shrikhande, S. S., and Parker, E. T. (1960). Further results on the construction of mutually orthogonal Lain squares and the falsity of Euler's conjecture. *Canadian Journal of Mathematics*, **12,** 189-203.

Buruk, S. Y., Aktar, D. E., and Burnak, N. (2016). Mixture Design: A review of recent applications in the food industry. *Pamukkale Univ Muh Bilim Derg*, **22**(**4**), 297-304.

Bush, K. A. (1952). Orthogonal arrays of index unity. *The Annals of Mathematical Statistics*, **23**, 426-434.

Cafaggi, S., Leardi, R., Parodi, B., Caviglioli, G., and Bignardi, G. (2003). An example of application of a mixture design with constraints to a pharmaceutical formulation. *Chemometrics and Intelligent Laboratory Systems*, **65**(**1**), 139-147.

Mann, H. B. (1942). The construction of orthogonal Latin squares. *The Annals of Mathematical Statistics*, **13**, 418-423.

McLean, R. A., and Anderson, V. L. (1966). Extreme vertices design of mixture experiments. *Technometrics*, **8**(**3**), 447 454.

Menon, P. K. (1961). Method of constructing two mutually orthogonal Latin squares of order 3*n*+1, *Sankhya*, **A23**, 281-282.

Mirabedini, S. M., Jamali, S. S., Hagheyegh, M., Sharifi, M., Mirabedini, A. S. and Hashemi-Nasab, R. (2012). Application of mixture experimental design to optimize formulation and performance of thermoplastic road markings. *Progress in Organic Coatings*, **75**(**4**), 549-559.

Murthy, M. S. R., and Murty, J. S. (1983). Restricted region simplex design for mixture experiments. *Communication in Statistics- Theory and Methods*, **12**(**22**), 2605-2615.

Parker, E. T. (1959a). Orthogonal Latin squares. *Proceedings of the National Academy of Sciences of the United States of America*, **45**, 859-862.

Parker, E. T. (1959b). Construction of some sets of mutually orthogonal Latin squares. *Proceedings of the American Mathematical Society*, **10**, 946-949.

Rao, C. R. (1946). Hypercubes of strength "*d*" leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society*, **38**, 67-78.

Rao, C. R. (1947). Factorial Experiments derivable from combinatorial arrangements of arrays. *Journal of Royal Statistical Society* (*Supplement*), **9**(**1**), 128-139.

Saxena, S. K., and Nigam, A. K. (1977). Restricted exploration of mixtures by symmetric simplex design. *Technometrics*, **19**(**1**), 47-52.

Scheffé, H. (1958). Experiments with Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, **20**(**2**), 344-360.

Scheffé, H. (1963). The simplex centroid design for experiments with mixtures. *Journal of the Royal Statistical Society, Series B (Methodological)*, **25**(**2**), 235-263.

Schrevens, E. and Cornell, J. (1993). Design and analysis of mixture systems: Applications in hydroponic, plant nutrition research. In: Fragoso, M.A.C., Van Beusichem, M.L., Houwers, A. (Eds) *Optimization of Plant Nutrition. Developments in Plant and Soil Sciences*, Vol. **53**. Springer, Dordrecht (Online ISBN: 978-94-017-2496-8).

Snee, R. D. (1975). Experimental designs for quadratic models in constrained mixture spaces, *Technometrics*, **17(2)**, 149-159.

Snee, R. D., and Marquardt, D. W. (1974). Extreme vertices designs for linear mixture models. *Technometrics*, **16**(**3**), 399-408.

Thompson, W. C., and Myers, R. H. (1968). Response surface designs for experiments with mixtures. *Technometrics*, **10**(**4**), 739-756.

Wallis W. D. (1984). Three orthogonal Latin squares. *Congressus Numerantium*, **42**, 69-86.

Wheeler, R. E. (1972). Efficient experimental design. Presented at the Annual Meeting of the *American Statistical Association*, Montreal, Canada.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 27-39

Factors Affecting on the Usage Pattern of Cloud Storage in Vadodara City of Gujarat

Khimya Tinani, Bhargav Choithwani, Bhagyashree Patil,
Pathan Faiyazkhan and Tanvi Salat
Department of Statistics, Faculty of Science,
The Maharaja Sayajirao University of Baroda, Vadodara 390002, India

Received: 27 March 2020; Revised: 01 July 2020; Accepted: 13 July 2020

Abstract

Cloud Computing is the trending topic in the field of research as well as in commercial environment. Currently there has been a focus on cloud storage and cloud security. Cloud computing model facilitates data backup, minimised cost, data centralization, free storage, data sharing capabilities and many more which leads to company's high performance. The purpose of this paper is to study the proportion of people's awareness on cloud storage, investigate the factors contributing to the usage of cloud storage, predicting the future scope of cloud storage usage using Multinomial Logistic Regression and to provide an overall security perspective of cloud computing and highlight the security concerns needed to be addressed. A survey based on questionnaire was used to gather data from determined sample size and some relevant hypothesis were derived and tested. Chi-square test for Multiple Response Analysis to study the data was done. Statistical Package for Social Sciences (SPSS) software and R- Programming was used for statistical analysis.

Key words: Cloud storage; Cloud security; Multinomial logistic regression; Multiple response analysis.

1. Introduction

In an era of information technology and globalisation, the great computing power is expected to develop business insights and competitive benefit. A customary way for enterprises to process their data is to use the computing power provided by their own in-house data centres. However, intervening a private data centre to keep up with rapidly growing data processing requests can be complicated and costly. Cloud computing vanishes the complexity of IT infrastructure to physically store data and manage huge data which is time consuming and demands huge man power. The challenge nowadays is to acquire, store and manage the data generating at every second. Cloud storage is the service model which not only maintains the data but also manages the data by storing the data on the internet via the cloud computing provider that operates and manages data storage as a service. It also provides anytime and anywhere data access facility. It assures to provide on-demand computing power with quick implementation, low maintenance, fewer IT staff, and consequently lower cost. Among the various definitions, the one by the National Institute of Standards and Technology (NIST) has gained recent recognition and popularity. Yang *et al.* (2012) define cloud storage is a model for enabling convenient, on-demand network access to

Corresponding Author: Khimya Tinani
E-mail: khimya27@yahoo.com

a shared pool of configurable computing resources (*e.g.*, networks, servers, applications, storage, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It also suggested that a cloud computing model should be composed of five vital characteristics, three service levels, and four deployment models. Those five essential characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. The service levels are software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS). And the deployment models are public cloud, private cloud, hybrid cloud and community cloud.

In this paper, from the questionnaire-based survey, the proportion of the people's awareness about cloud storage was found using proportion test and multinomial logistic regression was used to predict whether the user would go for paid cloud storage or not.

2. Review of Literature

Mohd Rahul *et al.* (2012) defines cloud computing as a concept where files and applications are hosted on a "cloud" consisting of thousands of computers and servers, all linked together and accessible via the Internet. Further he says a cloud is a pool of virtualized computer resources. Cloud Computing is the long-held dream of computing as a utility, has the embryonic to transform a large part of the IT industry, making software even more attractive as a service and framing the way for designing IT hardware. Cloud Computing can be considered as a pay and use service. More you Pay, more services you get. It has become the 5th essential utility (after water, electricity, gas, and telephone) with its growing features and demands the day is not far when the world would become the slave of cloud computing. Tinani *et al.* (2019) observes that cloud computing provides a surrounding for resource sharing in terms of ascendance frameworks, middleware's and application development platforms, and business applications. The operation models of cloud computing grasp free infrastructure services with value another platform services, subscription-based infrastructure services with supplemental application services, and free services for sellers. Pankaj Sareen (2013) concluded cloud computing infrastructures are next generation platforms that can provide tremendous value to companies of any size. Cloud computing provides platform, software, infrastructure, storage, security, data, test environment etc. as a service. Clients or customer would be able to access their applications and data from anywhere at any time. Data wouldn't be enclosing to a hard drive on one user's computer or even a corporation's internal network. It would also lead hardware costs down. You would not need a huge hard drive because you would store all your information on a remote computer. However, the substantial concerns about cloud computing are security and privacy. Since the idea of handling over significant data to another company worries some people. With the rapid development of cloud storage technology and applications, it has broken through the single point of the original IaaS (Infrastructure as a Service) layer definition. It establishes a technical system, contains three layers of cloud computing infrastructure services (IaaS, PaaS, SaaS). At present, cloud computing services are mainly concentrated in the IaaS and SaaS layers. The contents of IaaS and SaaS are not the same. From the viewpoint of IaaS, cloud storage provides a service for data storage, archiving, and backup. From the SaaS point of view, cloud storage service is very diverse, the service has online backup document notes save network disk business photo preservation and sharing home video. Cloud storage is the small layer of cloud computing system which supports the service of the other layers above it. In addition, it is an important way to save and manage large data. So, it occupied even more attentions from some researchers. Cloud storage is catering for pervasive storage requirements and massive storage wishes. The persistency and operation are supported by the

storage systems. Programs, texts, data, pictures, videos, etc. are all needed to accumulate in the storage systems. Mobile terminals, PC, consume electronics such as smart phones, cameras; MP3/MP4, etc. are all need more and more storage resources. Usually, local storage is not sufficient to store, and lost easily. Therefore, the necessary and pervasiveness of storage requirements results to that the properties of cloud storage must be low cost, easy maintenance, reliable, security, recoverable, *etc*. Spoorthy *et al.* (2014) defined importance of cloud storage as no need to invest any capital on storage devices, no need for technical expert to maintain the storage, backup, replication and importantly disaster management, grant others to access your data will result with collaborative working style instead of individual work. Aized Amin Soofi (2014) discussed about data security which plays major concern for the users who want to use cloud computing. Proper security measures are expected to be taken in order to eliminate this concern. It was found that many of the cloud services users have concerns about their private data that it may be used for other purposes or sent to other cloud service providers.

3. Problem Statement

This research study is aimed to identify the various factors which affect the usage of public cloud storage. The research is further aimed to find the proportion among users of cloud storage in different area, to find the association among those factors, finding most influencing factors among all those factors, developing model whether the user will use paid version or not and potential risk for the user.

4. Objectives

The objective of this paper is to gain insight about the usage patterns of public cloud storage by the people of Vadodara city. To know the facilities which are mostly used by people gender wise, age wise, profession wise.

a) To check whether the proportion of cloud storage users is same in 6 wards of Vadodara city.
b) To check whether any association between specific cloud storage facilities usage on cloud company.
c) To check whether use of different cloud storage facilities depends on profession or not.
d) To predict whether the user will use paid version cloud storage or not.
e) To gain insight whether users are aware of the associated potential risk or not with respect to gender wise and profession.

5. Research Methodology

The research is done by the methods of survey, based on the questionnaire which gives the quantitative and qualitative data such as age, gender, education, profession of users, the cloud storage facilities they use, the cloud service provider they prefer and other related questions. The review held with explicit sampling technique and sample size determination technique. Sample Size determination is technique used to choose the number of observations to include in a statistical sample study. It is significant for any statistical study to make inference about the population based on a sample. Practically sample size is usually determined based on the cost, time or method of collecting the data and the need for it to

offer sufficient statistical power. For our research study the sample size is dependent on the pilot survey results. In the pilot survey study, the sample of size 30 was selected, from the pilot survey study the proportion of cloud storage users were 0.7 and of non-users was 0.3. The sample size is estimated from the pilot survey by proportion method where p = proportion of users = 0.7 and q = proportion of non-users taking margin of error e = 0.05, thus sample size of the main survey study was 364. For this study the primary data was collected and investigated using the questionnaire. This research survey was conducted in 6 clusters which were selected randomly in Vadodara city *i.e.*, Raopura, Karelibaug, Waghodiya, Akota, Harni and Alkapuri. Two stage sampling technique was used for collecting the data. The population of Vadodara city was divided into 13 clusters based on wards. In the first stage cluster sampling and in the second stage probability proportion to size (PPS) sampling method were used. To draw a sample with PPS, cumulative total method was used to determine the number of samples selected from each ward. Thus, as a result we got the sample size 29 out of 43555 population from Raopura, the sample size 62 out of 98723 population from Karelibaug, the sample size 72 out of 236097 population from Waghodiya, the sample size 41 out of 130715 population from Akota, the sample size 89 out of 122741 population from Harni and the sample size 71 out of 177287 population from Alkapuri.

6. Data Visualization

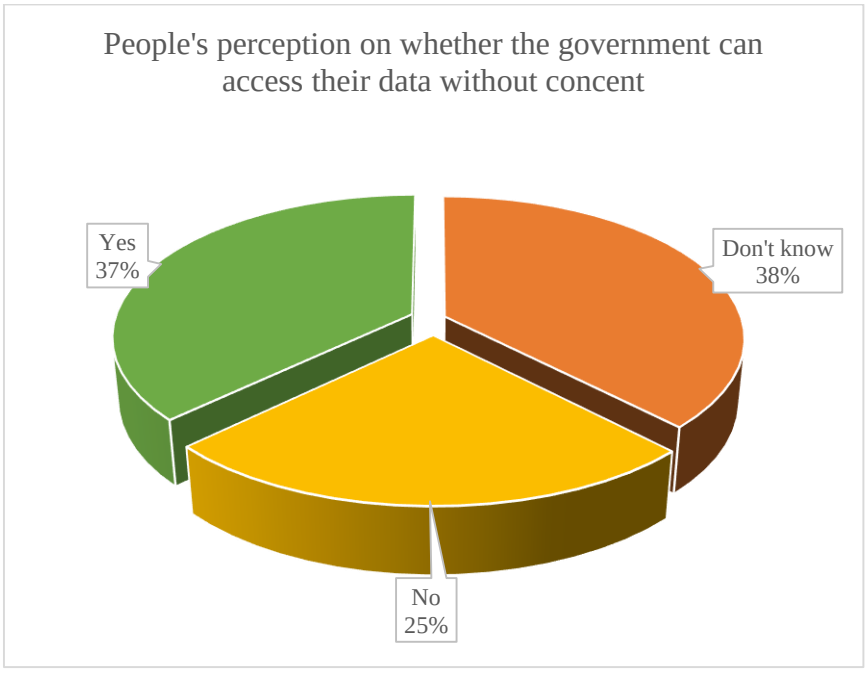


Figure 1: People's perception on cloud storage data access by government without their consent

From the above graph we observe that 38% of people don't know whether government can access their cloud storage data without their consent or not. Further 37% of people says government access their data without their consent.

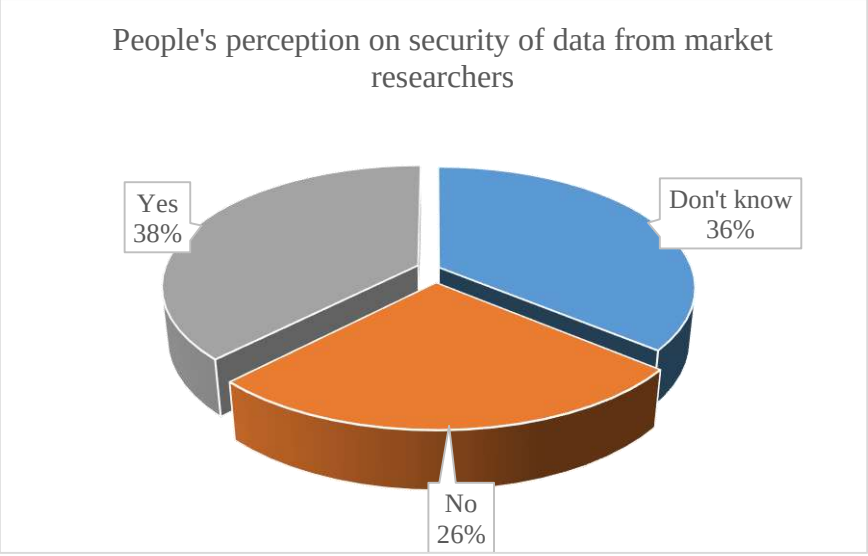


Figure 2: People's perception on security of their cloud storage data from market researchers

From Figure 2 we observe that 38% of people in our study says our cloud storage data is secure from market researchers and further follows that 36% of people don't know about it.

7. Results and Discussion

7.1. Proportion test Analysis

a) Objective: To check whether the proportion of cloud storage users is same in 6 wards of Vadodara city

In our research study, based on sample size we consider 6 areas of Vadodara city that is Raopura, Karelibaug, Waghodiya, Akota, Harni, and Alkapuri. There is total 178 users out of 364 total sample size. We observe that there are 11 users in Raopura out of 29 samples, 37 users in Karelibaug out of 62 samples, 27 users in Waghodiya out of 72 samples, 26 users in Akota out of 41 samples, 42 users in Harni out of 89 samples and 35 users in Alkapuri out of 71 samples. Hence to check equality of users proportion of 6 wards of Vadodara city we use proportion test in R programming by the function prop.test ().

H_0: $P_1 = P_2 = P_3 = P_4 = P_5 = P_6$
H_1: at least one may differ significantly.

```
#Proportion test is done using R programming.
#List of cloud storage users from each of 6 wards.
users <- c (11,37,27,26,42,35)

#List of total number of samples from each of 6 wards.
total <- c (29,62,72,41,89,71)

#Using prop.test function to perform multiple proportion test.
#syntax of prop.test function
```

```
str(prop.test)

## function (x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
##     conf.level = 0.95, correct = TRUE)

prop.test(x = users, n = total, alternative = "two.sided")

##
##  6-sample test for equality of proportions without continuity
##  correction
##
## data:  users out of total
## X-squared = 11.588, df = 5, p-value = 0.04089
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5    prop 6
## 0.3793103 0.5967742 0.3750000 0.6341463 0.4719101 0.4929577
```

Since *p*-value < alpha (0.05) there is sufficient evidence to reject the null hypothesis (H_0) and hence we conclude that proportion of users is not same in all the wards. We observe that maximum proportion of cloud storage users are in ward 4 that is in Akota.

7.2. Chi-square test for multiple response analysis

b) Objective: To check whether any association between specific cloud storage facilities usage on cloud company

Q1) which public cloud company you use? (Allow to tick multiple answers)
(a) Google Drive (b) Dropbox (c) One Drive (d) iCloud (e) Others (specify)----------

Q2) which are the cloud storage drive facilities you mostly use? (Allow to tick multiple ans.)
(a) Sync folder (b) Selective Sync (c) Folder sharing (d) Mobile Apps (e) Deleted file retention (f) File Editing (g) Versioning (h) others [specify]----------

H_0: There is no association between Cloud Company and specific cloud storage facilities usage.

H_1: There is association between Cloud Company and specific cloud storage facilities usage.

Facilities	Company				
	Google Drive	Dropbox	OneDrive	iCloud	Other company
Sync Folder	74	32	26	27	7
Selective sync	46	13	17	16	2
Folder sharing	65	32	20	17	3
Mobile apps	84	32	21	26	3
Deleted file retention	40	22	17	18	0
File editing	38	19	15	17	1
Versioning	5	5	5	2	0
Others	2	0	0	1	0

Pearson Chi-Square Tests	
Facilities	Company
Chi-square	143.442
Df	40
Sig.	0.000

Since *p*-value < alpha (0.05) the test provides enough evidence to reject the null hypothesis (H_0) and hence we conclude that there is association between cloud company and specific cloud storage facilities usage.

7.3. Chi-square test for multiple response analysis

c) Objective**:** To check whether the use of different cloud storage facilities depends on profession or not.

Q1) What is your profession?
(a) Student (b) Job (c) Business (d) Unemployed

Q2) Which are the cloud storage drive facilities you mostly use? (Allow to tick multiple answers)
(a) Sync folder (b) Selective Sync (c) Folder sharing (d) Mobile Apps (e) Deleted file retention (f) File Editing (g) Versioning (h) others [specify]----------

H_0: There is no association between profession and different cloud storage facilities usage.
H_1: There is association between profession and different cloud storage facilities usage.

Facilities	Profession			
	Student	Job	Business	Unemployed
Sync Folder	31	47	8	9
Selective sync	14	25	5	9
Folder sharing	28	42	5	2
Mobile apps	29	51	10	10
Deleted file retention	9	24	7	5
File editing	13	25	5	3
Versioning	3	2	0	0
Others	1	0	1	0

Pearson Chi-Square Tests	
Facilities	Profession
Chi-square	41.243
Df	24
Sig.	0.016

Since *p*-value < alpha (0.05) the test provides enough evidence to reject null hypothesis (H_0) and hence we conclude that there is association between profession and use of different cloud storage facilities.

7.4. Multinomial Logistic Regression analysis

d) Objective: To predict whether the user will use paid version cloud storage or not.

As per our domain knowledge the factors which may affect for the usage of paid version of cloud storage are: Income, profession, cloud storage usage, type of data on cloud, etc.

Using survey data, the response variable and 15 explanatory variables are as follows:

Y = Use of paid version (1 = Yes, 2 = No, 3 = May be)

X_1 = Income (1= No, 2 = Below one lac, 3 = 1-3 lac, 4 = 3-6 lac, 5 = 6-9 lac, 6 = Above 9 lac), X_2 = Profession (1 = Student, 2 = Job, 3 = Business, 4 = Unemployed),

X_3= Upload videos (1 = Yes, 2 = No), X_4 = Upload songs (1 = Yes, 2 = No),

X_5 = Upload images (1 = Yes, 2 = No), X_6 = Upload documents (1 = Yes, 2 = No),

X_7 = How frequently they use cloud (1 = (once in a week), 2 = (twice in a week), 3 = (thrice in a week), 4 = (four times in a week), 5 = (five times in a week)),

X_8= Apps backup on cloud (1= Yes with media, 2 = Without media, 3 = No),

X_9 = Apps media backup on cloud (1= Yes with media, 2 = Without media, 3 = No),

X_{10} = No. of cloud accounts they have (1 = 1, 2 = 2, 3 = 3, 4 = More than 3),

X_{11} = It provides me extra space (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5=Strongly agree),

X_{12} = Easy and convenient (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5= Strongly agree)**,** X_{13} = I can access my data from anywhere from any device (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5=Strongly agree),

X_{14} = I can easily share information with others (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5=Strongly agree),

X_{15} = I won't loss my data if my computer fails (1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5=Strongly agree)

In order to predict whether user will use paid version cloud storage or not, Multinomial logistic regression was used.

For multinomial logistic Regression, *Y*: Yes = 1, May be = 2 and No = 3.

No = 3 is taken as the base with ascending order of categories.

Model is fitted by using forward stepwise multinomial logistic regression.

Step Summary						
Model	Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
			–2 Log Likelihood	Chi-Square[a]	Df	Sig.
0	Entered	Intercept	372.201	.		
1	Entered	*X*15	346.012	26.189	8	0.001
2	Entered	*X*11	327.889	18.122	8	0.020
3	Entered	*X*1	308.477	19.412	10	0.035
Stepwise Method: Forward Entry						
a. The chi-square for entry is based on the likelihood ratio test.						

The fitted model gives the following hypothesis results.
H_0: No significance between null model (model with only intercept) and final model (model with all variable)
H1: Significance between null model (model with only intercept) and final model (model with all variables)

Model Fitting Information				
Model	Model Fitting Criteria	Likelihood Ratio Tests		
	–2 Log Likelihood	Chi-Square	Df	Sig.
Intercept	372.201			
Final	308.477	63.724	26	0.000

Here *p*-value < alpha (0.05) therefore the data provides enough evidence to reject null hypothesis (H_0) at 5% level of significance, thus the fitted model is significant.

Pseudo R-Square	
Cox and Snell	0.301
Nagelkerke	0.343
McFadden	0.171

Here Cox and Snell = 0.301, 30.1% of variation in *Y* is explained by fitted model.

Likelihood Ratio Tests				
Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	–2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	308.477[a]	0.000	0	0.000
*X*15	333.134	24.657	8	0.002
*X*11	326.186	17.709	8	0.024
*X*1	327.889	19.412	10	0.035

Since the probability is smaller than 0.05, hence it can be concluded that the three variables are significantly different from zero at 5 % level of significance.

Parameter Estimation

Future use		B	Std. Error	Wald	DF	P-value
Yes	Intercept	0.372	1.157	0.103	1	0.748
	[X15 = 1]	2.744	1.486	3.411	1	0.065
	[X15 = 2]	-0.134	1.096	0.015	1	0.902
	[X15 = 3]	1.798	1.02	3.107	1	0.078
	[X15 = 4]	-0.805	0.612	1.73	1	0.188
	[X15 = 5]	0	-	-	0	-
	[X11 = 1]	-1.886	1.206	2.447	1	0.118
	[X11 = 2]	-2.104	1.387	2.301	1	0.129
	[X11 = 3]	-2.429	1.079	5.067	1	0.024
	[X11 = 4]	0.648	0.781	0.688	1	0.407
	[X11 = 5]	0	-	-	0	-
	[X1 = 1]	1.173	1.047	1.256	1	0.262
	[X1 = 2]	1.915	1.198	2.555	1	0.11
	[X1 = 3]	0.584	1.071	0.297	1	0.586
	[X1 = 4]	0.893	1.124	0.631	1	0.427
	[X1 = 5]	1.471	1.806	0.663	1	0.415
	[X1 = 6]	0	-	-	0	-
May be	Intercept	2.831	0.927	9.327	1	0.002
	[X15 = 1]	2.692	1.665	2.615	1	0.106
	[X15 = 2]	-3.314	1.614	4.214	1	0.04
	[X15 = 3]	0.814	0.979	0.692	1	0.406
	[X15 = 4]	-0.936	0.572	2.675	1	0.102
	[X15 = 5]	0	-	-	0	-
	[X11 = 1]	-4.159	1.441	8.334	1	0.004
	[X11 = 2]	-0.618	1.49	0.172	1	0.679
	[X11 = 3]	-1.878	0.934	4.041	1	0.044
	[X11 = 4]	-1.149	0.731	2.466	1	0.116
	[X11 = 5]	0	-	-	0	-
	[X1 = 1]	-0.161	0.809	0.04	1	0.842
	[X1 = 2]	-0.672	1.071	0.394	1	0.53
	[X1 = 3]	-1.13	0.85	1.766	1	0.184
	[X1 = 4]	-0.512	0.907	0.318	1	0.573
	[X1 = 5]	1.892	1.441	1.726	1	0.189
	[X1 = 6]	0	-	-	0	-

Classification				
Observed	Predicted			
	Yes	Maybe	No	Percent Correct
Yes	22	27	6	40.0%
Maybe	9	69	7	81.2%
No	3	20	15	39.5%
Overall Percentage	19.1%	65.2%	15.7%	59.6%

59.6% times the fitted model does the correct classification.

Fitted Model in Logit Form

P: Whether user will use paid version in future or not.

= "Yes" or "May be "or "No"

Let *A* = Yes, *B* = May be, *C* = No

First logistic function is *A* with *C*

$$log\left(\frac{P(A)}{1-P(A)}\right) = 0.372 + 1.173X_{1_1} + 1.915X_{1_2} + 0.584X_{1_3} + 0.893X_{1_4} + 1.471X_{1_5} - 1.886X_{11_1} - 2.104X_{11_2} - 2.429X_{11_3} - 0.648X_{11_4} + 2.744X_{15_1} - 0.134X_{15_2} + 1.798X_{15_3} - 0.805X_{15_4}$$

where X_{ij} is the value of i^{th} variable at j^{th} level,

$$log\left(\frac{P(A)}{P(C)}\right) = P(A_{RHS})$$

Then logit function is written for *B* with *C*,

$$log\left(\frac{P(B)}{1-P(B)}\right) = 2.831 - 0.161X_{1_1} - 0.672X_{1_2} - 1.130X_{1_3} - 0.512X_{1_4} + 1.892X_{1_5} - 4.159X_{11_1} - 0.618X_{11_2} - 1.871X_{11_3} - 1.149X_{11_4} + 2.692X_{15_1} - 3.314X_{15_2} + 0.814X_{15_3} - 0.936X_{15_4}$$

$$log\left(\frac{P(B)}{P(C)}\right) = P(B_{RHS})$$

$$P(A) = \frac{P(A_{RHS})}{1 + P(A_{RHS}) + P(B_{RHS})}$$

$$P(B) = \frac{P(B_{RHS})}{1 + P(A_{RHS}) + P(B_{RHS})}$$

$$P(C) = \frac{1}{1 + P(A_{RHS}) + P(B_{RHS})}$$

The equation *P(A)* can be used to calculate the probability that whether the cloud storage user will use paid version in future or not.

7.5. Frequency analysis for security related issues

e) To gain insight whether users are aware of the associated potential risk or not with respect to gender wise and profession

Sr. No.	Security related issue	Category	GENDER		PROFFESSION			
			Male	Female	Business	Job	Student	Unemploy-ment
1.	People's perception on Security of data from market research	YES	39%	35%	30%	67%	35%	17%
		NO	27%	25%	4g5%	33%	24%	8%
		DON'T KNOW	34%	40%	25%	0%	41%	75%
2.	People's perception whether the government can access their data without their consent	YES	40%	30%	45%	100%	38%	17%
		NO	26%	24%	25%	0%	26%	0%
		DON'T KNOW	34%	46%	30%	0%	36%	83%
		NO	110	52	18	2	52	12
		DON'T KNOW	-	-	-	-	-	-

8. Conclusions

From the survey and above analysis, it can be concluded that proportion of users is not same in all the wards. We observe that maximum proportion of cloud storage users are in ward 4 that is in Akota. Usage of different cloud storage facilities depends on the cloud company. From multiple response analysis we observe that there is association between profession and different cloud storage facilities usage. In multinomial logistic regression, from equation *P(A), P(B), P(C)* we can predict whether the user will use paid version of cloud in future or not. We observe that perception of males and females on the security of data are nearby same, business profession are less aware about the risk of data on cloud, and IT profession people are more aware about the risk of data on cloud.

Acknowledgements

The authors express their gratefulness to the reviewer and the editor for their invaluable comments and suggestions, which have helped us to improve the paper substantially.

References

Yang, Haibo and Tate, Mary (2012). A descriptive literature review and classification of cloud computing research. *Communications of the Association for Information Systems*, **31**(**2**), 35-60. http://aisel.aisnet.org/cais/vol31/issl/2.

Aized Amin Soofi, M., Irfan Khan and Fazal-e-Amin (2014). A review on data security in cloud computing. *International Journal of Computer Applications*, **94** (**5**), 1-20.

Tinani, K., Choithwani, B., Patil, B., Pathan, F. and Salat T. (2019). Study on usage pattern of public cloud storage. *International Journal of Computer Sciences and Engineering*, **7(6)**, 922-927.

Spoorthy, V., Mamatha, M. and Santhosh, K. (2014). A survey on data storage and security in cloud computing. *International Journal of Computer Science and Mobile Computing*, **3**(**6**), 306-313.

Mohd, R., Mohd J. H. and Mohd M. (2012). Impact of cloud computing on IT industry: A Review and Analysis. *International Journal of Computer and Information Technology*, **1**(**2**), 1-5.

Sareen, Pankaj (2013). Cloud computing: Types, architecture, applications, concerns, virtualization and role of IT governance in cloud. *International Journal of Advanced Research in Computer Science and Software Engineering*, **3**(**3**), 533-538.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), 41-62

Generalized Lindley Shared Frailty Models

Arvind Pandey[1], **David D. Hanagal**[2] **and Shikhar Tyagi**[1]

[1]*Department of Statistics, Central University of Rajasthan, Rajasthan, India*
[2]*Department of Statistics, Savitribai Phule Pune University, Pune-411007, India.*

Received: 03 June 2020; Revised: 22 July 2020; Accepted: 24 July 2020

Abstract

We propose a new frailty distribution named as the generalized Lindley frailty distribution with generalized Weibull and exponential power as baseline distributions. To estimate the parameters in the models, the Bayesian paradigm of the Markov Chain Monte Carlo technique was designed. Bayesian comparison techniques have been performed for the comparison of models. We analyze kidney infection data and suggest a better model.

Key words: Bayesian estimation; Exponential power distribution; Generalized Lindley frailty; Generalized Weibull distribution; MCMC; Random censoring.

1. Introduction

In survival data, a common approach is that each individual under study experiencing the same risk factors which act as multiplicatively. Sometimes, in real-life situations risk (hazard rate) changes from one family to another family, one group to another group, one cluster to another cluster. Heterogeneity in the population exists, because of the mixture of groups of individuals with different risk factors. This heterogeneity is called as a frailty. Ignoring frailty may have adverse consequences. A random impact that is unobservable risk shared by the subject characterized as frailty which was introduced by Vaupel *et al.* (1979). To handle such kind of problems, many models have been derived in survival analysis. Since the establishment of the proportional hazard model given by Cox (1972), survival function has been dominated by hazard rate models. The reason behind the popularity of this model is, the significance of known covariates can be tested, also a relationship between lifetimes and covariates can be incorporated. Cox (1972) gave the following proportional hazard model or multiplicative hazard model as

$$\phi(t|\underline{K}) \;=\; \phi_0(t)e^{\underline{K}^{'}\underline{\beta_0}} \tag{1}$$

where, $\phi(t|\underline{K})$ stands for conditional hazard rate given the covariates, $\phi_0(t)$ stands for baseline hazard rate. $\underline{K}^{'} = (K_{1j}, K_{2j}, ..., K_{mj})$ are vector of known covariate and $\underline{\beta_0}$ is the vector of

Corresponding author: David D. Hanagal
E-mail: david.hanagal@gmail.com

regression parameters of order m corresponding to $\underline{K}$. Augmentation of Cox's proportional hazard model provided away to introduce the unknown covariates,

$$\phi(t|\underline{V}) = \phi_0(t)e^{\underline{K}^{'}\underline{\beta_0}+\underline{V}^{'}\underline{\beta_1}}$$

or

$$\phi(t|w) = w\phi_0(t)e^{\underline{K}^{'}\underline{\beta_0}} \tag{2}$$

where, $\underline{V}^{'} = (V_{1j}, V_{2j}, ..., V_{mj})$ are considered as the vector of unknown covariates respectively, $\underline{\beta_1}$ are indicated as the vector of regression coefficients of order m corresponding to $\underline{V}$. $w = e^{\underline{V}^{'}\underline{\beta_1}}$ called as frailty random effect. The conditional cumulative hazard function is given by

$$\Phi(t|w) = w\Phi_0(t)e^{\underline{K}^{'}\underline{\beta_0}} \tag{3}$$

where $\Phi_0(t) = \int_0^t \phi_o(t)dt$. The conditional survival function is given by

$$S(t|w) = \exp\left(-w\Phi_0(t)e^{\underline{K}^{'}\underline{\beta_0}}\right) \tag{4}$$

Frailty models firstly introduced by Vaupel *et al.* (1979) in univariate survival models that can be separated into multiplicative components. It has been assumed that the baseline hazard function has a multiplicative effect of frailties. Several frailty models had been proposed by Oakes (1989). As a frailty distribution, gamma, inverse Gaussian, positive stable distributions had been claimed by Hougaard (1986). Hougaard (1985, 1991, 2000) had discussed the different aspects of frailty on a broad scale. Log-normal distribution was proposed as frailty distribution by Flinn and Hackman (1982). In the last decade, frailty regression models in mixture distribution have been discussed by Hanagal (2008). Hanagal and Dabade (2013, 2015) proposed modeling of the inverse Gaussian frailty model and comparison of different frailty models for analyzing kidney infection data. Modeling kidney infection data for inverse Gaussian shared frailty was done by Hanagal and Pandey (2014a). Gamma frailty models for bivariate survival data were given by Hanagal and Pandey (2015a). Hanagal and Pandey (2017a) were used the shared inverse Gaussian frailty models based on additive hazard. For reversed hazard rate setup, Hanagal and Pandey (2014b, 2015b, 2016a, 2016b, 2017b) have contemplated gamma and inverse Gaussian shared frailty models with different baseline distribution functions. 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. Compound Poisson frailty was used by Hanagal and Kamble (2015) for Bayesian estimation. Analysis of kidney infection data and Australian twin data were done by Hanagal and Bhambure (2014, 2015, 2016) with different frailty distributions. Hanagal (2011, 2017, 2019) gave extensive literature review on different shared frailty models.

The main aim of this article has three objectives. First, generalized Lindley (GL) shared frailty models for hazard rate with generalized Weibull and exponential power as baseline distributions have been introduced. Second, Bayesian approach of estimation has been employed to estimate the unknown parameters under random censoring. Third, simulation study and data analysis have been done for the kidney infection data set.

2. Generalized Lindley Frailty Model

Lindley (1958) proposed a distribution with one parameter. Because of having only one parameter, the Lindley distribution does not provide enough flexibility for modeling purposes. It will be useful to consider further alternatives of this distribution. Zakerzadeh and Dolati (2009) proposed generalized Lindley distribution which generalizes Lindley distribution and includes exponential and gamma distributions as special cases. For a frailty distribution, generalized Lindley (GL) distribution has been considered in this paper. This distribution is the mixture of two gamma distributions G(θ,μ) and G(θ,η) with mixing coefficient $\theta/(\theta+1)$. That is the reason why GL frailty model is more adaptable in comparison with gamma frailty model. Probability density function of GL distribution has been specified below:

$$f_W(w) = \begin{cases} \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}w^{\mu-1}}{\Gamma\mu} + \frac{\theta^{\eta}w^{\eta-1}}{\Gamma\eta}\right]e^{-\theta w} & ; w \in \mathbb{R}^+, \mu, \eta, \theta \in \mathbb{R}^+ \\ 0 & ; \textit{otherwise} \end{cases}$$

with mean $E[W] = \frac{1}{1+\theta}\left[\mu + \frac{\eta}{\theta}\right]$. And corresponding variance is,

$$V(W) = \frac{1}{(1+\theta)}\left[\left(\mu^2 + \frac{\eta^2}{\theta}\right)\left(\frac{1}{\theta(1+\theta)}\right) + \left(\frac{\mu+\eta}{\theta}\right) - \left(\frac{2\mu\eta}{\theta(1+\theta)}\right)\right]$$

after applying identifiability property, *i.e.*, $E[W] = 1$ we get a relation between parameters $\eta = \theta\,(1+\theta-\mu) > 0$. Consequently, the density function, Laplace transformation and variance for GL reduced to,

$$f_W(w) = \begin{cases} \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}w^{\mu-1}}{\Gamma\mu} + \frac{\theta^{\theta(1+\theta-\mu)}w^{\theta(1+\theta-\mu)-1}}{\Gamma\theta(1+\theta-\mu)}\right]e^{-\theta w} & ;\ w, \theta \in \mathbb{R}^+, \mu \in (0, 1+\theta) \\ 0 & ;\ otherwise. \end{cases}$$

$$L_W(s) = \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(s+\theta)^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(s+\theta)^{\theta(1+\theta-\mu)}}\right] \tag{5}$$

$$V(W) = \frac{\theta^4 - \theta^3\mu + 3\theta^2(1+\theta) - 4\theta^2\mu + 3\theta\mu(\mu-1) + \mu^2}{\theta(1+\theta)^2} \tag{6}$$

n objects are postulated to be under study. (T_{1j}, T_{2j}) are contemplated as first and second survival time of $i^{th}(i = 1, 2)$ component of $j^{th}(j = 1, 2, ..., n)$ objects. The unconditional bivariate survival function at time $t_{1j} \in \mathbb{R}^+$ and $t_{2j} \in \mathbb{R}^+$ can be written as,

$$\begin{aligned} S(t_{1j}, t_{2j}) &= \int_{w_j \in \mathbb{R}^+} S(t_{1j}, t_{2j}|w_j) f_W(w_j) dw_j \\ &= \int_{w_j \in \mathbb{R}^+} e^{-W_j(\Phi_{01}(t_{1j}) + \Phi_{02}(t_{2j}))\rho_j} f_W(w_j) dw_j \\ &= L_{W_j}\left[(\Phi_{01}(t_{1j}) + \Phi_{02}(t_{2j}))\,\rho_j\right] \end{aligned} \tag{7}$$

where, $L_{W_j}(.)$ is Laplace transformation of frailty variable W_j. $\Phi_0(.)$ stands for cumulative baseline hazard rate and $\rho_j = e^{\underline{K}_j^{'}\underline{\beta}_j}$ is the therm containing the regression coefficients corresponding to known covariates. To get unconditional survival function, using equations (5) and (7),

$$S(t_{1j}, t_{2j}) = \frac{1}{(1+\theta)}\left[\frac{\theta^{\mu+1}}{(\theta+\rho(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j})))^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{(\theta+\rho(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j})))^{\theta(1+\theta-\mu)}}\right] \tag{8}$$

corresponding cross-ratio function given by Clayton (1978) and Oakes (1989) is given by,

$$\theta^{*}(t_{1j}, t_{2j}) = \frac{A * B}{\left(\theta^{\theta(\theta+1)}(-\mu+\theta+1)(C*\rho+\theta)^{\mu} + \mu\theta^{\mu(\theta+1)}(C*\rho+\theta)^{\theta(-\mu+\theta+1)}\right)^{2}}$$

where,

$$A = \theta^{\mu\theta-1}\left(\theta^{\theta(-\mu+\theta+1)}\left(\left(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j})\right)\rho+\theta\right)^{\mu} + \theta^{\mu+1}\left(\left(e^{\lambda_1 t_{1j}{}^{\alpha_1}} + e^{\lambda_2 t_{2j}{}^{\alpha_2}} - 2\right)\rho+\theta\right)^{\theta(-\mu+\theta+1)}\right)$$

$$B = \left(\theta^{\theta^2+\theta+1}(\mu^2 - 2\mu(\theta+1) + \theta(\theta+2)+2) - (\mu-1)\theta^{\theta(\theta+1)}\right)\left(\left(\Phi_1(t_1)+\Phi_2(t_2)\right)\rho+\theta\right)^{\mu} + \mu(\mu+1)\theta^{\mu(\theta+1)}\left(\left(\Phi_1(t_1)+\Phi_2(t_2)\right)\rho+\theta\right)^{\theta(-\mu+\theta+1)}$$

$$C = \left(\left(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j})\right)\right)$$

in the absence of frailty effect, model in the hazard rate setup will be,

$$S(t_{1j}, t_{2j}) = \exp\left(-\rho_j(\Phi_{01}(t_{1j})+\Phi_{02}(t_{2j}))\right) \tag{9}$$

One can have different baseline distributions for T_1 and T_2. After substituting different cumulative hazard functions in (8), we get different generalized Lindley frailty distributions.

3. Baseline Distributions

3.1 Generalized Weibull distribution

Here, the generalized Weibull distribution has been postulated as a baseline distribution. If a continuous random variable T follows the generalized Weibull distribution then the survival, hazard, and cumulative hazard function, are respectively,

$$S(t) = \begin{cases} 1-\left(1-e^{-\delta t^{\xi}}\right)^{\zeta} & ; t \in \mathbb{R}^{+}, \delta, \zeta, \xi \in \mathbb{R}^{+} \\ 1 & ; \textit{otherwise} \end{cases} \tag{10}$$

$$\phi_0(t) = \begin{cases} \frac{\xi\zeta\delta t^{\xi-1}e^{-\delta t^{\xi}}(1-e^{-\delta t^{\xi}})^{\zeta-1}}{1-\left(1-e^{-\delta t^{\xi}}\right)^{\zeta}} & ; t \in \mathbb{R}^{+}, \delta, \zeta, \xi \in \mathbb{R}^{+} \\ 1 & ; \textit{otherwise} \end{cases} \tag{11}$$

$$\Phi_0(t) = \begin{cases} -\log\left(1-\left(1-e^{-\delta t^{\xi}}\right)^{\zeta}\right) & ; t \in \mathbb{R}^{+}, \delta, \zeta, \xi \in \mathbb{R}^{+} \\ 0 & ; \textit{otherwise} \end{cases} \tag{12}$$

3.2 Exponential power distribution

Another baseline distribution we considered is exponential power distribution. A continuous random variable T is said to follow exponential power distribution if survival, hazard, and cumulative hazard function is,

$$S(t) = \begin{cases} e^{(1-e^{\delta t^{\zeta}})} & ; t \in \mathbb{R}^{+}, \delta, \zeta \in \mathbb{R}^{+} \\ 1 & ; \textit{otherwise} \end{cases} \tag{13}$$

$$\phi_0(t) = \begin{cases} \zeta \delta t^{\zeta-1} e^{\delta t^{\zeta}} & ; t \in \mathbb{R}^{+}, \delta, \zeta \in \mathbb{R}^{+} \\ 0 & ; \textit{otherwise} \end{cases} \tag{14}$$

$$\Phi_0(t) = \begin{cases} e^{\delta t^{\zeta}} - 1 & ; t \in \mathbb{R}^{+}, \delta, \zeta \in \mathbb{R}^{+} \\ 0 & ; \textit{otherwise} \end{cases} \tag{15}$$

Kolmogorov–Smirnov (K–S) statistic for goodness of fit shows that both baseline distributions are fitting well to kidney infection data set(see section 7, Figure 1-4).

4. Proposed Model

Due to group variation or frailty and individual variation described by the hazard function, a shared frailty model can be considered as a mixture model in survival analysis. After substituting cumulative hazard function for generalized Weibull and exponential power baseline distributions in equations (8) and (9), we get the following four survival functions.

$$S(t_{1j}, t_{2j}) = \frac{1}{(1+\theta)} \left[\frac{\theta^{\mu+1}}{\left[\theta + \left\{ \log\left(1 - \left(1 - e^{\delta_1 t_{1j}^{\xi_1}}\right)^{\zeta_1}\right) + \log\left(1 - \left(1 - e^{\delta_2 t_{2j}^{\xi_2}}\right)^{\zeta_2}\right)\right\}\rho\right]^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{\left[\theta + \left\{ \log\left(1 - \left(1 - e^{\delta_1 t_{1j}^{\xi_1}}\right)^{\zeta_1}\right) + \log\left(1 - \left(1 - e^{\delta_2 t_{2j}^{\xi_2}}\right)^{\zeta_2}\right)\right\}\rho\right]^{\theta(1+\theta-\mu)}} \right] \tag{16}$$

$$S(t_{1j}, t_{2j}) = e^{\rho_j \left(\log\left(1 - \left(1 - e^{\delta_1 t_{1j}^{\xi_1}}\right)^{\zeta_1}\right) + \log\left(1 - \left(1 - e^{\delta_2 t_{2j}^{\xi_2}}\right)^{\zeta_2}\right)\right)} \tag{17}$$

$$S(t_{1j}, t_{2j}) = \frac{1}{(1+\theta)} \left[\frac{\theta^{\mu+1}}{\left(\theta + \rho\left\{e^{\delta_1 t_{1j}^{\zeta_1}} + e^{\delta_2 t_{2j}^{\zeta_2}} - 2\right\}\right)^{\mu}} + \frac{\theta^{\theta(1+\theta-\mu)}}{\left(\theta + \rho\left\{e^{\delta_1 t_{1j}^{\zeta_1}} + e^{\delta_2 t_{2j}^{\zeta_2}} - 2\right\}\right)^{\theta(1+\theta-\mu)}} \right] \tag{18}$$

$$S(t_{1j}, t_{2j}) = e^{-\rho\left\{e^{\delta_1 t_{1j}^{\zeta_1}} + e^{\delta_2 t_{2j}^{\zeta_2}} - 2\right\}} \tag{19}$$

Here, equations (16), (17) can be called as Model-I, Model-III respectively that have been established for generalized Weibull baseline distribution with and without frailty and equations (18), (19) can be called as Model-II and Model-IV respectively that have been established for exponential power baseline distribution with and without frailty.

5. Likelihood Design and Bayesian Paradigm

For the study, n individuals have been considered. Observed failure times have been indicated by (t_{1j}, t_{2j}). We are using the random censoring scheme. Censoring time, supposed to be indicated by c_{1j} and c_{2j} for j^{th} individual $(j = 1, 2, 3, ..., n)$. Independence between censoring schemes and lifetimes of individuals has been presumed. Likelihood function can be described for bivariate lifetime random variable of the j^{th} individual as,

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

and likelihood function will be,

$$L(\underline{\Theta}, \underline{\beta}, \theta, \mu) = \prod_{j=1}^{n_1} f_1(t_{1j}, t_{2j}) \prod_{j=1}^{n_2} f_2(t_{1j}, c_{2j}) \prod_{j=1}^{n_3} f_3(c_{1j}, t_{2j}) \prod_{j=1}^{n_4} f_4(c_{1j}, c_{2j}) \tag{20}$$

where,$\underline{\Theta}$, $\underline{\beta}$, θ and μ are vector of baseline parameters and the vector of regression coefficients and frailty parameters respectively. Likelihood function for without frailty model is,

$$L(\underline{\Theta}, \underline{\beta}) = \prod_{j=1}^{n_1} f_1(t_{1j}, t_{2j}) \prod_{j=1}^{n_2} f_2(t_{1j}, c_{2j}) \prod_{j=1}^{n_3} f_3(c_{1j}, t_{2j}) \prod_{j=1}^{n_4} f_4(c_{1j}, c_{2j}) \tag{21}$$

let n_1, n_2, n_3 and n_4 be the number of pairs for which first and second failure times (t_{1j}, t_{2j}) lie in the ranges $t_{1j} < c_{1j}, t_{2j} < c_{2j}$; $t_{1j} < c_{1j}, t_{2j} > c_{2j}$; $t_{1j} > c_{1j}, t_{2j} < c_{2j}$ and $t_{1j} > c_{1j}, t_{2j} > c_{2j}$ respectively and let

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}} \\ f_2(t_{1j}, c_{2j}) &= -\frac{\partial S(t_{1j}, c_{2j})}{\partial t_{1j}} \\ f_3(c_{1j}, t_{2j}) &= -\frac{\partial S(c_{1j}, t_{2j})}{\partial t_{2j}} \\ f_4(c_{1j}, c_{2j}) &= S(c_{1j}, c_{2j}). \end{aligned} \tag{22}$$

substituting cumulative hazard rates $\Phi_{01}(t_{1j})$ and $\Phi_{02}(t_{2j})$ and survival function $S(t_{1j}, t_{2j})$ in equation (22) for Model-I and Model-II and by differentiating we get the likelihood function. The maximum likelihood method has a crucial importance in computing efficient estimators. Inappropriately, due to a convergence problem, maximum likelihood failed to estimate the parameters, because of Model-I has thirteen-dimensional, Model-II, Model-III have eleven-dimensional and Model-IV has nine-dimensional optimization problem. The Bayesian scenario has been discussed by several researchers for estimating parameters of the frailty models. For gamma and log-normal frailty models, the Bayesian paradigm has been contemplated by Santos and Achcar (2010). Weibull and piecewise exponential model have been discussed by Ibrahim *et al.* (2001) with gamma frailty. The joint posterior density function of parameters for given failure times is obtained as,

$$\pi(\Theta, \theta, \mu, \underline{\beta_0}) \propto L(\Theta, \mu, \underline{\beta}_0) g_1(\zeta) g_2(\xi) g_3(\delta) g_4(\theta) g_5(\mu) \prod_{i=1}^{5} p_i(\beta_{0i\times 1})$$

where $g_i(.)$ indicates the prior density function with known hyperparameters of corresponding argument for baseline parameters and frailty variance; $p_i(.)$ is prior density function for regression coefficient β_{0i} and likelihood function is $L(.)$. An important assumption here is, all the parameters are independently distributed. In a similar way, joint posterior density function can be written for without frailty models. To estimate the parameters of the models, Metropolis-Hastings algorithms and Gibbs samplers have been used. Geweke test (see Geweke, 1992) and Gelman-Rubin (see Gelman and Rubin, 1992) statistics have been used to monitor the convergence of a Markov chain to a stationary distribution.

Due to the high-dimensions of conditional distributions, it is not unproblematic to integrate out. Thus, it has been considered that full conditional distributions can be obtained as they are proportional to the joint distribution of the parameter of the model. The conditional distribution for single parameter δ with frailty as,

$$\psi_1(\delta \mid \xi, \zeta, \theta, \mu, \underline{\beta_0}) \quad \propto \quad L(\delta, \xi, \zeta, \theta, \mu, \underline{\beta_0}) \cdot g_1(\delta) \tag{23}$$

and the conditional distribution for single parameter δ without frailty as,

$$\psi_1(\delta \mid \xi, \zeta, \underline{\beta_0}) \quad \propto \quad L(\delta, \xi, \zeta, \underline{\beta_0}) \cdot g_1(\delta)$$

similarly full conditional distributions can be obtained.

6. Simulation Study

A simulation study has been executed to appraise the Bayesian estimation paradigm for Model-I and Model-II. Single covariate K_1 has been considered as follows normal distribution. The frailty variable W is assumed to follow generalized Lindley distribution. Independence between lifetimes of individuals has been considered. Samples are generated using the subsequent mechanism,

1. From the binomial distribution with probability 0.6, 25 values for K_1 has been generated.

2. For known covariates, compute $\rho = e^{K_1\beta_1}$.

3. Lifetimes reckoned to follows generalized Weibull and exponential power baseline distributions for given frailty W_j. 25 values of lifetimes have been spawned after using ensuing manners.
Conditional survival function for lifetime t_j $(j = 1, 2, ..., n)$ for given frailty $W_j = w_j$ and covariate K_1 is,

$$S(t_j \mid w_j, K_1) \quad = \quad e^{-w_j H_0(t_j)\rho}$$

Equating $S(t_j \mid w_j, K_1)$ to random number, say $v_j (0 < v_j < 1)$ spawned from $U(0, 1)$ over $t_j > 0$ we get,
for Model-I,

$$t_j \quad = \quad \left(-\frac{1}{\delta}\log(1-(1-v_j^{\frac{1}{w_j\rho_j}})^{\frac{1}{\zeta}})\right)^{\frac{1}{\xi}}$$

for Model-II,

$$t_j \quad = \quad \left(\frac{1}{\delta}\log(1-\frac{1}{w_j\rho_j}\log(v_j))\right)^{\frac{1}{\zeta}}$$

4. Censoring time c_j has been spawned from $G(0.9, 0.01)$ for Model-I.

5. Observe the j^{th} survival time $t_j^* = min(t_j, c_j)$ and the censoring indicator δ_j for the j^{th} individual $(j = 1, 2, ..., 25)$ where,

$$\delta_j = \begin{cases} 1, & ; t_j < c_j \\ 0, & ; t_j > c_j \end{cases}$$

thus we have data consisting of 25 pairs of survival times t_j^* and the censoring indicator δ_j.

Concurrently, with different priors and starting points, two chains based on two priors (one is based on gamma prior and another is based on uniform prior) have been operated. Both chains recapitulated 100,000 times. Gelman-Rubin test (see Gelman and Rubin, 1992) values are very close to one. Due to small values of Geweke test statistic (see Geweke, 1992) and corresponding p-values, the chains reach stationary distribution for both prior sets. In view of, estimates of parameters were about the same, no impact of prior distributions has been founded on posterior summaries. Here, the analysis for one chain has been exhibited because both the chains have shown similar results. Tables 1 and 2 present the estimates and the credible intervals of the parameters for the Models I and II 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.

7. Analysis of Kidney Infection Data

To elucidate the Bayesian estimation paradigm, kidney infection data of McGilchrist and Aisbett (1991) has been considered. This data consists of 38 patients, recurrence times (in days) of infection are given which can be outlined as these are recorded from the insertion of the catheter until it has to be removed due to infection. Data having five known covariates age, sex (Female=1, Male=0), and disease type Glomerulo Neptiritis (GN), Acute Nephritis (AN) and Polycystic Kidney Disease (PKD). Opine first and second time to infection is symbolized by T_{1j} and T_{2j}. Five covariates age, sex, GN, AN and PKD are symbolized by K_1, K_2, K_3, K_4 and K_5. To check goodness of fit of kidney data set, we consider Kolmogrove-Smirnov (K-S) test for two baseline distributions. Table 3 gives the p-values of goodness of fit test for Model I and Model II. Thus from p-values of K-S test we can say that there is no statistical evidence to reject the hypothesis that data are from the Model I and Model II in the marginal case and we assume that they also fit for bivariate case. Figures 1-4 show the parametric plot with semi-parametric plot for models (Model I and Model II) with frailty for T_1 and T_2 separately and both lines are close to each other.

For frailty parameters, gamma prior distribution with very small shape and scale parameters (say, 0.0001) has been used. Additionally, it can be considered, regression coefficients are normally distributed with mean zero and high variance (say 1000). A similar type of prior was used in Ibrahim et al (2001) and Santos and Achcar (2010). That's why for frailty parameters θ, μ and regression coefficients $\beta_{0i}, i = 1, ..., 5$, vague priors have been used. Because of no information about baseline parameter having, therefore, prior distribution corresponding to baseline parameters are also considered flat. We considered two different vague prior distributions for baseline parameters, one is gamma distribution with shape and scale hyperparameters ϵ_1, ϵ_2 respectively and another is uniform distribution with interval (ν_1, ν_2). All the hyperparameters are known. Under the Bayesian paradigm, for both models, two parallel chains have been run. Also, two sets of prior distributions have been used with different starting points using the Metropolis-Hastings algorithm and Gibbs sampler based on normal transition kernels. It can be said that estimates are independent of the different prior distributions because, for both sets of priors, estimates of parameters are approximately similar. We got almost similar convergence rate of Gibbs sampler for both sets of priors. Here, the analysis for one chain has been exhibited because both the chains have shown similar results.

Markov chain has seemed to reach the stationary state because of the zigzag pattern of the trace plots for all the parameters that gesture parameters move and mix more freely (See Figure 5). Coupling from the past plot has been applied to fix up the burn-in period (See Figure 6). A sequence of draws may have serial correlation after the burn-in period. Randomness may not be shown in successive draws. But almost independence can be seen in values at the extensive split. After using the values from the single run of the Markov chain, a vague sample can be obtained from the posterior distribution. Because of the burn-in period, it has been founded at extensive spaced time points. Autocorrelation function (ACF) plots can be utilized to examine the appropriate blend of our chains (See Figure 7). ACF plot for each parameter is converging to the posterior mean of the parameter, thus, represents a good mixing of the chain. Thus, our diagnostic plots suggest that the MCMC chains are mixing very well. After a certain lag, the serial correlation of the parameters turns out to almost negligible for all the parameters. Observations are shown independently after thinning the serial correlation function plot (See Figure 8). For visual approximate estimates as

confirmative measures such as posterior density plots also drawn for Model-I. It has been observed in some of posterior densities of the parameters depict multi-modal shapes which are quite possible in frailty models. The Gelman-Rubin convergence statistic values are closely equal to one. The Geweke test statistic values are somewhat small, and the corresponding p-values are large enough to say that the chains reach stationary distribution. Tables 4-7 give the values of posterior mean and the standard error with 95% credible intervals, the Gelman-Rubin statistics values and the Geweke test with p-values for Model I, II, III and IV. Table 8 present the values of AIC, BIC and DIC values for both models. Values of AIC, BIC, and DIC, given in Table 8, have been used to the comparison of all models. Model-I holds the lowest possible values of AIC, BIC, and DIC. For all models, regression coefficients contained different values. For Model-I and Model-II, the credible interval of $\beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}$ are not contained zero. It indicates that covariates sex, diseases GN, AN and PKD have a significant effects on all four models. It is being indicated that sex (β_2), disease PKD (β_5) are significant factors for kidney infection, having negative effects for all the four models. Negative value of β_2 indicates that the female patients have a slightly lower risk for infection. Negative value of β_5 indicates that the patients with the disease PKD has a slightly lower risk for infection.

8. Conclusions

Generalized Lindley frailty model under generalized Weibull and exponential power baseline distributions have been proposed. To fit the proposed models M-H algorithm and Gibbs samplers have been applied. Analysis has been done in R statistical software with self-written programs. The value of both frailty parameters for Model-I ($\theta = 3.08680, \mu = 2.89438$) and Model-II ($\theta = 2.89271, \mu = 2.49934$) are very high (See Tables 4 and 5) and corresponding variances are 1.38334 and 1.42248 by using equation (2.2). This exhibits that there is a strong indication of heterogeneity among the patient in the population for the data set. To take the decision about all models, different tools have been utilized. With the lowest value of AIC, BIC and DIC, given by Table 8, it can be said that Model-I and Model-II are better than Model-III and Model-IV for analyzing kidney infection data. The generalized Lindley frailty with generalized Weibull baseline (Model-I) is the best among all four models. For kidney infection data, sex, diseases AN, GN, and PKD have been found statistically significant factors for both with frailty and without frailty models (See Tables 4-7). Our proposed frailty model (Model-I) has been founded better in compare to Hanagal and Pandey's (2015a) frailty model with baseline generalized Weibull distribution. In a similar way, with a minimum value of AIC, our proposed frailty model (Model-II) has been founded better in compare of Hanagal and Dabade's (2015) frailty model.

Acknowledgments

The authors are thankful to the reviewer for the valuabe suggestions and comments.

References

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. *Biometrika*, **65**, 141-151.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society*, **B34**, 187-220.

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

Flinn, C. J. and Heckman, J. J. (1982). New methods for analyzing individual event histories. *Sociological Methodology*, **13**, 99-140.

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

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. (2008). Frailty regression models in mixture distributions. *Journal of Statistical Planning and Inference*, **138(8)**, 2462-68.

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

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

Hanagal, D. D. (2019). *Modeling Survival Data Using Frailty Models*. 2nd Edition. Springer; Singapore.

Hanagal, D. D., and Bhambure, S. M. (2014). Analysis of kidney infection data using shared positive stable frailty models. *Advances in Reliability*, **1**, 21-39.

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 and Methods*, **42(20)**, 3744-3769.

Hanagal, D. D., and Dabade, A. D. (2015). Comparison of shared frailty models for kidney infection data under exponential power baseline distribution. *Communications in Statistics-Theory and Methods*, **44(23)**, 5091-5108.

Hanagal, D. D., and Kamble, A. T. (2015). Bayesian estimation in shared compound poisson frailty models. *Journal of Reliability and Statistical Studies*, **8(1)**, 159-180.

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 bivariate 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. (2016a). Gamma shared frailty model based on reversed hazard rate. *Communications in Statistics-Theory and Methods*, **45(7)**, 2071-2088.

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

Hanagal, D. D., and Pandey, A. (2017a). Shared inverse Gaussian frailty models based on additive hazards. *Communications in Statistics-Theory and Methods*, **46(22)**, 11143-11162.

Hanagal, D. D., and Pandey, A. (2017b). Shared frailty models based on reversed hazard rate for modified inverse Weibull distribution as baseline distribution. *Communications in Statistics-Theory and Methods*, **46(1)**, 234-246.

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. (1985). Discussion of the paper by D.G. Clayton and J. Cuzick. *Journal of the Royal Statistical Society*, A, **148**, 113-14.

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

Hougaard, P. (1991). Modeling heterogeneity in survival data. *Journal of Applied Probability*, **28**, 695-701.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. *Springer:* New York.

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

Lindley, D. V. (1958). Fiducial distributions and Bayes's theorem. *Journal of the Royal Statistical Society*, B, **20**, 102-107.

McGilchrist, C. A. and Aisbett, C. W.(1991): Regression with frailty in survival analysis. *Biometrics*, **47**, 461-466.

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

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.

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

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

Zakerzadeh, H. and Dolati, A. (2009). Generalized Lindley distribution. *Journal of Mathematical Extension*, **3(2)**, 1-17.

Appendix
Summary of Tables and Figures

Table 1: Posterior Summary of Generalized Lindley Frailty with Baseline Generalized Weibull (Simulation Study: Model I)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
$\zeta_1(3.1)$	3.08997	0.17672	2.68349	3.44157	0.00101	0.50040	0.99999
$\delta_1(0.15)$	0.15417	0.01551	0.12216	0.17852	-0.01145	0.49543	1.00378
$\xi_1(0.85)$	0.85822	0.03033	0.79395	0.91321	0.01542	0.50615	1.00239
$\zeta_2(5.0)$	4.99795	0.55329	4.07576	5.95502	-0.00825	0.49671	0.99998
$\delta_2(0.26)$	0.27186	0.02817	0.20969	0.31781	0.00043	0.50017	1.00494
$\xi_2(0.74)$	0.74822	0.02897	0.69046	0.81146	-0.00174	0.49930	1.00096
$\theta(3.0)$	2.99983	0.15562	2.71870	3.28869	0.01826	0.50729	1.00002
$\mu(2.5)$	2.49520	0.08979	2.33103	2.65688	0.00415	0.50166	1.00201
$\beta_1(0.005)$	0.00458	0.00349	-0.00237	0.01135	-0.00530	0.49788	1.00021

Table 2: Posterior Summary of Generalized Lindley Frailty with Baseline Exponential Power (Simulation Study: Model II)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
$\zeta_1(0.75)$	0.71106	0.03077	0.65061	0.77292	0.00752	0.50300	1.00123
$\delta_1(0.09)$	0.09774	0.00960	0.07452	0.10972	-0.00332	0.49868	1.01043
$\zeta_2(0.7)$	0.75231	0.05995	0.62533	0.86400	0.00035	0.50014	1.00299
$\delta_2(0.06)$	0.06937	0.00764	0.05270	0.07915	0.00347	0.50138	1.00051
$\theta(1.2)$	1.19076	0.09615	1.01631	1.37833	-0.00444	0.49823	1.00352
$\mu(0.7)$	0.70361	0.03565	0.63321	0.76945	-0.00327	0.49870	1.00050
$\beta_1(0.003)$	0.00303	0.00173	-0.00049	0.00664	-0.00322	0.49871	0.99997

Table 3: p-values of K-S Statistics for goodness of fit test for Kidney Infection data set

	Recurrence Time	
Distribution	**First**	**Second**
Model I	0.5174	0.6060
Model II	0.1184	0.4185

Table 4: Posterior Summary of Generalized Lindley Frailty with Baseline Generalized Weibull for Kidney Infection Data (Model I)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
ζ_1	2.99002	0.15637	2.70025	3.29677	-0.00910	0.49637	1.00016
δ_1	0.18320	0.01455	0.15510	0.21414	-0.00871	0.49653	1.00094
ξ_1	0.78464	0.02354	0.73846	0.83274	0.00234	0.50094	1.00005
ζ_2	8.99452	0.97203	7.14211	10.92913	-0.00559	0.49777	0.99998
δ_2	0.30022	0.01887	0.26317	0.34148	-0.00381	0.49848	0.99996
ξ_2	0.67188	0.02468	0.62380	0.71751	-0.00787	0.49686	1.00005
θ	3.08680	0.12306	2.84842	3.35547	0.00019	0.50008	1.00048
μ	2.89438	0.12997	2.63819	3.16898	0.00218	0.50087	1.00010
β_1	0.00091	0.00054	-0.00004	0.00185	0.00781	0.50312	1.00117
β_2	-2.02839	0.22274	-2.46437	-1.58541	0.00661	0.50312	1.00306
β_3	-0.00446	0.00259	-0.00928	-0.00005	-0.00167	0.49933	1.00001
β_4	0.44121	0.20265	0.07885	0.80228	-0.00178	0.49929	1.00117
β_5	-1.06697	0.25150	-1.49606	-0.59086	0.00912	0.50364	1.00081

Table 5: Posterior Summary of Generalized Lindley Frailty with Baseline Exponential Power for Kidney Infection Data (Model II)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
ζ_1	0.59126	0.01185	0.56663	0.61349	-0.00265	0.49894	1.00119
δ_1	0.08507	0.00315	0.07859	0.09102	0.00065	0.50026	1.00011
ζ_2	0.67026	0.01459	0.64434	0.69888	-0.00376	0.49850	0.99998
δ_2	0.05114	0.00301	0.04569	0.05758	-0.00204	0.49919	1.00012
θ	2.89271	0.09667	2.70694	3.09136	0.00038	0.49965	1.00047
μ	2.49934	0.03404	2.43316	2.56418	-0.00088	0.49831	1.00025
β_1	-0.00044	0.00205	-0.00431	0.00306	0.00464	0.50185	1.00006
β_2	-1.67836	0.10168	-1.86985	-1.49478	-0.00517	0.50185	1.00406
β_3	0.21870	0.02916	0.17235	0.26824	0.00703	0.50280	1.00004
β_4	0.75038	0.11075	0.55099	0.93289	-0.01367	0.49455	1.00029
β_5	-0.67039	0.05381	-0.76419	-0.57894	-0.00423	0.49831	0.99998

Table 6: Posterior Summary of Generalized Weibull Distribution for Kidney infection Data (Model III)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
ζ_1	2.48490	0.30943	1.89920	3.08310	-0.00829	0.49670	1.00010
δ_1	0.20310	0.06863	0.09050	0.35890	-0.00614	0.49750	0.99990
ξ_1	0.60490	0.07812	0.46010	0.76200	0.01195	0.50470	0.99990
ζ_2	5.04010	0.50556	4.09990	5.94920	0.00073	0.50030	1.00030
δ_2	0.32220	0.08144	0.17580	0.49420	-0.00887	0.49650	1.00000
ξ_2	0.51290	0.06162	0.38820	0.63330	0.01124	0.50450	1.00000
β_1	0.00070	0.00279	-0.00440	0.00630	-0.00968	0.49610	1.00040
β_2	-1.07160	0.31695	-1.67560	-0.46080	-0.01568	0.49370	0.99990
β_3	-0.01590	0.02781	-0.06770	0.03750	0.00845	0.50340	1.00040
β_4	-0.00410	0.00660	-0.01670	0.00780	-0.00533	0.49780	0.99990
β_5	0.00120	0.00185	-0.00210	0.00460	0.00589	0.50240	1.00000

Table 7: Posterior Summary of Exponential Power Distribution for Kidney infection Data (Model IV)

Parameter	Estimate	S.E.	L.C.L.	U.C.L.	Geweke test	p-value	Gelman-Rubin test
ζ_1	0.61387	0.01685	0.57520	0.64394	0.00142	0.50057	1.00138
δ_1	0.06108	0.00327	0.05428	0.06775	0.00493	0.50197	1.00050
ζ_2	0.63406	0.01645	0.60266	0.66541	0.00632	0.50252	1.00090
δ_2	0.05032	0.00305	0.04411	0.05608	-0.00380	0.49849	1.00051
β_1	-0.00193	0.00178	-0.00579	0.00084	-0.00567	0.49774	1.00285
β_2	-1.61959	0.09413	-1.81647	-1.48638	0.00300	0.49774	0.99998
β_3	0.21887	0.02533	0.17269	0.26366	-0.01487	0.49407	0.99999
β_4	0.76834	0.09830	0.57241	0.93181	0.00916	0.50365	1.00803
β_5	-0.67269	0.03430	-0.74420	-0.60866	-0.00258	0.49897	1.00013

Table 8: AIC, BIC and DIC Comparison

Model	AIC	BIC	DIC
Model-I	682.2537	703.5423	661.4339
Model-II	689.5993	707.6128	670.3702
Model-III	690.2814	708.2949	678.103
Model-IV	690.3103	705.0486	676.7688

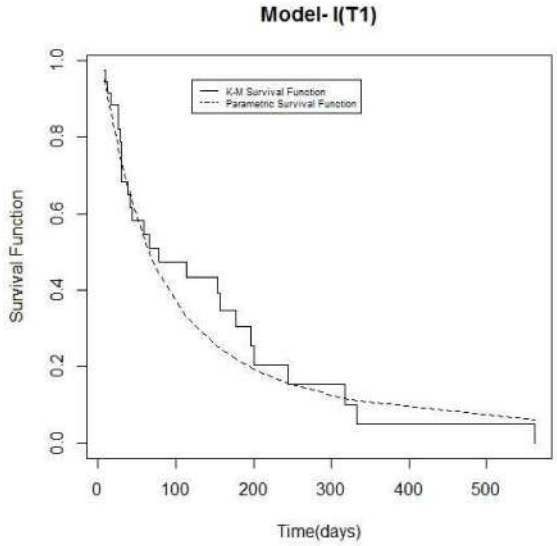


Figure 1

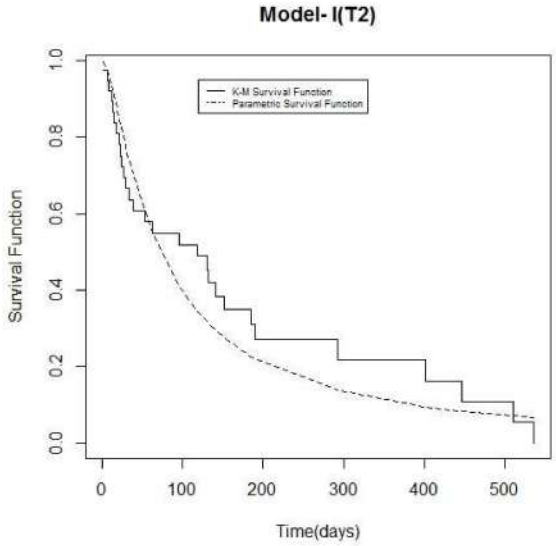


Figure 2

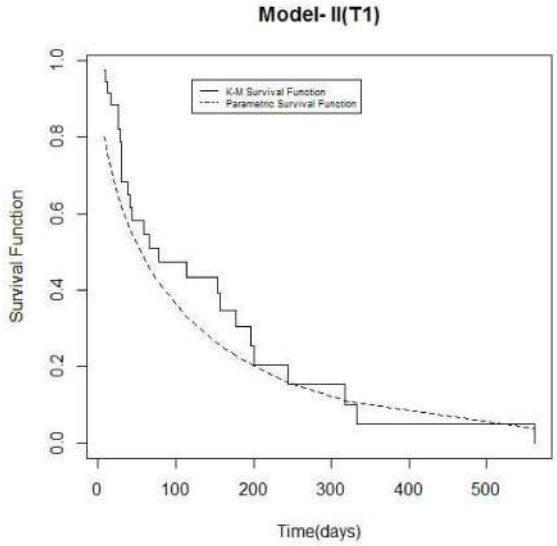


Figure 3

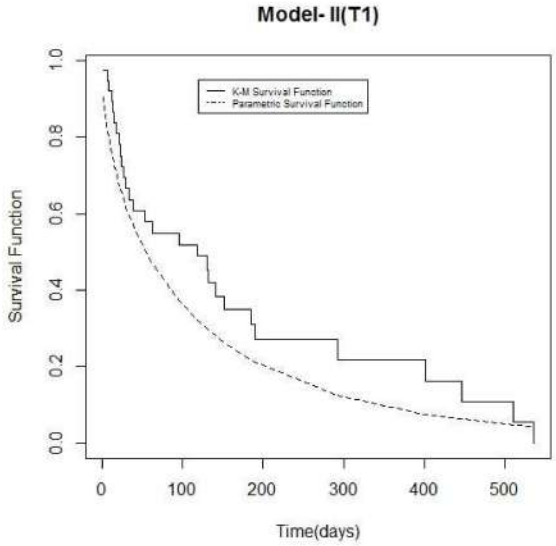


Figure 4

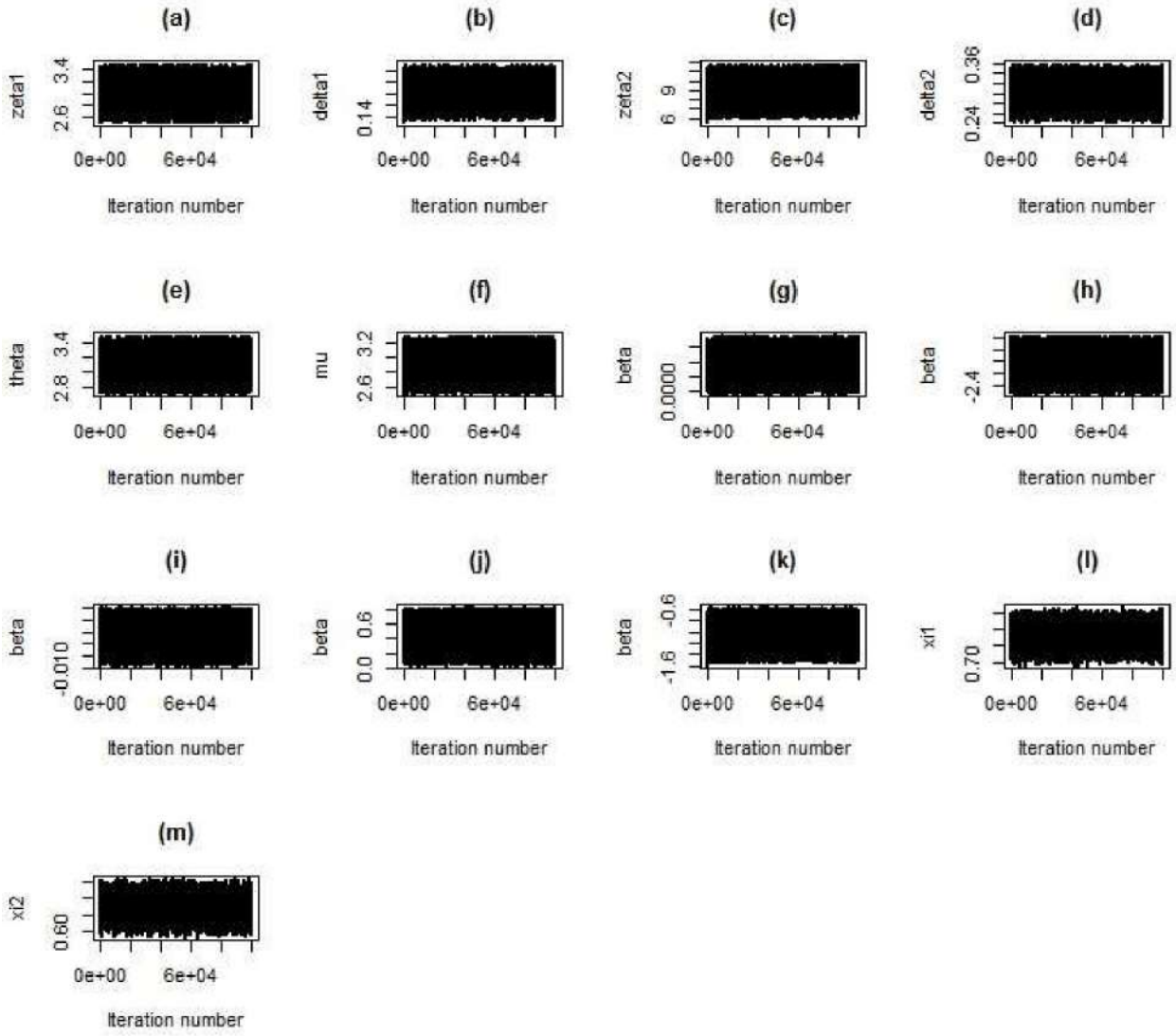


Figure 5: Trace plots for Model-I

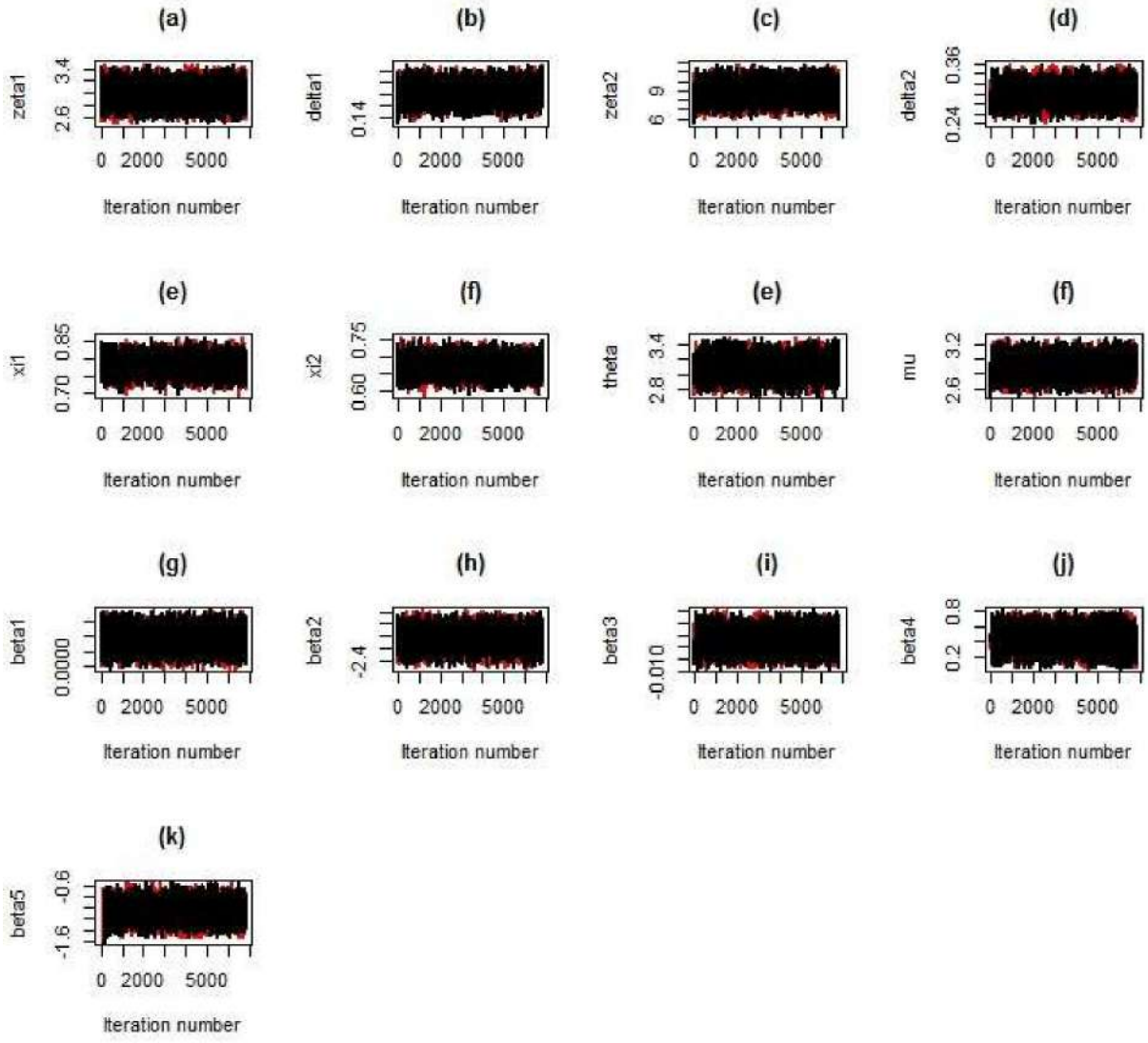


Figure 6: Coupling from the past plots for Model-I

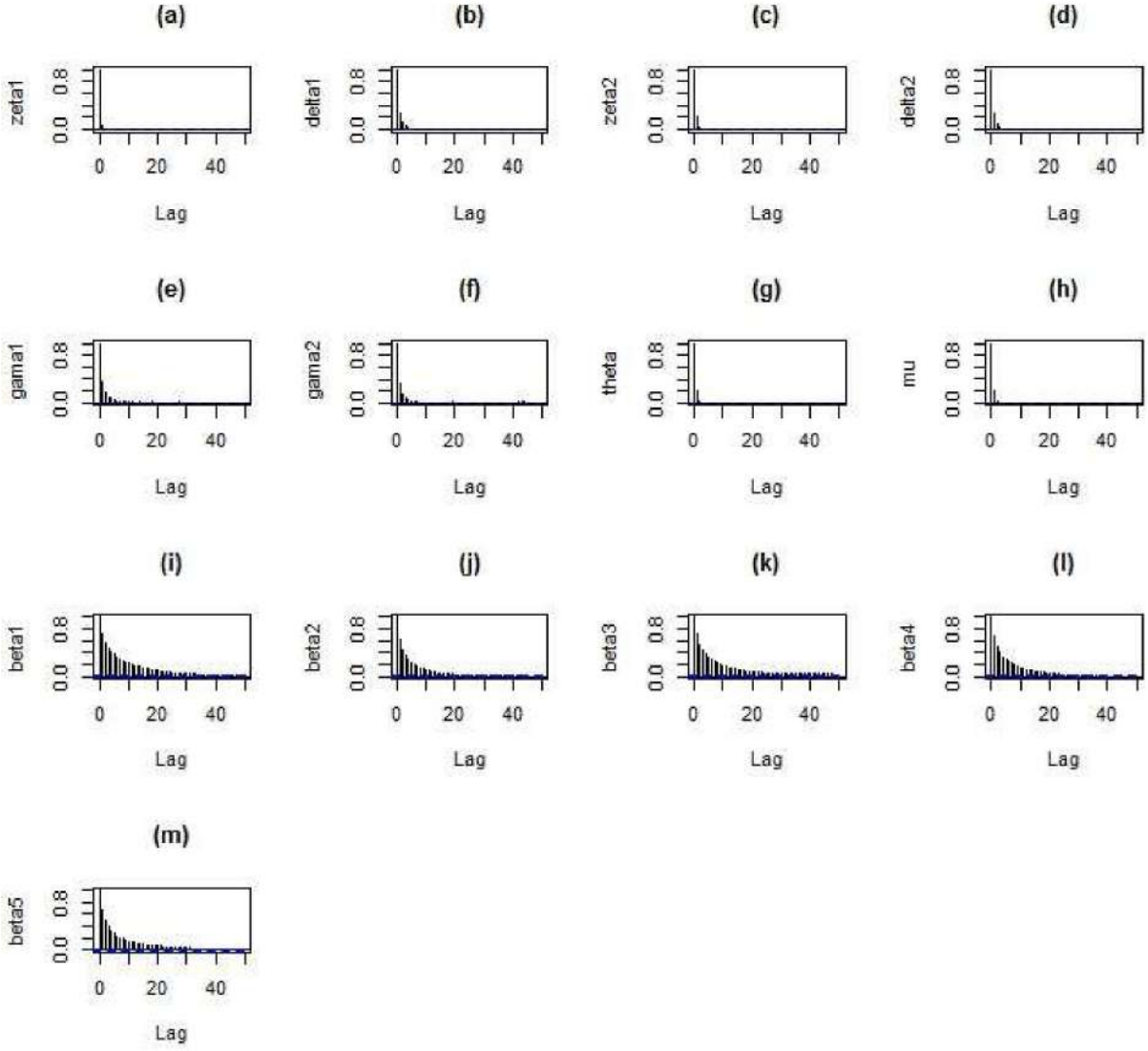


Figure 7: ACF plots for Model-I

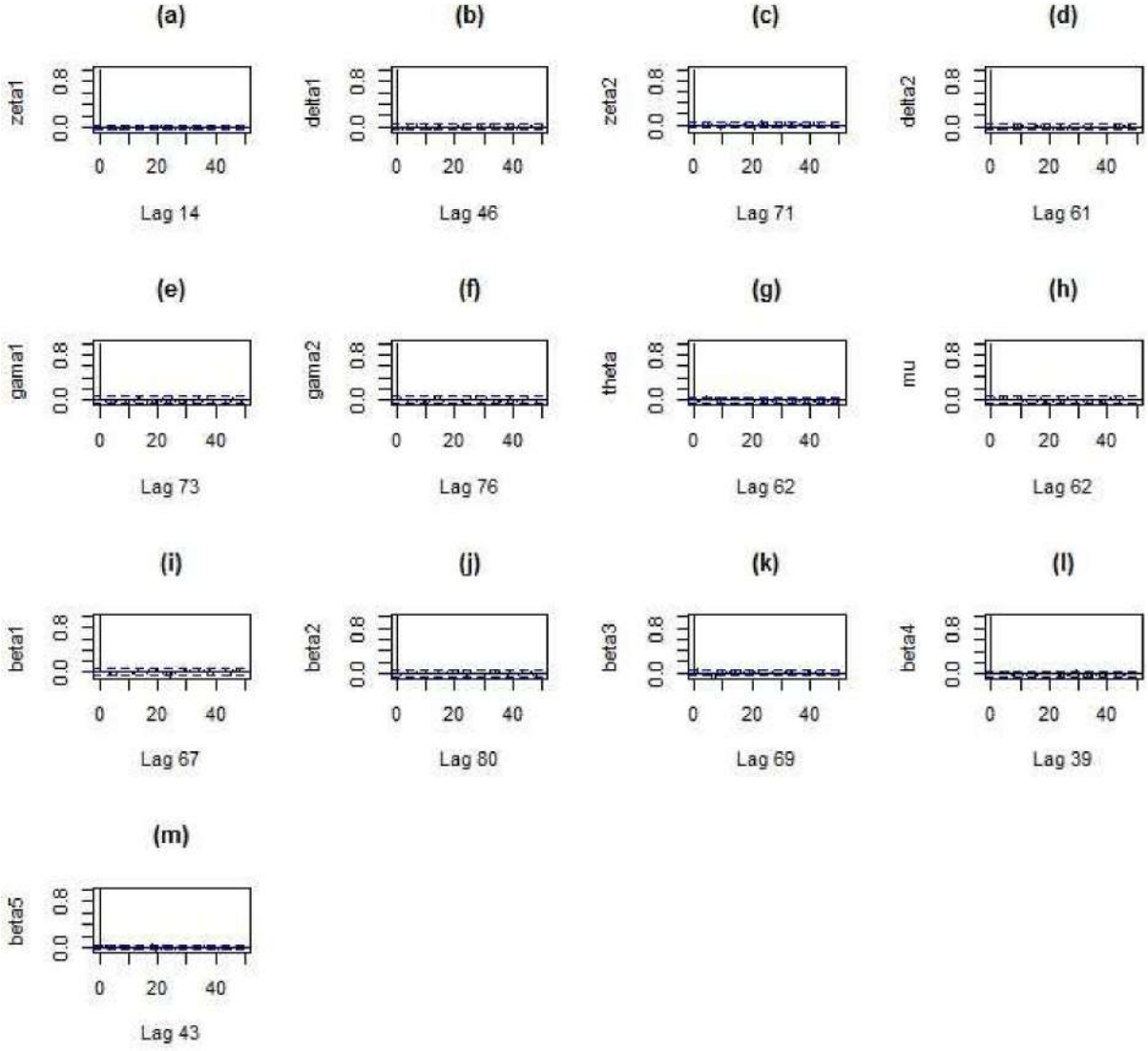


Figure 8: ACF plots After thinning for Model-I

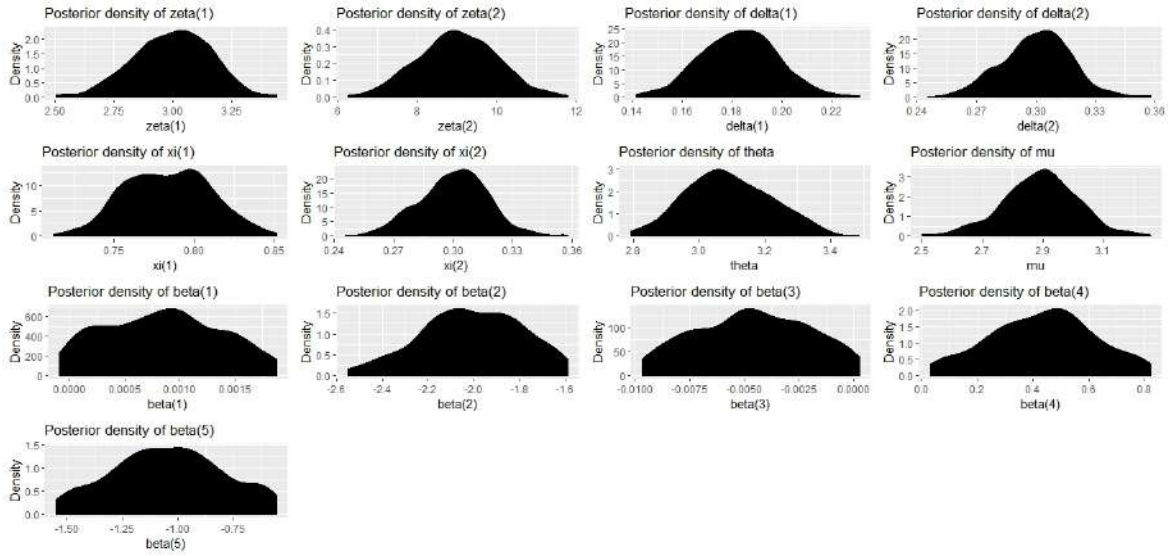


Figure 9: Posterior density plots for Model-I

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 63–76

Combinatorial Patterns of D-Optimal Weighing Designs Using a Spring Balance

Monica Pena Pardo and Jyotirmoy Sarkar

Department of Mathematical Sciences
Indiana University–Purdue University Indianapolis, Indiana, USA

Received: 15 May 2020; Revised: 02 August 2020; Accepted: 06 August 2020

Abstract

Given a spring balance that reports the true total weight of items plus a white noise of an unknown variance, which n subsets of n items will you weigh in order to estimate the true weights of each item with the highest possible precision?

For $n \leq 6$, we classify all D-optimal weighing designs according to the combinatorial patterns they exhibit (modulo permutation), we count the D-optimal designs exhibiting each pattern, and we explain how a D-optimal design for n items may arise out of a D-optimal design for $(n-1)$ items. For $n = 7, 11$ we exhibit D-optimal designs obtained from balanced incomplete block designs (BIBDs). We discuss some strategies to construct D-optimal designs of larger sizes, and pose some unsolved problems.

Key words: Design of experiments; Estimable parameter; Information matrix; Credibility region; Symmetric BIBD; Hadamard matrix.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

This story has a humble beginning in a classroom activity, then a surprising discovery, and finally an unexpected entry into the fascinating world of combinatorial designs.

While teaching a master's level first course in *Design of Experiments*, one day we brought to class four books, A–D, and asked the students: "If you want to estimate the true weight of each book, but you will only receive the true weight of each subset plus a white noise of an unknown variance, which four subsets of books will you weigh?"

Once a student would make his/her choices of any 4 out of the 15 subsets A, B, C, D, AB, AC, AD, BC, BD, BC, ABC, ABD, ACD, BCD, $ABCD$, we would give him/her the true weight of each chosen subset plus a white noise. We wanted to demonstrate that a haphazard choice of four subsets may not yield an estimate of $\boldsymbol{\mu}_{4\times 1}$, the vector of true weights of all four books; rather the subsets should be chosen with care, not only for estimating, but also for lowering the Euclidean volume of the estimated confidence region for $\boldsymbol{\mu}$. A more elaborate discussion on this classroom activity is given in the technical report, which we will happily share with the interested reader. Here we develop the main research ideas and their extensions to more general problems.

Corresponding Author: Jyotirmoy Sarkar
Email: jsarkar@iupui.edu

Initially, we had thought that only one design is optimal in producing a confidence region for $\boldsymbol{\mu}$ having the smallest Euclidean volume. When we tried to establish this optimal property of our preconceived choice, we had hoped to show that no other design had the same property. So, we carried out a complete search of all $\binom{15}{4} = 1365$ viable binary designs that render $\boldsymbol{\mu}$ estimable. Although it was somewhat counter-intuitive to us at that time, we were pleasantly surprised to find several other optimal designs (to be revealed in Section 3)!

Naturally, curiosity took a hold of us and we wanted to study the problem not just for 4 books, but for any n books, allowing selection of n out of $(2^n - 1)$ possible non-empty subsets. The rest of the paper documents what we found. Not only did we find multiple optimal designs in most cases, but also we categorized the optimal designs into distinct patterns (modulo permutation) and counted the number of optimal designs within each pattern. Additionally, we discovered connections between the optimal designs for n books and the optimal designs for $(n-1)$ books, for some values of n.

The origin of this optimal design problem can be traced back to almost a century ago when Yates introduced the experiment in 1935, which lead to a precise formulation by Hotelling in 1944. Since then, weighing designs have been thoroughly studied for both the spring balance problem and the chemical balance problem, with and without bias. This paper focuses on the spring balance problem where the scale has no bias.

In Section 2, we summarize the mathematical basis to estimate the true weights of the n books based on the experimental design. Among several reasonable criteria for determining the optimal design, we adopt D-optimality for our problem. In Section 3, we count the number of D-optimal designs of size $n \leq 5$, and classify them into distinct patterns (modulo permutation). In Section 4, we describe how sometimes a D-optimal design of size n is related to that of size $(n-1)$, illustrating the feature for $n \leq 6$. In Section 5, we discuss D-optimal designs of size $n = 4k - 1$ (for $k > 1$) using balanced incomplete block designs (BIBDs), and illustrate the same for $n = 7, 11$. Section 6 gives some strategies to construct D-optimal designs of larger sizes, and poses some unsolved problems, hoping to inspire young researchers to study this fascinating topic. All computations are done using the freeware R.

2. Mathematical Background

Let μ_j denote the true weight of item $j \in S \equiv \{1, 2, \ldots, n\}$. When any subset of items $S_i \subset S$ is suspended from a spring balance, the reported weight y_i equals $\sum_{\{j \in S_i\}} \mu_j + \epsilon_i$, where ϵ_i is a white noise; that is, it is normally distributed with mean 0 and unknown variance σ^2. The white noises are assumed independent. For $1 \leq i \leq n$, let us write $x_{ij} = 1$ if $j \in S_i$ and $x_{ij} = 0$ if $j \notin S_i$. Then the linear model, in matrix notation, can be written as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\epsilon}$, where the binary matrix $\mathbf{X} = (x_{ij})$ is called the design matrix for weighing with a spring balance. Since each of the n^2 elements of $\mathbf{X}$ can be chosen to be either 0 or 1, there are altogether 2^{2n} possible design matrices, of which only a subset of $\binom{2^n-1}{n}$ design matrices render $\boldsymbol{\mu}$ estimable. For details on the statistical model behind this estimation problem, see Banerjee (1975).

For our in-class book-weighing activity, intending to estimate each parameter μ_i with the highest possible precision, we prefer small values on the diagonal of the inverse of the information matrix $(\mathbf{X}^T\mathbf{X})^{-1}$, which correspond to the variances of the estimates $\hat{\mu}_i$. How-

ever, it may not be possible to minimize all diagonal elements simultaneously. We must thoughtfully choose an optimality criterion. Among the various notions of optimality discussed in Nishii (1993) and Pukelsheim (2006), we like D-optimality the most: It minimizes the determinant of $(\mathbf{X}^T\mathbf{X})^{-1}$; or equivalently, maximizes the determinant of $\mathbf{X}^T\mathbf{X}$. Moreover, when we adopt a Bayesian point of view, then under a non-informative prior distribution, the posterior credibility region of the *smallest Euclidean volume* turns out to be an ellipsoid with center $\hat{\boldsymbol{\mu}}$. Therefore, we adopt D-optimality as the criterion for choosing the best design.

3. Classifying D-optimal Designs into Patterns

Among the $\binom{2^n-1}{n}$ designs that render $\boldsymbol{\mu}$ estimable, many exhibit similar patterns. We will present these patterns by showing a characteristic illustrative design, along with its incidence matrix. However, before we construct and classify the different D-optimal weighing designs (DWDs) into patterns, it is helpful to know how many binary square matrices of degree n achieve the maximal determinant, and what is the value of that maximal determinant. Let us denote the determinant of $\mathbf{X}$ by $\det(\mathbf{X})$. Then $\det(\mathbf{X}^T\mathbf{X})^{-1} = |\det(\mathbf{X})|^{-2}$. Among the 2^{2n} binary matrices $\mathbf{X}$, what is $\delta = \max\det(\mathbf{X})$, and how many binary matrices achieve this maximum determinant δ?

In fact, the answers are well known for small n as summarized by Weisstein (no date) and presented in Table 1. Using this information, let us explain how to determine the number of DWDs (modulo permutation). Any binary design matrix $\mathbf{X}$ represents a weighing design, but a weighing design is invariant under row-permutation, since the order in which we weigh the subsets is irrelevant. As there are $n!$ permutations of the n rows, each weighing design can be represented by $n!$ binary matrices. Moreover, since the number of matrices achieving the maximal absolute value of determinant is twice the number of binary matrices achieving the maximal determinant, the total number of DWDs is given by the relation

$$\#\text{ D-optimal weighing designs} = \frac{\#\text{ D-optimal matrices}}{n!} = \frac{2\cdot\#\{\mathbf{X} : \det(\mathbf{X}) = \delta\}}{n!},$$

where D-optimal matrices are those matrices $\mathbf{X}$ that attain the maximal determinant δ in absolute value. We summarize the information in Table 1, where $\mathbf{X}$ is a binary matrix of size n. Next, Table 2 shows the number of DWDs per pattern.

Table 1: The number of D-optimal matrices (DMs) and weighing designs

n	$\max\det(\mathbf{X})$	$\frac{1}{2}$# DMs	# DWDs	# patterns
2	1	3	3	2
3	2	3	1	1
4	3	60	5	2
5	5	3600	60	3
6	9	529200	1470	7
7	32	75600	30	1
8	56	195955200	9720	?*
9	144	13716864000	75600	??*

*The undisclosed # patterns for $n = 8, 9$ are offered as exercise to the interested reader.

Table 2: The number of D-optimal weighing designs per pattern

Pattern	# DWDs
$D_{2,1}$	1
$D_{2,2}$	2
$D_{3,1}$	1
$D_{4,1}$	1
$D_{4,2}$	4
$D_{5,1}$	20
$D_{5,2}$	10
$D_{5,3}$	30

Pattern	# DWDs
$D_{6,1}$	360
$D_{6,2}$	180
$D_{6,3}$	180
$D_{6,4}$	180
$D_{6,5}$	360
$D_{6,6}$	180
$D_{6,7}$	30
$D_{7,1}$	30

Table 3 shows an illustrative design characterizing each pattern for $n = 2, 3, 4$, along with its incidence matrix, information matrix and inverse of information matrix.

Table 3: The patterns of D-optimal weighing designs for $n = 2, 3, 4$ illustrated

Pattern	Illustrative Design	Corresponding $\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$(\mathbf{X}^T\mathbf{X})^{-1}$
$D_{2,1}$	$\{P_1, P_2\}$	$\begin{pmatrix}1&0\\0&1\end{pmatrix}$	$\begin{pmatrix}1&0\\0&1\end{pmatrix}$	$\begin{pmatrix}1&0\\0&1\end{pmatrix}$
$D_{2,2}$	$\{P_1P_2, P_1\}$	$\begin{pmatrix}1&1\\1&0\end{pmatrix}$	$\begin{pmatrix}2&1\\1&1\end{pmatrix}$	$\begin{pmatrix}1&-1\\-1&2\end{pmatrix}$
$D_{3,1}$	$\{P_1P_2, P_1S_1, P_2S_1\}$	$\begin{pmatrix}1&1&0\\1&0&1\\0&1&1\end{pmatrix}$	$\begin{pmatrix}2&1&1\\1&2&1\\1&1&2\end{pmatrix}$	$\frac{1}{4}\begin{pmatrix}3&-1&-1\\-1&3&-1\\-1&-1&3\end{pmatrix}$
$D_{4,1}$	$\{P_1P_2Q_1,\ P_1P_2Q_2,$ $P_1Q_1Q_2,\ P_2Q_1Q_2\}$	$\begin{pmatrix}1&1&1&0\\1&1&0&1\\1&0&1&1\\0&1&1&1\end{pmatrix}$	$\begin{pmatrix}3&2&2&2\\&3&2&2\\&&3&2\\&&&3\end{pmatrix}$	$\frac{1}{9}\begin{pmatrix}7&-2&-2&-2\\&7&-2&-2\\&&7&-2\\&&&7\end{pmatrix}$
$D_{4,2}$	$\{Q_1Q_2Q_3,\ Q_1S_1,$ $Q_2S_1,\ Q_3S_1\}$	$\begin{pmatrix}1&1&1&0\\1&0&0&1\\0&1&0&1\\0&0&1&1\end{pmatrix}$	$\begin{pmatrix}2&1&1&1\\&2&1&1\\&&2&1\\&&&3\end{pmatrix}$	$\frac{1}{9}\begin{pmatrix}7&-2&-2&-1\\&7&-2&-1\\&&7&-1\\&&&4\end{pmatrix}$

Returning to our classroom book-weighing activity, we note that there are exactly five DWD's. The first pattern $D_{4,1}$ represents only one DWD, namely $\{ABC, ABD, ACD, BCD\}$, which we had anticipated beforehand; and the second pattern $D_{4,2}$ represents the following four DWD's whose discovery surprised us and propelled us into this research:

$$\{ABC, AD, BD, CD\}, \{ABD, AC, BC, CD\}, \{ACD, AB, BC, BD\}, \{BCD, AB, AC, AD\}.$$

For $n = 5$ and 6, we simply present the incidence matrices, the information matrices and the inverse information matrices in Table 4 below and in Table B1 in Annexure B, respectively, leaving the reader to find the corresponding illustrative designs. Details of these designs, along with ways to construct them, can be found in the technical report, which we will be happy to share with the interested reader, if needed.

Table 4: The three patterns of D-optimal designs for $n = 5$ illustrated

Pattern	Corresponding $\mathbf{X}$	$\mathbf{X}^T\mathbf{X}$	$25(\mathbf{X}^T\mathbf{X})^{-1}$
$D_{5,1}$	$\begin{pmatrix}1&0&0&0&1\\1&1&1&0&0\\1&1&0&1&0\\1&0&1&1&0\\0&1&1&1&1\end{pmatrix}$	$\begin{pmatrix}4&2&2&2&1\\&3&2&2&1\\&&3&2&1\\&&&3&1\\&&&&2\end{pmatrix}$	$\begin{pmatrix}11&-3&-3&-3&-1\\&19&-6&-6&-2\\&&19&-6&-2\\&&&19&-2\\&&&&16\end{pmatrix}$
$D_{5,2}$	$\begin{pmatrix}1&1&1&0&0\\1&1&0&1&0\\1&1&0&0&1\\1&0&1&1&1\\0&1&1&1&1\end{pmatrix}$	$\begin{pmatrix}4&3&2&2&2\\&4&2&2&2\\&&3&2&2\\&&&3&2\\&&&&3\end{pmatrix}$	$\begin{pmatrix}16&-9&-2&-2&-2\\&16&-2&-2&-2\\&&19&-6&-6\\&&&19&-6\\&&&&19\end{pmatrix}$
$D_{5,3}$	$\begin{pmatrix}1&1&1&0&0\\0&1&0&1&0\\0&1&0&0&1\\0&0&1&1&1\\1&0&0&1&1\end{pmatrix}$	$\begin{pmatrix}3&1&1&1&1\\&2&1&1&1\\&&2&1&1\\&&&3&2\\&&&&3\end{pmatrix}$	$\begin{pmatrix}11&-3&-3&-1&-1\\&19&-6&-2&-2\\&&19&-2&-2\\&&&16&-9\\&&&&16\end{pmatrix}$

4. Interrelations Between DWD's of Sizes $(n-1)$ and n

As n gets larger, patterns become more complicated. However, we have found that all designs of size n for $n = 2, \ldots, 6$ are related to at least one pattern of size $(n-1)$, and thus, can be constructed by simply adding a new letter to some words of a design of size $(n-1)$, and then adding a new word (or equivalently, by adding a row and column to a binary matrix representing the design of size $(n-1)$). Alternatively, we can think of a pattern or design of size n to have a D-optimal design of size $(n-1)$ embedded in it; or in terms of matrices, a D-optimal matrix of order n to have a minor of order $(n-1)$ which attains the maximal determinant for that order. However, this feature fails for $n = 7$ and $n = 11$.

Surely, all DWDs of size 2 embed in them a DWD of size 1. We illustrate how new designs are constructed from a lower order design for $n = 3, \ldots, 6$, by taking matrices of the previous order displayed in Section 3 and adding a new row and a new column to them.

4.1. From $n = 2$ to $n = 3$

Recall that for $n = 2$ and $n = 3$ some of the D-optimal matrices we found were

$$\mathbf{X}_{2,1} = \begin{pmatrix}1&0\\0&1\end{pmatrix}, \qquad \mathbf{X}_{2,2} = \begin{pmatrix}1&1\\1&0\end{pmatrix}, \qquad \mathbf{X}_{3,1} = \begin{pmatrix}1&1&0\\1&0&1\\0&1&1\end{pmatrix}$$

Note that $\mathbf{X}_{2,1}$ and $\mathbf{X}_{2,2}$ are both embedded in $\mathbf{X}_{3,1}$, as shown below:

$$\mathbf{X}_{3,1} = \left(\begin{array}{c|cc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ \hline 0 & 1 & 1 \end{array}\right) = \left(\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 0 & 1 \\ \hline 0 & 1 & 1 \end{array}\right)$$

4.2. From $n = 3$ to $n = 4$

From $D_{3,1}$, we can construct both, $D_{4,1}$ and $D_{4,2}$. Below we give an illustrative matrix for each pattern (permute the rows/columns to see $D_{3,1}$ embedded in $D_{4,2}$):

$$\mathbf{X}_{4,1} = \left(\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array}\right) \qquad \mathbf{X}_{4,2} = \left(\begin{array}{ccc|c} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 \end{array}\right)$$

4.3. From $n = 4$ to $n = 5$

Patterns $D_{5,1}$ and $D_{5,2}$ are constructed from either $D_{4,1}$ or $D_{4,2}$; but $D_{5,3}$ comes only from $D_{4,2}$. Refer to the technical report for details about how to construct these patterns. Below are illustrative matrices for each of these cases. Within the first two cases, permute the rows/columns to see that the two incidence matrices represent the same pattern.

1) Illustrative matrices for $D_{5,1}$ coming from $\mathbf{X}_{4,1}$ and $\mathbf{X}_{4,2}$, respectively:

$$\left(\begin{array}{cccc|c} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 1 \end{array}\right) \qquad \left(\begin{array}{cccc|c} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ \hline 1 & 1 & 1 & 0 & 1 \end{array}\right)$$

2) Illustrative matrices for $D_{5,2}$ coming from $\mathbf{X}_{4,1}$ and $\mathbf{X}_{4,2}$, respectively:

$$\left(\begin{array}{cccc|c} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 & 1 \end{array}\right) \qquad \left(\begin{array}{cccc|c} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 0 \end{array}\right)$$

3) An illustrative matrix for $D_{5,3}$ coming from $D_{4,2}$:

$$\left(\begin{array}{cccc|c} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ \hline 0 & 1 & 1 & 0 & 1 \end{array}\right)$$

4.4. From $n = 5$ to $n = 6$

Patterns $D_{6,1}$ and $D_{6,2}$ come from $D_{5,1}$, patterns $D_{6,3}$ and $D_{6,4}$ come from $D_{5,2}$, and patterns $D_{6,5}$, $D_{6,6}$ and $D_{6,7}$ come from $D_{5,3}$. The matrices representing these designs can be found in Table B1 in Annexure B. We invite the astute reader to contemplate how to obtain these extensions from size 5 to size 6, how to count the number of DWDs, and how to construct the different patterns. All of these topics and more are thoroughly addressed in the technical report, which we will gladly share, if needed.

What we found absolutely delightful, we offer as a gift to our dear readers: We present a D-optimal matrix of size 6, coming from a D-optimal matrix of size 5, coming from a D-optimal matrix of size 4, coming from ... you get the idea. Here it is:

$$\left(\begin{array}{cccccc} 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right)$$

Without first memorizing it, can you reconstruct this DWD?

5. DWD's for Cases $n = 4k - 1$ where $k = 2, 3, \ldots$

These special cases are related to a well-studied group of chemical balance weighing designs. A square sign matrix (all whose elements are -1 and 1) of size n that attains the maximal determinant is known as a Hadamard matrix. The order of a Hadamard matrix is $n = 1, 2$ or $n = 4k$ for $k \geq 1$, and its determinant is $n^{n/2}$ (see Brenner, 1972). Mood (1946) and Banerjee (1975) show that there is a one-to-one correspondence between Hadamard matrices of size n and square binary matrices of size $(n - 1)$ with maximal determinant. Thus, the number of inequivalent Hadamard matrices of size n is also the number of patterns for DWD's of size $(n - 1)$.

The existence of Hadamard matrices is known for all n divisible by 4 up to $n < 668$, thereby implying the existance of D-optimal binary matrices of size $n = 4k - 1 < 667$. Moreover, Raghavarao (1971) provides methods for constructing Hadamard matrices of order $n \leq 100$, and the number of inequivalent Hadamard matrices is known for orders $n \leq 32$, as given in the On-Line Encyclopedia of Integer Sequences (OEIS).

Since square binary matrices with maximal determinants can be constructed from Hadamard matrices, as shown by Mood (1946) and as found in Stinson (2004), we can construct DWDs of size $n = 4k - 1$, $k = 1, 2, ...$ starting from Hadamard matrices of size $n = 4k$. Here, we illustrate two of them, $n = 7$ and $n = 11$, leaving the rest to the reader.

5.1. Case $n = 7$

There are more reasons that make this case very, very special. First, there is only one possible pattern, yielding 30 possible DWDs. Note that both the number of patterns and the number of DWDs is much smaller than the previous case of $n = 6$. Moreover, the only

pattern for $n = 7$ does not come from any of the patterns for $n = 6$, as proven by Williamson (1946). Surprisingly, that single pattern for $n = 7$ is rather easy to construct: First, find the incidence matrix of the symmetric BIBD$(7, 3, 1)$ associated with the Hadamard matrix of size 8, and then take its complement, which gives a symmetric BIBD$(7, 4, 2)$.

Recall that the symmetric BIBD$(7, 3, 1)$ can be found from the Fano plane shown below. To construct a particular design, we label the vertices of the Fano plane with distinct letters and make words consisting of the three letters on each line on the graph. (Here, the central circle also counts as a line.) There are 30 distinct ways to label of the Fano plane, not counting rotation and reflection symmetries. By taking the complements of each such labelled symmetric BIBD$(7, 3, 1)$, we obtain 30 possible DWDs of size $n = 7$.

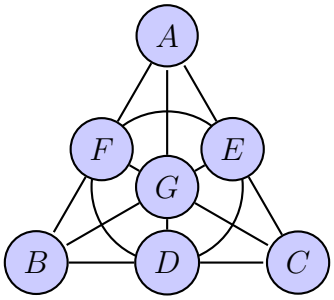


Figure 1: The Fano plane yields the symmetric BIBD$(7, 3, 1)$.

For example, using the labeling shown in Figure 1 above, we obtain the symmetric BIBD $\{ABF, ACE, ADG, BCD, BEG, CFG, DEF\}$. Thereafter, its complement yields the following DWD of size $n = 7$:

$$\{CDEG, BDFG, BCEF, AEFG, ACDF, ABDE, ABCG\}.$$

Another DWD is given below along with its corresponding matrix. Can you find the labeling of the Fano plane that leads to this DWD?

Illustrative Design: $\{ABCD, ABEF, ACEG, ADFG, BCFG, BDEG, CDEF\}$.

$$\mathbf{X}_7 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \qquad \mathbf{X}_7^T\mathbf{X}_7 = \begin{pmatrix} 4 & 2 & 2 & 2 & 2 & 2 & 2 \\ & 4 & 2 & 2 & 2 & 2 & 2 \\ & & 4 & 2 & 2 & 2 & 2 \\ & & & 4 & 2 & 2 & 2 \\ & & & & 4 & 2 & 2 \\ & & & & & 4 & 2 \\ & & & & & & 4 \end{pmatrix},$$

$$(\mathbf{X}_7^T\mathbf{X}_7)^{-1} = \frac{1}{16}\begin{pmatrix} 7 & -1 & -1 & -1 & -1 & -1 & -1 \\ & 7 & -1 & -1 & -1 & -1 & -1 \\ & & 7 & -1 & -1 & -1 & -1 \\ & & & 7 & -1 & -1 & -1 \\ & & & & 7 & -1 & -1 \\ & & & & & 7 & -1 \\ & & & & & & 7 \end{pmatrix}.$$

5.2. Case $n = 11$

According to the OEIS, there is only one distinct (up to permutation of rows and columns) Hadamard matrix of order 12, implying that there is only one possible pattern of DWDs for $n = 11$. This pattern can be found by using the Paley biplane (shown below), which leads to a symmetric BIBD$(11, 5, 2)$, as explained in the next paragraph. The Paley biplane can be labelled in 60,480 distinct ways not counting symmetries.

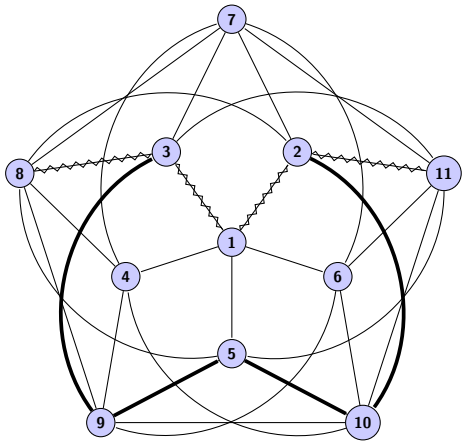


Figure 2: The Paley biplane yields the symmetric BIBD$(11, 5, 2)$.

After the Paley biplane is labelled, Brown (2004) explains that each of the 11 subsets (rows of the incidence matrix of the BIBD) can be found by traveling on three types of paths in the graph: The first type, shown in bold line gives rise to 5 subsets, via a 1/5 rotation about the center. Similarly, the second type of path, shown with the zigzag lines gives rise to 5 more subsets, via a 1/5 rotation about the center. Finally, any four edges of the outer pentagon constitute the last subset needed to construct the symmetric BIBD$(11, 5, 2)$.

Thereafter, we take its complement, a symmetric BIBD$(11, 6, 3)$, to obtain the pattern of DWDs for $n = 11$. We give the incidence matrix of one such DWD; but leave to the avid reader the task of finding other such designs by relabelling the Paley biplane.

$$\mathbf{X}_{11} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Note that the above matrix has determinant 1458, which means that it achieves the maximal determinant for a square binary matrix of order 11. Alternatively, a matrix for this

design can be found by using the standardized Hadamard matrix of order 12 and deleting its first row and column, or by using the Paley construction. Moreover, we know from Williamson (1946) that this pattern cannot be constructed using a DWD of size $n = 10$. In fact, there is no DWD of order $n = 8, 9$ or 10 embedded in this design.

We constructed DWDs of sizes $n = 7$ and $n = 11$ using symmetric BIBDs. This is no coincidence: Raghavarao (1971) proves that when estimating the weights of $n = 4k - 1$ objects using exactly n weighings, the incidence matrix of a symmetric BIBD$(4k - 1, 2k, k)$ is D-optimal as a spring balance weighing design of size $n = 4k - 1$.

This result can be strengthened. Mood (1946) proves that there is a one-to-one correspondence between D-optimal binary matrices and Hadamard matrices. Additionally, Stinson (2004) states that for $k > 1$, there exists a Hadamard matrix of order $4k$ if and only if there exists a symmetric BIBD$(4k - 1, 2k - 1, k - 1)$. Hence, the following lemma holds.

Lemma 1. For $k > 1$, there exists a D-optimal binary matrix of size $n = 4k - 1$ if and only if there exists a symmetric BIBD$(4k - 1, 2k, k)$.

Lemma 1 aids us in counting DWDs associated with each pattern (or equivalently, with each symmetric BIBD). For $n = 4k - 1$, the number of DWDs associated with a particular symmetric BIBD is given by $n!$ divided by the number of symmetries of the symmetric BIBD(v, k, λ); that is, the number of permutations of the v treatments (columns) that simultaneously permute the blocks (rows). (This number is also known as the order of the automorphism group of the design). This explains our counting of DWDs for $n = 7$ and $n = 11$:

$$\frac{7!}{168} = 30 \text{ DWDs} \qquad \frac{11!}{660} = 60480 \text{ DWDs}$$

In other words, there were 7! and 11! ways to relabel the Fano plane and the Paley biplane, respectively; but to remove all duplicates, we divided by 168 and 660—the number of symmetries of the plane/biplane.

6. Future Work

We have found, classified and counted all D-optimal weighing designs of sizes $n = 2, \ldots 7$. One may now consider other types of optimality, such as A-optimality and E-optimality, mentioned in Nishii (1993) and Pukelsheim (2006), to choose a preferred design depending on the research goal. We leave this task to the interested reader, aiding them with the following table of traces of the $(\mathbf{X}^T\mathbf{X})^{-1}$ matrices that we used in Sections 3 and 4. The reader can also find a thorough discussion of this topic in Shah and Sinha (1989).

A natural extension of this work is to study designs of sizes $n > 7$. As mentioned in Section 5, designs of sizes $n = 4k - 1$ are very well studied, given the study of Hadamard matrices and the association between Hadamard matrices and D-optimal binary matrices.

Moreover, to construct designs of sizes $n > 7$ where $n \neq 4k - 1$, Bose and Nair (1939) and Banerjee (1952) use some partial BIBDs as weighing designs. Also, one can try to extend designs of size $(n - 1)$, following strategies used in Section 4, and further explained in the technical report. For instance, starting with the matrix $\mathbf{X}_7$ in Section 5 illustrating the case

Table 5: Traces of $(\mathbf{X}^T\mathbf{X})^{-1}$ matrices for D-optimal weighing designs

Pattern	$tr((\mathbf{X}^T\mathbf{X})^{-1})$
$D_{2,1}$	2
$D_{2,2}$	3
$D_{3,1}$	9/4
$D_{4,1}$	28/9
$D_{4,2}$	25/9
$D_{5,1}$	84/25
$D_{5,2}$	89/25
$D_{5,3}$	81/25

Pattern	$tr((\mathbf{X}^T X)^{-1})$
$D_{6,1}$	298/81
$D_{6,2}$	310/81
$D_{6,3}$	309/81
$D_{6,4}$	309/81
$D_{6,5}$	295/81
$D_{6,6}$	319/81
$D_{6,7}$	306/81
$D_{7,1}$	49/16

for $n = 7$, and adding an extra row and column leads to a matrix of order $n = 8$ with determinant 56, which, as reported in Table 1, is the largest determinant a binary matrix of this size can attain.

$$\mathbf{X}_8 = \left(\begin{array}{ccccccc:c} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ \hdashline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{array}\right)$$

As mentioned in Section 3, we leave to the reader the task of finding all patterns and counting all DWDs for cases $n = 8, 9$. We empower them with yet another potent idea described below. But first let us explain a geometric representation of a design matrix: We construct an $n \times n$ grid of unit squares; color a unit square if the corresponding matrix entry is 1, but leave the unit square colorless if the entry is 0. We may also partition the grid suitably to enhance the pattern. For example, designs $D_{6,3}$ and $D_{6,4}$, after suitable rearrangements of rows and columns, are represented as shown in Figure 3.

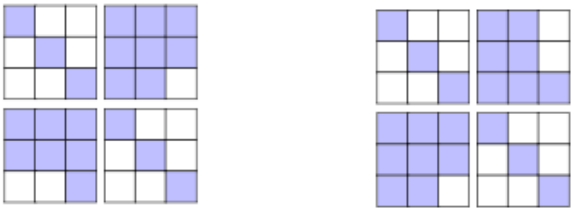


Figure 3: These geometric patterns for $D_{6,3}$ and $D_{6,4}$ are obtained by coloring and spacing unit squares.

Having given such a geometric representation of a design matrix, let us describe another type of extension from a smaller size design to a larger size design that is worth exploring: Carefully observe the patterns themselves, and replicate them for a larger n. For example, the geometric representation of $\mathbf{X}_7$, shown in Figure 4, reveals a visual pattern, which may be time-consuming to express verbally. Instead, we invite the astute reader to imitate the same pattern for a 9×9 grid. The resulting picture is shown in Figure 5 in Annexure A. Please do not peek at it too early, lest you miss the joy of discovery.

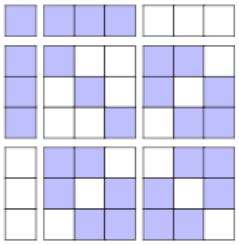


Figure 4: Here is a geometric pattern for $n = 7$. Can you imitate it for $n = 9$?

Figure 5 serves as a design matrix for the case $n = 9$, and it has determinant 144, which is the largest determinant a binary matrix of this size can attain. Although performing such extensions multiple times may not result in D-optimal matrices of larger orders (for example, this pattern does not work for $n = 11$ or $n = 13$), the idea is worth testing with other patterns of different sizes.

Let us briefly mention yet another avenue of research involving designs that do not necessarily involve square matrices. In this case, to estimate all parameters, we allow *more weighings* than objects to weigh (see Mood, 1946, Banerjee 1975, or Neubauer *et al.*, 1998 for designs corresponding to four and five objects). While the investment is higher than absolutely necessary in terms of increased number of weighings, the added benefit is that the estimates have smaller standard errors and one can also estimate the error variance σ^2. For an introduction to a very general description of weighing designs in this direction and related results, we refer to Raghavarao (1971) and Shah and Sinha (1989).

Acknowledgments

We thank our students for participating in the classroom activity which eventually led to this paper. We are grateful to an anonymous referee who suggested many improvements and furthermore generously listed many useful references. Finally, we thank the Chair Editor for his encouragement, guidance and counsel.

References

Banerjee, K. S. (1952). Weighing designs and partially balanced incomplete blocks. *Calcutta Statistical Association Bulletin*, **4(1)**, 36–38.

Banerjee, K. S. (1975). *Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics*. Marcel Dekker Inc., New York.

Bose, R. C. and Nair, K. R. (1939). Partially balanced incomplete block designs. *Sankhyā: The Indian Journal of Statistics*, **4(3)**, 337–372.

Brown, E. (2004). The fabulous $(11, 5, 2)$ biplane. *Mathematics Magazine*, **11**, 87–100.

Brenner, J. and Cummings, L. (1972). The Hadamard maximum determinant problem. *The American Mathematical Monthly*, **79(6)**, 626–630. doi:10.2307/2317092

Hotelling, H. (1944). Some improvements in weighing and other experimental techniques. *Annals of Mathematical Statistics* **15(3)**, 297–306.

Mood, A. M. (1946). On Hotelling's weighing problem. *Annals of Mathematical Statistics*, **17(4)**, 432–446.

Neubauer, M. G., Watkins, W. and Zeitlin, J. (1998). D-optimal weighing designs for four and five objects. *The Electronic Journal of Linear Algebra*, **4**, 48–72.

Nishii, R. (1993). Optimality of experimental designs. *Discrete Mathematics*, **116**, 209–225.
OEIS Foundation Inc. (2020). Number of Hadamard Matrices of Order 4n. *The On-Line Encyclopedia of Integer Sequences*. http://oeis.org/A007299.
Pena Pardo, M. and Sarkar, J. (2020). *D-Optimal Weighing Designs Using an Unbiased Spring Balance*. Technical Report, Mathematical Sciences, Indiana University-Purdue University Indianapolis.
Pukelsheim, F. (2006). *Optimal Design of Experiments*. Classics in Applied Mathematics Series, Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898719109
R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. Wiley, New York (ISBN: 0-486-65685-3).
Shah, K. R. and Sinha, B. K. (1989). *Theory of Optimal Designs. Lecture Notes in Statistics*, **54**. Springer-Verlag, New York (ISBN 978-0-387-96991-6).
Stinson, D. R. (2004). *Combinatorial Designs: Constructions and Analysis*. Springer-Verlag, New York (ISBN: 978-0-387-95487-5).
Weisstein, E. W. (no date) Hadamard's Maximum Determinant Problem. *MathWorld–A Wolfram Web Resource*. Retrieved 17 April 2020. https://mathworld.wolfram.com/HadamardsMaximumDeterminantProblem.html
Weisstein, E. W. (no date) Hadamard Matrix. *MathWorld–A Wolfram Web Resource*. Retrieved 17 April 2020. https://mathworld.wolfram.com/HadamardMatrix.html
Weisstein, E. W. (no date) Paley Construction. *MathWorld–A Wolfram Web Resource*. Retrieved 17 April 2020. https://mathworld.wolfram.com/HadamardsMaximumDeterminantProblem.html
Williamson, J. (1946). Determinants whose elements are 0 and 1. *The American Mathematical Monthly*, **53(8)**, 427–434. doi:10.2307/2306240
Wikipedia Contributors (no date). Hadamard's Maximal Determinant Problem. *Wikipedia, The Free Encyclopedia*. Retrieved 17 April 2020.
Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society* **2(2)**, 181–247.

ANNEXURE A

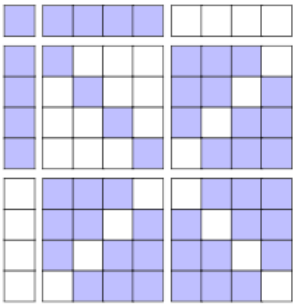


Figure 5: This geometric pattern for $n = 9$ imitates the pattern shown in Figure 4.

ANNEXURE B

Table B1: The seven patterns of D-optimal designs for $n = 6$ illustrated

Pattern	Illustrative **X**	$\mathbf{X}^T\mathbf{X}$	$81(\mathbf{X}^T\mathbf{X})^{-1}$
$D_{6,1}$	$\left(\begin{array}{cccccc} 1&1&1&0&0&0\\ 1&1&0&1&0&0\\ 1&0&1&1&0&1\\ 0&1&1&1&1&0\\ 1&0&0&0&1&1\\ \hdashline 0&1&0&0&0&1 \end{array}\right)$	$\begin{pmatrix} 4&2&2&2&1&2\\ &4&2&2&1&1\\ &&3&2&1&1\\ &&&3&1&1\\ &&&&2&1\\ &&&&&3 \end{pmatrix}$	$\begin{pmatrix} 45&-6&-12&-12&3&-21\\ &35&-11&-11&-4&1\\ &&59&-22&-8&2\\ &&&59&-8&2\\ &&&&56&-14\\ &&&&&44 \end{pmatrix}$
$D_{6,2}$	$\left(\begin{array}{cccccc} 1&1&1&0&0&0\\ 1&1&0&1&0&0\\ 1&0&1&1&0&1\\ 0&1&1&1&1&1\\ 1&0&0&0&1&1\\ \hdashline 1&0&1&1&1&0 \end{array}\right)$	$\begin{pmatrix} 5&2&3&3&2&2\\ &3&2&2&1&1\\ &&4&3&2&2\\ &&&4&2&2\\ &&&&3&2\\ &&&&&3 \end{pmatrix}$	$\begin{pmatrix} 35&-6&-11&-11&-4&-4\\ &45&-12&-12&3&3\\ &&59&-22&-8&-8\\ &&&59&-8&-8\\ &&&&56&-25\\ &&&&&56 \end{pmatrix}$
$D_{6,3}$	$\left(\begin{array}{cccccc} 1&1&0&0&1&1\\ 1&0&1&0&1&1\\ 1&0&0&1&1&0\\ 0&1&1&1&1&0\\ 1&1&1&1&0&0\\ \hdashline 0&0&0&1&0&1 \end{array}\right)$	$\begin{pmatrix} 4&2&2&2&3&2\\ &3&2&2&2&1\\ &&3&2&2&1\\ &&&4&2&1\\ &&&&4&2\\ &&&&&3 \end{pmatrix}$	$\begin{pmatrix} 56&-8&-8&-4&-25&-14\\ &59&-22&-11&-8&2\\ &&59&-11&-8&2\\ &&&35&-4&1\\ &&&&56&-14\\ &&&&&44 \end{pmatrix}$
$D_{6,4}$	$\left(\begin{array}{cccccc} 1&1&0&0&1&0\\ 1&0&1&0&1&0\\ 1&0&0&1&1&1\\ 0&1&1&1&1&0\\ 1&1&1&1&0&0\\ \hdashline 0&1&1&0&0&1 \end{array}\right)$	$\begin{pmatrix} 4&2&2&2&3&1\\ &4&3&2&2&1\\ &&4&2&2&1\\ &&&3&2&1\\ &&&&4&1\\ &&&&&2 \end{pmatrix}$	$3\begin{pmatrix} 17&-1&-1&-3&-10&-1\\ &17&-10&-3&-1&-1\\ &&17&-3&-1&-1\\ &&&18&-3&-3\\ &&&&17&-1\\ &&&&&17 \end{pmatrix}$
$D_{6,5}$	$\left(\begin{array}{cccccc} 1&1&0&0&1&0\\ 1&0&1&0&0&0\\ 1&0&0&1&0&1\\ 0&1&1&1&0&0\\ 0&0&1&1&1&0\\ \hdashline 0&1&1&0&1&1 \end{array}\right)$	$\begin{pmatrix} 3&1&1&1&1&1\\ &3&2&1&2&1\\ &&4&2&2&1\\ &&&3&1&1\\ &&&&3&1\\ &&&&&2 \end{pmatrix}$	$\begin{pmatrix} 35&-4&1&-6&-4&-11\\ &56&-14&3&-25&-8\\ &&44&-21&-14&2\\ &&&45&3&-12\\ &&&&56&-8\\ &&&&&59 \end{pmatrix}$
$D_{6,6}$	$\left(\begin{array}{cccccc} 1&1&0&0&1&1\\ 1&0&1&0&0&1\\ 1&0&0&1&0&1\\ 0&1&1&1&0&1\\ 0&0&1&1&1&1\\ \hdashline 1&1&1&1&1&0 \end{array}\right)$	$\begin{pmatrix} 4&2&2&2&2&3\\ &3&2&2&2&2\\ &&4&3&2&3\\ &&&4&2&3\\ &&&&3&2\\ &&&&&5 \end{pmatrix}$	$\begin{pmatrix} 45&-12&3&3&-12&-21\\ &59&-8&-8&-22&2\\ &&56&-25&-8&-14\\ &&&56&-8&-14\\ &&&&59&2\\ &&&&&44 \end{pmatrix}$
$D_{6,7}$	$\left(\begin{array}{cccccc} 1&1&0&0&1&0\\ 1&0&1&0&0&1\\ 1&0&0&1&0&1\\ 0&1&1&1&0&0\\ 0&0&1&1&1&0\\ \hdashline 0&1&0&0&1&1 \end{array}\right)$	$\begin{pmatrix} 3&1&1&1&1&2\\ &3&1&1&2&1\\ &&3&2&1&1\\ &&&3&1&1\\ &&&&3&1\\ &&&&&3 \end{pmatrix}$	$3\begin{pmatrix} 17&-1&-1&-1&-1&-10\\ &17&-1&-1&-10&-1\\ &&17&-10&-1&-1\\ &&&17&-1&-1\\ &&&&17&-1\\ &&&&&17 \end{pmatrix}$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 77-87

Depicting Bivariate Relationship with a Gaussian Ellipse

Jyotirmoy Sarkar[1] **and Mamunur Rashid**[2]
[1]*Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, USA*
[2]*Department of Mathematics, DePauw University, Greencastle, Indiana, USA*

Received: 15 June 2020; Revised: 01 August 2020; Accepted: 08 August 2020

Abstract

For data on two continuous variables, how should one depict the summary statistics (means, SDs, correlation coefficient, coefficient of determination, regression lines) so that their values can be read off easily from the depiction and potential outliers can be flagged also? We propose the Gaussian covariance ellipse as an answer that will benefit all users of statistics.

Key words: Five-number summary; Variance; Least squares lines; Bivariate normal distribution; x-, y- and regression outliers.

AMS Subject Classification: 62H20

1. Introduction

Appropriate graphical representation of data is necessary for easy comprehension of underlying information. For each type of variable and for each objective, one must choose the correct graph to depict the data. In this paper, we focus only on *quantitative variables* which take values on a continuous scale; that is, even though measurement limitations may force us (and ease of comprehension may prompt us) to report the value correct to an integer or up to a few decimal places, we recognize that finer values are surely possible. Some quantitative variables are measured only on a *difference scale*, where the difference between two values has a meaningful interpretation, but not their ratio; and other quantitative variables are measured on an *interval* or *ratio scale*, where the ratio of two values has a proper physical interpretation.

The objective of this paper is to depict the summary statistics of two quantitative variables that are related via the linear regression model or that exhibit a bivariate normal distribution. Section 2 identifies some commonly used bivariate statistics and poses the problem of depicting them efficiently. Section 3 depicts bivariate linear association for standardized data using a *correlation ellipse*; and Section 4 depicts bivariate summary statistics for raw data using a *covariance ellipse*. Section 5 highlights the sufficient statistics from which other bivariate summary statistics can be reconstructed; and Section 6 further reduces the sufficient statistics. Section 7 concludes the paper, interprets the covariance ellipse and poses some directions of future research.

Corresponding Author: Jyotirmoy Sarkar
Email: jsarkar@iupui.edu

2. Depicting Bivariate Summary Statistics: Statement of the Problem

Methods are well-known for depicting summary statistics of a quantitative variable. We refer the reader to Maverick (1932), Embse and Engebretsen (1996), and Sarkar and Rashid (2016, 2019). Also, above (or below) a dot plot or a histogram one can easily superimpose an arrow, whose tail shows the mean and length the SD. See, for example, Devore (2015) and Rashid and Sarkar (2018). Likewise, to depict the interrelations between two quantitative variables, the commonly used scatter plot can be augmented by the five-number-summary, the mean and the SD of each variable in the margins; that is, when the scatter plot is projected along each coordinate axis, the corresponding dot plots can be summarized using univariate methods. See an example given in Figure 1 with details found in Sarkar and Rashid (2020).

Throughout the paper we illustrate some visualization techniques using the following example involving the midterm exam score (x) and the final exam score (y) of 23 students in an *Introduction to Statistics* course.

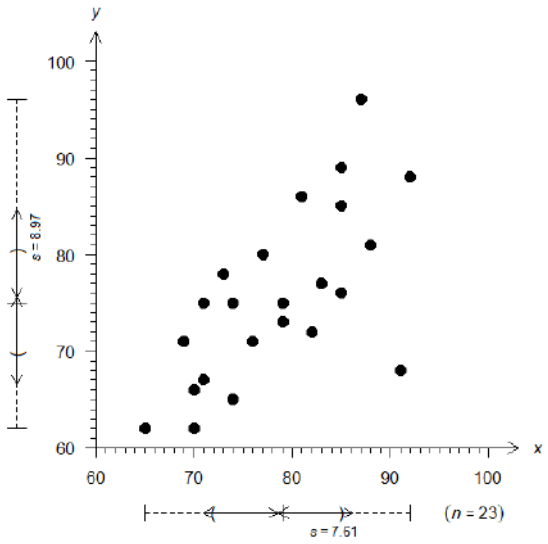


Figure 1: The midterm exam (x) and the final exam (y) scores of 23 students, with the five-number-summary, mean and SD of each score shown in the margins

To the astute reader we pose a quiz: "Projection of a scatter plot in which direction will cause the corresponding dot plot to exhibit the largest (or the smallest) SD? Alternatively, which linear combination of x and y has the largest (or the smallest) SD?" We urge the readers to jot down their answers before reading further. Our answer is given later in this paper.

Frequently used bivariate summary statistics include the correlation coefficient r, the least squares regression line $\hat{y}$ as a function of x, the *inverse* least squares regression line $\hat{x}$ as a function of y, and the coefficient of determination r^2. How should these bivariate statistics be depicted so that their numerical values can be easily read off from the scales of the axes?

As a solution to this question, we propose to draw the Gaussian covariance ellipse that fits inside the c-SD rectangle given by the c-SD boundaries $y = \bar{y} \pm cs_y$ and $x = \bar{x} \pm cs_x$. The diagonals of this rectangle intersect at the mean vector $(\bar{x}, \bar{y})$. We shall exhibit the Gaussian covariance ellipse in Section 4. But first, in Section 3, let us look at the Gaussian correlation ellipse, which only considers the standardized variables. It strips out the central location vector

$(\bar{x}, \bar{y})$, suppresses the scale factors s_x and s_y, and focuses on the correlation coefficient r, the coefficient of determination r^2 and the two regression lines (in standardized units).

3. Gaussian Correlation Ellipse

Let us first focus on the correlation coefficient r. To do so, we replace the variables by their standardized versions: Replace x by $\tilde{x} = (x - \bar{x})/s_x$ and y by $\tilde{y} = (y - \bar{y})/s_y$. Consequently, the mean vector for $(\tilde{x}, \tilde{y})$ is (0, 0), and $s_{\tilde{y}} = s_{\tilde{x}} = 1$. In particular, the c-SD rectangle for $(\tilde{x}, \tilde{y})$ is a square (so long as the scales of the two axes in the diagram are chosen to be the same). We inscribe in this square the c-SD Gaussian correlation ellipse whose two axes pass through $(0, 0)$ and have slopes 1 and -1, and is internally tangential to the c-SD square at exactly four points: bottommost point $B = (-rc, -c)$, topmost point $T = (rc, c)$, leftmost point $L = (-c, -rc)$ and rightmost point $R = (c, rc)$. Then LR is the regression line $\hat{\tilde{y}} = r\tilde{x}$ line, BT is the inverse regression line $\hat{\tilde{x}} = r\tilde{y}$ line. Furthermore, these two regression lines LR and BT intersect at the center (0, 0) of the c-SD correlation ellipse, which is also the point of intersection of the two diagonals of the c-SD square and the *center of gravity* of the scatter plot of standardized variables $(\tilde{x}, \tilde{y})$.

For the example data, shown in Figure 1, after standardizing the scores, Figure 2 depicts the *standard* correlation ellipse, where we have chosen $c = 1$, so that 39.35% of the data are expected to fall inside. We will say more about the choice of c towards the end of Section 4.

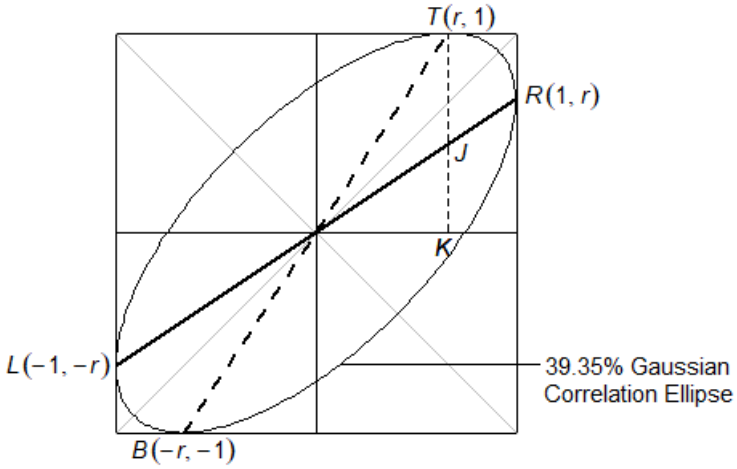


Figure 2: The 1-SD square and the 1-SD correlation ellipse for midterm and final exam scores of 23 students in an *Introduction to Statistics* course

Based on the standardized variables, the standard (that is, with $c = 1$) correlation ellipse is centered at the origin, has one axis with half-length a (to be determined in the next paragraph) in the direction $(1, 1)$ starting from the center, and has the other (orthogonal) axis with half-length b (also to be determined shortly) in the direction $(1, -1)$ starting from the origin. After normalizing the direction vectors (that is, dividing each vector by $\sqrt{2}$), the standard correlation ellipse can be described by the equation

$$\frac{(\tilde{x}+\tilde{y})^2}{2a^2} + \frac{(\tilde{x}-\tilde{y})^2}{2b^2} = 1. \qquad (1)$$

Next, focusing on the upper $\tilde{y}$-value as a function of $\tilde{x}$ and implicitly differentiating with respect to $\tilde{x}$, we have

$$\frac{(\tilde{x}+\tilde{y})}{a^2}(1+\tilde{y}') + \frac{(\tilde{x}-\tilde{y})}{b^2}(1-\tilde{y}') = 0. \tag{2}$$

The standard correlation ellipse is internally tangential to the 1-SD square at the point $(r, 1)$. Hence, we have $\tilde{y}'(r) = 0$ and $\tilde{y}(r) = 1$, and equations (1) and (2) yield

$$\frac{(r+1)^2}{2a^2} + \frac{(r-1)^2}{2b^2} = 1; \text{ and } \quad \frac{(r+1)}{a^2} + \frac{(r-1)}{b^2} = 0. \tag{3}$$

Solving the two equations in (3) simultaneously, we determine $a = \sqrt{1+r}$ and $b = \sqrt{1-r}$.

Having determined a and b, the standard correlation ellipse can be described by any one of the following equivalent equations {of which we prefer the last; that is, expression (4)}:

$$\frac{(\tilde{x}+\tilde{y})^2}{2(1+r)} + \frac{(\tilde{x}-\tilde{y})^2}{2(1-r)} = 1$$

$$\begin{pmatrix} \frac{\tilde{x}+\tilde{y}}{\sqrt{2}} & \frac{\tilde{x}-\tilde{y}}{\sqrt{2}} \end{pmatrix} \begin{bmatrix} 1+r & 0 \\ 0 & 1-r \end{bmatrix}^{-1} \begin{pmatrix} (\tilde{x}+\tilde{y})/\sqrt{2} \\ (\tilde{x}-\tilde{y})/\sqrt{2} \end{pmatrix} = 1$$

$$(\tilde{x} \quad \tilde{y}) \begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & -\sqrt{1/2} \end{bmatrix} \begin{bmatrix} 1+r & 0 \\ 0 & 1-r \end{bmatrix}^{-1} \begin{bmatrix} \sqrt{1/2} & \sqrt{1/2} \\ \sqrt{1/2} & -\sqrt{1/2} \end{bmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = 1$$

$$(\tilde{x} \quad \tilde{y}) \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}^{-1} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = 1. \tag{4}$$

The reader can verify that the correlation matrix $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ has eigen-values $(1+r)$ and $(1-r)$, and eigen-vectors $\left(\sqrt{1/2}, \sqrt{1/2}\right)$ and $\left(\sqrt{1/2}, -\sqrt{1/2}\right)$, respectively. The standard correlation ellipse is also given (among several equivalent expressions) by

$$(1-r^2)\tilde{x}^2 + (\tilde{y}-r\tilde{x})^2 = 1-r^2$$

or,

$$(\tilde{x}-r\tilde{y})^2 + (1-r^2)\tilde{y}^2 = 1-r^2.$$

The standard correlation ellipse has the following properties: It passes through, not just the already mentioned four points L, R, B, T, but also through other recognizable points on the $\tilde{x}$-axis, the $\tilde{y}$-axis, the major- and the minor axis such as

$$\pm\left(\sqrt{1-r^2}, 0\right), \pm\left(0, \sqrt{1-r^2}\right), \pm\left(\sqrt{\tfrac{1+r}{2}}, \text{sign}(r)\sqrt{\tfrac{1+r}{2}}\right), \pm\left(\sqrt{\tfrac{1-r}{2}}, -\text{sign}(r)\sqrt{\tfrac{1-r}{2}}\right).$$

Any vertical line segment terminated by the correlation ellipse is bisected by the $\hat{\tilde{y}} = r\tilde{x}$ line LR; and similarly, any horizontal line segment terminated by the correlation ellipse is

bisected by the $\hat{\tilde{x}} = r\tilde{y}$ line BT. Hence, the standard correlation ellipse also passes through the following four points: $\pm(r, 2r^2 - 1)$ and $\pm(2r^2 - 1,\ r)$.

When the 1-SD square and the 1-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$ are both horizontally and vertically dilated (magnified or expanded) by the same factor c we obtain the c-SD square and the c-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$. To reiterate, the major axis of the c-SD correlation ellipse falls precisely on that diagonal of the c-SD square whose slope has the same sign as r. The ratio of the lengths of the two axes is $ca/[cb] = \sqrt{1+r}/\sqrt{1-r}$. Hence, the c-SD Gaussian correlation ellipse (for all c) has eccentricity

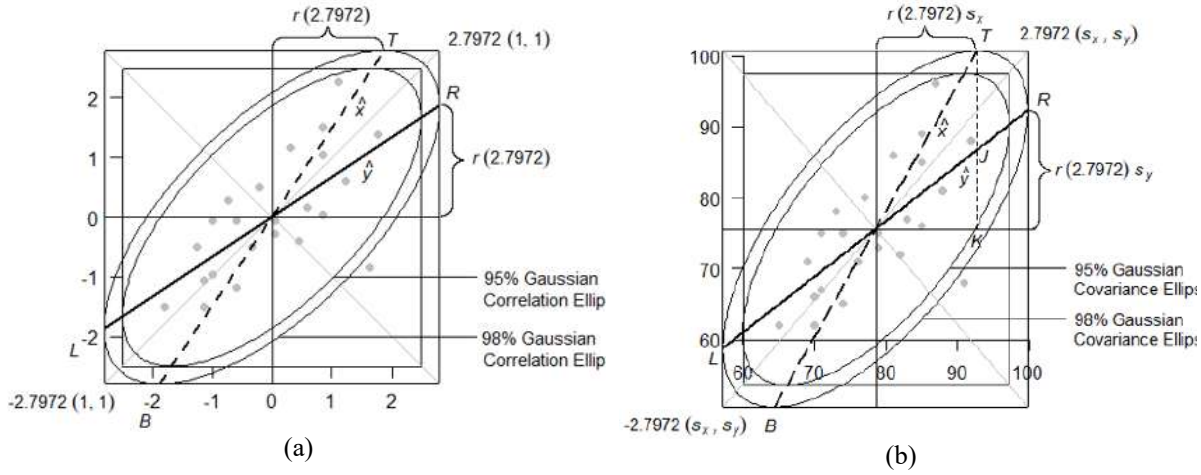
$$e = \frac{\sqrt{a^2-b^2}}{a} = \sqrt{\frac{2|r|}{1+|r|}}. \tag{5}$$

Specifically, when $r = 0$, eccentricity is 0 and the correlation ellipse is a circle; and when $|r| = 1$, eccentricity is 1 and the ellipse with a major axis of half-length $\sqrt{2}$ and a minor axis of half-length 0 collapses into a line segment of length $2\sqrt{2}$.

4. Gaussian Covariance Ellipse

In the more general case, when $s_x \neq s_y$, let us consider the shifted variables $u = x - \bar{x}$ and $v = y - \bar{y}$. Note that the mean vector for (u, v) is (0, 0), and $s_u = s_x \neq s_y = s_v$.

Starting from the c-SD square and the c-SD Gaussian correlation ellipse of $(\tilde{x}, \tilde{y})$, shown in Figure 3(a), if both are horizontally dilated by a factor $s_u = s_x$ and vertically dilated by a factor $s_v = s_y$, and then the image is translated by $(\bar{x}, \bar{y})$, the *transformed regions* are shown in Figure 3(b). We have chosen two different values, $c = 2.448$ and $c = 2.7972$, for reasons given at the end of this section. What exactly are the shapes of these transformed regions?



(a) (b)

Figure 3: A scatter plot of midterm and final exam scores, together with (a) the c-SD square and the c-SD correlation ellipse for <u>standardized</u> data, and (b) the corresponding transformed regions after dilations and translation for <u>raw</u> data, choosing $c = 2.448$ and $c = 2.7972$. What shapes are these transformed regions?

It is trivial to see that each c-SD square in Figure 3(a) turns into a c-SD *rectangle* in Figure 3(b). But it is not easy to recognize that each c-SD correlation ellipse turns into another

ellipse, which we shall call the c-SD covariance ellipse. Why is the dilations-translation of an ellipse another ellipse? How do the major- and the minor axes of the correlation ellipse morph into the corresponding axes of the covariance ellipse?

The answers to these questions are straightforward in the special case when $s_u = s_v = s$, say: The axes of the covariance ellipse coincide with the SD lines and their half-lengths are s-multiples of those of the correlation ellipse. To answer the questions in the more general case when $s_u \neq s_v$, we use matrix algebra. Note that after dilations, the standard correlation ellipse, given in (4), changes into

$$\begin{pmatrix} \frac{u}{s_u} & \frac{v}{s_v} \end{pmatrix} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}^{-1} \begin{pmatrix} u/s_u \\ v/s_v \end{pmatrix} = 1$$

or equivalently,

$$(u \quad v) \begin{bmatrix} s_u^2 & rs_u s_v \\ rs_u s_v & s_v^2 \end{bmatrix}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = 1, \tag{6}$$

which is an ellipse (called the *shifted* standard covariance ellipse, shifted because the mean is (0, 0) and standard because $c = 1$).

Let the eigen-values of the covariance matrix $S = \begin{bmatrix} s_u^2 & rs_u s_v \\ rs_u s_v & s_v^2 \end{bmatrix}$ be α and β; let the associated (orthonormal) eigen-vectors be $(e_{11}, \quad e_{12})$ and $(e_{21}, \quad e_{22})$ respectively; that is,

$$\begin{bmatrix} s_u^2 & rs_u s_v \\ rs_u s_v & s_v^2 \end{bmatrix} = \begin{bmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix};$$

and

$$\begin{bmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then, reversing the steps shown for the standard correlation ellipse, the shifted standard covariance ellipse, given in (6), becomes

$$\frac{(e_{11}u + e_{12}v)^2}{\alpha} + \frac{(e_{21}u + e_{22}v)^2}{\beta} = 1. \tag{7}$$

Returning to the original variables $(x = u + \bar{x}, \; y = v + \bar{y})$ with mean vector $(\bar{x}, \quad \bar{y})$, the shifted standard correlation ellipse (7) becomes the general covariance ellipse (with arbitrary c) given by

$$\frac{(e_{11}(x-\bar{x}) + e_{12}(y-\bar{y}))^2}{\alpha} + \frac{(e_{21}(x-\bar{x}) + e_{22}(y-\bar{y}))^2}{\beta} = c^2. \tag{8}$$

Thus, matching (8) with (1), we note that the c-SD covariance ellipse has major and minor axes given by

$$\begin{aligned} &\text{Major axis: } y - \bar{y} = (e_{12}/e_{11})\,(x - \bar{x}) \text{ with half-length } c\sqrt{\alpha} \\ &\text{Minor axis: } y - \bar{y} = (e_{22}/e_{21})\,(x - \bar{x}) \text{ with half-length } c\sqrt{\beta}. \end{aligned} \tag{9}$$

To complete the discussion on the major and the minor axes, shown in (9), it remains to obtain the eigen-decomposition of the covariance matrix S. The eigen-values are the solutions τ of the quadrative equation $det\begin{bmatrix} s_x^2-\tau & rs_xs_y \\ rs_xs_y & s_y^2-\tau \end{bmatrix}=0$; or equivalently,

$$\tau^2-\left(s_x^2+s_y^2\right)\tau+s_x^2s_y^2(1-r^2)=0$$

or,

$$\tau=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)\pm\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}.$$

The larger eigen-value is $\alpha=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)+\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}$, and the eigen-vector associated with α satisfies $(s_x^2-\alpha)e_{11}+rs_xs_ye_{12}=0$, whence

$$\frac{e_{12}}{e_{11}}=\frac{\alpha-s_x^2}{rs_xs_y}=\frac{\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}-\left(s_x^2-s_y^2\right)}{2rs_xs_y}$$

which has the same sign as that of r.

Likewise, the smaller eigen-value is $\beta=\frac{1}{2}\left\{\left(s_x^2+s_y^2\right)-\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}\right\}$, and the eigen-vector associated with β satisfies $(s_x^2-\beta)e_{21}+rs_xs_ye_{22}=0$, whence

$$\frac{e_{22}}{e_{21}}=\frac{\beta-s_x^2}{rs_xs_y}=-\frac{\sqrt{\left(s_x^2-s_y^2\right)^2+\left(2rs_xs_y\right)^2}+\left(s_x^2-s_y^2\right)}{2rs_xs_y}$$

which has the opposite sign as that of r.

Moreover, instead of documenting the two eigen-vectors, it may suffice to record only the slope $m=e_{12}/e_{11}$ of the major axis (since the minor axis is orthogonal to the major axis, the slope of the minor axis is $-1/m=e_{22}/e_{21}$). Let us compare m with the ratio s_y/s_x of the two SDs in all possible cases:

1. When $r=0$, the major- and minor axes of the covariance ellipse may be chosen to coincide with the two coordinate axes. In this case, $m=0$ if $s_x>s_y$ and $m=\infty$ if $s_x<s_y$.
2. When $s_x=s_y$ and $r\neq 0$, we have $m=sign(r)$; that is, the major axis falls on the SD line if $r>0$, and the minor axis falls on the SD line if $r<0$.
3. When $s_x<s_y$ and $r>0$, we have $m>s_y/s_x$; but when $s_x<s_y$ and $r<0$, we have $m<-s_y/s_x$; that is, the major axis is steeper than the SD line.
4. When $s_x>s_y$ and $r>0$, we have $m<s_y/s_x$; but when $s_x>s_y$ and $r<0$, we have $m>-s_y/s_x$; that is, the major axis is less steep than the SD line.

How should one choose the multiplier c to construct the c-SD rectangle? If one desires a fraction p of points to fall outside the covariance ellipse, one can choose c as the $(1-p)^{\text{th}}$ percentile of a chi-square distribution with two degrees of freedom, obtained from R using code: sqrt(qchisq(1–p,2). For example, 60.65% of points fall outside the 1–SD covariance ellipse; 5% of points will fall outside the 2.448–SD covariance ellipse; 1.11% outside the 3–SD covariance ellipse; 1% outside the 3.035–SD covariance ellipse. In Figure 3, we have used $c = 2.448$ and $c = 2.7972$ to flag the farthest (in 2-d sense) 5% and 2% of the scatter points. We recommend using $c = 2.8$, since this value is easy to remember and since with this choice, for a bivariate normal distribution, roughly 2% of the points fall outside the covariance ellipse.

Whereas the x-outliers and the y-outliers are already detected using the single variable c-SD line segments, the c-SD covariance ellipse is a handy tool to detect the regression outliers or bivariate outliers.

We mention a few properties of the c-SD Gaussian covariance ellipse: As mentioned before, the c-SD Gaussian covariance ellipse is internally tangent to the c-SD rectangle at four points: bottommost point $B = (\bar{x} - rcs_x, \bar{y} - cs_y)$, topmost point $T = (\bar{x} + rcs_x, \bar{y} + cs_y)$, leftmost point $L = (\bar{x} - cs_x, \bar{y} - rcs_y)$ and rightmost point $R = (\bar{x} + cs_x, \bar{y} + rcs_y)$. Moreover, LR is the $\hat{y}$ line, BT is the $\hat{x}$ line. These two regression lines LR and BT intersect at the center of the ellipse, which is also the point of intersection of the two diagonals of the c-SD rectangle and is also the mean vector $(\bar{x}, \bar{y})$. As it was for the correlation ellipse, any vertical line segment terminated by the covariance ellipse is bisected by the $\hat{y}$-line LR; and any horizontal line segment terminated by the covariance ellipse is bisected by the $\hat{x}$-line BT. Admittedly, the directions and lengths of the major- and the minor-axis of the c-SD Gaussian covariance ellipse, given in (9) and the discussion afterwards, are relatively more difficult to fathom. Nonetheless, astute students of statistics will do wisely to learn them.

5. Sufficiency

So far, we established that the c-SD rectangle and the c-SD covariance ellipse summarize all bivariate statistics mentioned in Section 2. Now we go a step further to claim that it suffices to draw only one c-SD Gaussian covariance ellipse (for any value of c) since all summary statistics can be recovered from it. How so?

Here is how: Refer to Figure 3(b) again. Given the c-SD covariance ellipse, the c-SD rectangle can be reconstructed by sandwiching the ellipse between lines parallel to the two coordinate axes. Hence, we can locate the four points of tangency B, T, L, R between the c-SD covariance ellipse and the c-SD rectangle. Then, using the points of tangency, we obtain the regression lines LR (for $\hat{y}$) and BT (for $\hat{x}$). The center of the ellipse is found either as the point of intersection between LR and BT, or the point of intersection of the two diagonals of the c-SD rectangle. The correlation coefficient r is the ratio of the horizontal distance between B and T to the horizontal distance between L and R of the c-SD rectangle (with sign positive if T is to the right of B, and negative otherwise); or equivalently, it is the ratio of the vertical distance between L and R to the vertical height BT of the c-SD rectangle (with sign positive if L is below R, and negative otherwise). The major and the minor axes are found (at least visually) as the largest and the smallest diameters (line segments passing through the center and terminated by the ellipse). If the two axes have half-lengths a and b respectively, we can also calculate $r = \frac{a^2 - b^2}{a^2 + b^2}$, since one can verify that $\frac{a}{b} = \sqrt{\frac{1+r}{1-r}}$. Finally, if the vertical line from T to

the horizontal line $y = \bar{y}$ intersects LR (the $\hat{y}$ line) at J and ends at K, then the coefficient of determination is $r^2 = JK/TK$.

6. Further Reduction

To help the user decipher all bivariate summary statistics, we recommend superposing the entire c-SD covariance ellipse on the scatter plot. However, for mathematical completeness, we must mention that it suffices to superpose the four points of tangency B, T, L, R between the covariance ellipse and the c-SD rectangle. In fact, any three of these points will also suffice. For instance, given B, T, L, you can discover R as follows: Join BT; find its midpoint M; join LM and produce it to R such that $LM = MR$. Using these four points, we can obtain the c-SD rectangle, as explained below and shown in Figure 4.

Draw horizontal lines through B and T and vertical lines through L and R. Their points of intersection form the c-SD rectangle $EFGH$, whose vertices are labeled clockwise starting from the north-west corner. Then we have

$$s_x = \frac{1}{2c}EF,\ s_y = \frac{1}{2c}FG, \qquad \text{and} \quad r = 2\frac{ET}{EF} - 1 = 1 - 2\frac{TF}{EF}. \tag{10}$$

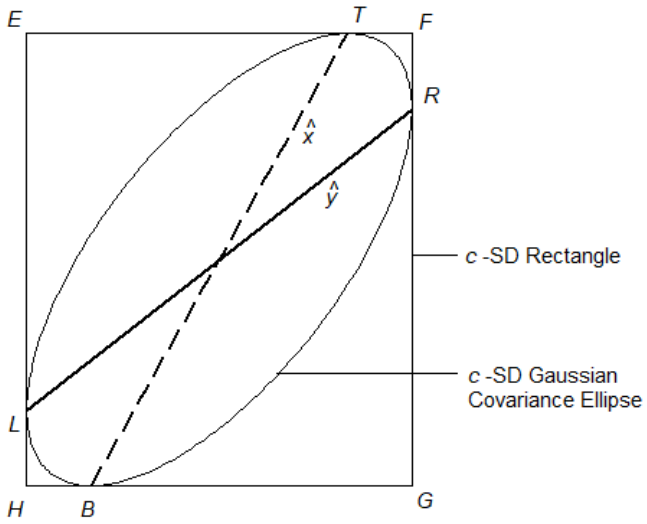


Figure 4: Any three points, out of the four points of tangency T, B, L, R, suffice to reconstruct the c-SD rectangle and the two regression lines

As already mentioned before, LR is the $\hat{y}$ line, BT is the $\hat{x}$ line. Furthermore, if we draw vertical lines through B and T, and horizontal lines through L and R, then their intersections form the inner rectangle $E'F'G'H'$ whose area as a proportion of the area of the outer rectangle $EFGH$ represents the coefficient of determination r^2. See Figure 5. In particular, as r approaches 0, the inner rectangle $E'F'G'H'$ reduces in size until it coincides with the center; and as r approaches 1, the inner rectangle $E'F'G'H'$ increases in size until it coincides with the outer rectangle $EFGH$.

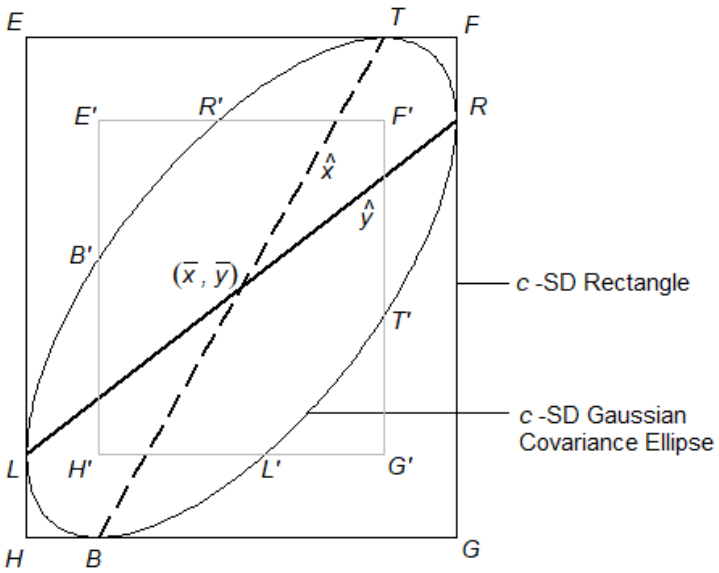


Figure 5: Any three of the four points of tangency B, T, L, R, suffice to calculate the coefficient of determination and the correlation coefficient.

Thereafter, we impose a coordinate system such that the point of intersection between BT and LR represents $(\bar{x}, \bar{y})$. Then using the SDs and the correlation given in (10) and scaling both variables in (6) by the same factor c, the c-SD covariance ellipse is given by

$$(x - \bar{x}, y - \bar{y}) \begin{bmatrix} s_x^2 & rs_x s_y \\ rs_x s_y & s_y^2 \end{bmatrix}^{-1} \begin{pmatrix} x - \bar{x} \\ y - \bar{y} \end{pmatrix} = c^2. \tag{11}$$

We can draw (at least a free-hand sketch of) the c-SD covariance ellipse internally tangential to the c-SD rectangle at the four points B, T, L, R and passing through B', T', L', R' obtained by moving vertically points B, T towards the regression line $\hat{y}$ (or LR) and continuing equally far on the opposite side of LR, and similarly moving points L, R horizontally towards the regression line $\hat{x}$ (or BT) and continuing equally far on the opposite side of BT. Additional points on the ellipse are found by repeating the process.

7. Conclusion

While a box plot and a mean-SD arrow (or alternatively, a c-SD line) offer graphical summaries of one continuous variable, our proposed c-SD covariance ellipse does the same for two continuous variables. Using the c-SD covariance ellipse, we can recover the means, the SDs, the correlation coefficient r, the regression line $\hat{y}$ (as a linear function of x), the regression line $\hat{x}$ (as a linear function of y), and the coefficient of determination r^2. Thereafter, the equation of the ellipse can be recovered from (11). Moreover, scatter points outside the c-SD covariance ellipse (with a desired choice of c) are flagged as potential outliers. We hope that the c-SD covariance ellipse (or simply any three points of tangency between the c-SD covariance ellipse and the c-SD rectangle) will help users develop better intuitions about the important concepts of correlation, regression and bivariate outliers.

It is worth mentioning that the c-SD covariance ellipse, given in (11), is the shortest area region (of any shape) that captures inside it a specific fraction of the bivariate normal distribution approximately equal to the cumulative distribution function of a chi-square

variable with two degrees of freedom {given by R codes: pchisq(c^2, 2) = pexp(c^2/2)}. Equivalently, the contour plots of a bivariate normal distribution form a family of ellipses for various values of c. What we have demonstrated in this paper is that given any one of these contour ellipses (even when not knowing the value of c) we can discover the mean vector, the SD line, the SD-ratio s_y/s_x, the correlation coefficient r, the two regression lines $\hat{y}, \hat{x}$, and the coefficient of determination r^2.

Here is the answer to the quiz we posed in the Section 2 regarding the largest (or the smallest) SD of a suitable projection of a scatter plot: The largest SD is attained when the scatter plot is projected on to the major axis of the c-SD covariance ellipse, for any c. Similarly, the smallest SD is attained when the scatter plot is projected on to the minor axis of the c-SD covariance ellipse.

We invite the interested reader to depict simultaneously the summary statistics involving three or more quantitative variables—specifically focusing on multiple correlation coefficient partial correlation coefficients, and principal components.

Acknowledgements

We thank our fellow faculty members and graduate students who participated in multiple pop quizzes on guessing the summary statistics based on the covariance ellipse. Special thanks are due to an anonymous reviewer for suggesting a more focused exposure of the methodology.

References

Devore, J. (2015). *Probability and Statistics for Engineering and Sciences*. Ninth Edition, Boston, MA: Brooks/Cole, Cengage Learning.

Embse, C. and Engebretsen, A. (1996). Visual representations of mean and standard deviation. *The Mathematics Teacher*, **89**(**8**), 688–692.

Maverick, L. A. (1932). Graphic presentation of standard deviation. *Journal of the American Statistical Association*, **27**(**179**), 287–297.

Rashid, M. and Sarkar, J. (2018). Cyber mentoring in an online *Introductory Statistics* course. *Educational Research Quarterly*, **41**(**3**), 25–38.

Sarkar, J. and Rashid, M. (2016). Visualizing mean, median, mean deviation and standard deviation of a set of numbers. *The American Statistician*, **70**(**3**), 304–312.

Sarkar, J. and Rashid, M. (2019). Portraying standard deviation via revolution. *Journal of Probability and Statistical Science*, **17**(**1**), 109–119.

Sarkar, J. and Rashid, M. (2020). Shutter plot: A visual display of summary statistics over a scatter plot. *International Journal of Statistical Sciences*. To appear.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 89–100

Modelling Pepper Export Income in Sri Lanka Using Deterministic Decomposition and Seasonal ARIMA Models

W.P.M.C.N. Weerasinghe and D.D.M. Jayasundara
Department of Statistics and Computer Science, University of Kelaniya, Kelaniya 11 600, Sri Lanka

Received: 27 June 2020; Revised: 16 August 2020; Accepted: 08 October 2020

Abstract

Pepper is the most used spice and flavoring agent in the food industry. Sri Lanka is the fifth largest exporter of pepper in the world. Variations in export income of pepper is a huge matter for farmers, consumers, investors and policy makers in the country. Hence an accurate forecast of export income is extremely important. This study presents a time series modelling approach for forecasting the income of pepper exports in Sri Lanka. Two different models are adopted: deterministic decomposition model which is built decomposing the trend, seasonality and the random variations and Seasonal Auto Regressive Integrated Moving Average (SARIMA) which belongs to the stochastic class of models. The data used in this study are monthly export income of pepper in Sri Lanka from January 2000 to December 2018. Among the two types of models, deterministic decomposition model with Mean Absolute Error (MAE) of 4.36 has a strong potential in forecasting the income of pepper exports in Sri Lanka. As the forecasts from the model shows an increasing pepper export market which will need a higher production of pepper, the government can improve the awareness of farmers about the requirements of pepper in export market by providing infra-structure facilities.

Key words: Deterministic decomposition; Forecasting; MAE; Seasonality; Trend.

1. Introduction

International trade and finance are important economic concepts in a country where international trade in goods and services allows nation to raise their standards of living by exporting and importing goods. Exporting and importing helps in growing national economies and expanding the global market. The more a country exports, the more domestic economic activity is occuring. So, export sector plays a vital role in every country's economy influencing the level of economic growth, employment opportunities, balance of payments and international relations.

According to the Economic Complexity Index (ECI), Sri Lanka is the 79th largest export economy in the world and the 101st most complex economy. In 2017, Sri Lanka exported \$12B and imported \$21.1B, resulting in a negative trade balance of \$9.1B in net imports ("Sri Lanka (LKA) Exports, Imports, and Trade Partners | OEC - The Observatory of Economic Complexity" , 2020.). Sri Lanka is an island and borders India and Maldives by sea that creates a large possibility in export market. The top export destinations of Sri Lanka

Corresponding Author: W.P.M.C.N. Weerasinghe
Email: wpmcn.weerasinghe@gmail.com

are the United States, United Kingdom, Germany, India and Italy. Sri Lanka exports 204 products with revealed comparative advantage which includes Ceylon tea, apparel, spice and allied products, aquarium fish, seafood, rubber and rubber-based products, wooden products. Sri Lanka was historically known as the Spice Island and further attracted by the Western nations for its spices. Sri Lankan Spices and Allied products Suppliers export the most sought-after cinnamon, pepper, cloves, cardamoms, nutmeg, mace and vanilla. 56% of Sri Lankan Agricultural exports consists of spices, allied products and essential oils. These varieties are used to season, flavour and aromatise various forms of cuisines across the world.

Pepper is the most significant and widely used spice in the world. There are but a few countries in Asia including Sri Lanka and the Pacific, Brazil and Madagascar that produce pepper. Currently, Sri Lanka ranks at fifth place in terms of area under pepper cultivation and seventh place in terms of production with a world share of 5.7% in production. Pepper is mainly used as a spice and flavoring agent in the food industry. It also has industrial uses in perfumery and pharmaceutical industries. Sri Lanka exports pepper as: black pepper, white pepper, black pepper oil, pepper ground pepper crushed, light berries, organic pepper (Institute of Policy Studies in Sri Lanka and Department of National Planning, 2017). However, majority of the exports are black pepper – neither crushed/ grounded (53% of total export earnings of pepper products in 2012), followed by light berries.

In Sri Lanka, pepper is mainly cultivated in Low and mid country wet and Intermediate agro-climatic zones. The total extent of pepper in Sri Lanka is about 29,378 ha and Matale, Kandy, Kegalle, Badulla, Ratnapura, Monaragala, and Kurunegala are the major districts. Unlike other pepper growing countries, Sri Lanka pepper flowers twice a year during the Maha and Yala rains. Currently about 60% of pepper production of the country is exported, while the remainder is consumed domestically according to government sources (Institute of Policy Studies in Sri Lanka and Department of National Planning, 2017). Sri Lanka is the fifth largest exporter of pepper in the world. India buys 62% of pepper exports from Sri Lanka followed by European, American and other Asian and African countries (Institute of Policy Studies in Sri Lanka and Department of National Planning, 2017). Given that majority of black pepper from Sri Lanka is exported to the low end of the market (India), Sri Lanka should explore high end markets elsewhere. Its widely known that the demand for Sri Lankan pepper is increasing rapidly due to its richer in piperine content, which lends it a superior quality and greater pungency (Ministry of Development Strategies and International Trade and Sri Lanka Export Development Board, 2018). Thus, Sri Lanka has the potential to become a key player in the global pepper market by raising the country's pepper production penetrating through high value export markets.

The unforeseen variations in export value (export income) can complicate budgetary planning in a country. Hence an accurate forecasting method of export income is extremely important for efficient monitoring and planning of export commodities. Many attempts have been made in the past to develop forecasting models for export quantity and price of various commodities around the world. Most of the past research have focused on forecasting the export quantity or export volume of a commodity. Autoregressive Integrated Moving Average (ARIMA) and Bayesian Artificial Neural Network (BANN) were used in forecasting the imports and exports of the Philippines and the two models were compared using comparison methods (Urrutia *et al*., 2019). Seasonal autoregressive integrated moving average (SARIMA) methodology has been applied for modelling and forecasting of monthly export of meat and meat products from India (Paul *et al*., 2013). ARIMA model by using Box-Jenkins methodology was used to forecast the export/import of wood based panel in

India (Plywood, 2015). A forecasting model was built for Exports of Dates in Pakistan for the next 15 years using ARIMA technique (Naz, 2012). An analysis has been done in India about the production and export performance of black pepper (Mokshapathy, 2017).

There are some research papers related to pepper in Sri Lanka in current literature. Analysis and a comparison were done among the small and large scale pepper farmers considering production levels and costs, income and profitability of pepper cultivation. The study has concluded that there is a significant difference between small and large scale pepper farmers in the extent cultivated, incomes received, cost of production and profits earned (Sivarajah, 2018). No studies have been conducted on modelling and forecasting the export income of pepper in any country so far. In this context, it is necessary to know to what extent the export income is being fluctuated and to draw meaningful policy conclusion. Hence, this study focuses on the objective of modelling and forecasting the pepper income in Sri Lanka by using time-series models.

2. Materials and Methodology

Two main approaches to the research problem with their methodologies are discussed in here: deterministic decomposition method and is built decomposing the trend, seasonality and the random variations and second method is based on SARIMA, and it belongs to the stochastic class of models.

2.1. Data collection

In the context of export sector, export volume and export value are important terms. Export volume refers to the quantity of goods exported. It is usually in kilograms, Metric tons. Generally, Tons is used for calculating export volume in worldwide. Export value is the value of export goods in currency and it may be measured in any currency such as USD, EURO, and RUPEES. Export value represents the total worth of a commodity hence this is the total export income of that commodity to the country. This study is based on the export income of pepper in Sri Lanka. Based on the varieties of pepper, they are compiled according to the trade classification harmonized system with coding as HS 090411, HS 090420, HS 090412 which represents dried pepper (crushed or ground), fruits of genus capsicum or pimento (dried, crushed or ground) and pepper (crushed or ground) respectively. Monthly data on total pepper export value (US Dollar Millions) in Sri Lanka were collected from Sri Lanka Exports Development Board for the period of January 2000 to December 2018 (228 observations). The data set was divided into two parts as 80% and 20% for the model building and model validation respectively. Historical data in the period from January 2000 to March 2015 was used in the model identification and the data from April 2015 to December 2018 was used for the model validation. R statistical package was used in analysis and modelling of pepper export income.

2.2. Preliminary analysis

One of the most common data pre-processing technique is data cleaning. It includes fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies. In this study, the data set is first explored to identify the outliers and the missing values. There were no any missing values or outliers detected in the data set. Therefore, the original series was taken for model fitting. A stationary time series is defined as a time series whose mean and variance are constant over time. In order to identify the stationary of a time series,

statistical tests are used. The three tests Kwiatkowski-Phillips-Schmidt-Shin (KPSS), Augmented Dickey Fuller (ADF) and Phillips Perron (PP) were used in this study to check the stationary of the time series.

2.3. Time series forecasting methods

Time series is a set of time-ordered observations of a process where the intervals between observations remain constant such as hours, weeks, quarters, months, years. Time series analysis can be applied to any variable that changes over time. The main objective of time series analysis is to develop models that best capture an observed time series in order to understand the underlying causes. Time series models used for forecasting include decomposition models, exponential smoothing models, ARIMA models and SARIMA models.

2.4. Components of a time series

The components of a time series are trend, seasonal variation, cyclic changes, and irregular factors. Trend is the long-term change in the mean level and often thought of as the underlying growth or decline component in the series. Cyclic changes within a time series are similar to the seasonal component in that it is revealed by a wavelike pattern. Once the trend and cyclic variations have been accounted for, the remaining movement is attributed to irregular fluctuations.

2.5. Seasonality in a time series

Seasonal variation, or seasonality, are changes that repeat themselves within a fixed period such as weekly, monthly, quarterly. Seasonality may be caused by various factors, such as weather, vacation, and holidays and consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series. Seasonality of a time series can be detected in many ways including graphical methods and statistical tests. The run sequence plot is considered as a first step in analyzing any time series plot. Seasonal subseries plot, box plot, Auto Correlation Function (ACF) plot do an excellent job in showing seasonal variations. In this study, ACF plot is used to identify seasonality among graphical methods. If there is a significant seasonality, ACF plot shows spikes at lags equal to the seasonal period. Student *t*-test and Wilcoxon Signed-Ranks test are some popular statistical tests for detecting seasonality in a time series (Nwogu *et al*., 2016).

In R software, Webel-Ollech overall seasonality test combines results from different seasonality tests. It combines the results of the QS-test and the Kruskal-Wallis (KW) test. If the p-value of the QS-test is below 0.01 or the p-value of the KW test is below 0.002, the WO-test will classify the corresponding time series as seasonal. This test is used in this study to identify seasonality in theoretically (Ollech, 2019).

2.6. SARIMA model

SARIMA models are used in cases where the time series exhibits a seasonal variation. It is formed by including additional seasonal terms in the ARIMA model. A seasonal autoregressive notation (*P*) and a seasonal moving average notation (Q) will form the multiplicative process of SARIMA as ARIMA *(p, d, q)(P, D, Q)s* where subscripted letter ‘s’ shows the length of seasonal period. The multiplicative SARIMA model has the form in (1).

$$\Phi_P(B^s)\varphi_p(B)\nabla_s^D\nabla_{z_t}^d = \Theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (1)$$

where $\Phi_P(B^s)$ is the seasonal AR operator of order P, φ_p is the regular AR operator of order p, ∇_s^D represents the seasonal differences, ∇^d represents the regular differences, $\Theta_Q(B^s)$ the seasonal moving average of order Q, $\Theta_q(B)$ is the regular MA order of order q and ε_t is a white noise process.

The Box-Jenkins (BJ) methodology of estimating a time series model consists of four iterative steps: Model identification, Estimation of model parameters, Diagnostic checking and forecasting. First tentative model parameters are identified through ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function), then coefficients of the most likely model are determined, next steps involves is to forecast, validate and check the model performance by observing the residuals through Ljung Box test and ACF plot of residuals.

2.7. Decomposition method

Time series can be decomposed into various sub-components and their effects can be checked in the data in series. Mainly, time series data composed of seasonal pattern and trend pattern. There are two different decomposition models possible.

Additive Decomposition: Here, the total data are taken as the sum of the decomposed components.

$$X_t = seasonal(S_t) + trend(T_t) + random$$

Multiplicative Decomposition: Here, the given time series data are treated as the product of the decomposed components.

$$X_t = seasonal(S_t) \times trend(T_t) \times random$$

An additive model is appropriate if the magnitude of the seasonal fluctuations does not vary with the level of time series. The multiplicative model is appropriate if the seasonal fluctuations increase or decrease proportionally with increases and decreases in the level of the series. Multiplicative decomposition is more prevalent with economic series because most seasonal economic series do have seasonal variations which increase with the level of the series. Often the transformed series can be modeled additively when the original data is not additive. In particular, logarithm turn a multiplicative relationship into an additive relationship. So, a multiplicative relationship can be fitted by fitting a more convenient additive relationship to the logarithms of the data and then to move back to the original series by exponentiating.

The steps involved in developing the multiplicative decomposition model in this study are listed below.

- *Estimating the trend and the seasonal factors*

  Here, both trend and the seasonal effects were estimated specifying a regression equation. To decide upon the mathematical form of a trend, one must first draw the

plot of the time series. The number of seasonal factors is equal to the frequency of the series (e.g. monthly data = 12 seasonal factors, quarterly data = 4, etc.)

- *Calculating the irregular component; for an additive model* $\varepsilon_t = Y_t - T_t - S_t$
- *Analyzing the residual component.*
- *Whichever method was used to decompose the series, the aim is to produce stationary residuals.*
- *Choosing a model to fit the stationary residuals.*
- *Forecasting can be achieved by forecasting the residuals and combining with the forecasts of the trend and seasonal components.*

2.8. Forecasting accuracy

The forecast errors are the difference between the actual values in the test set and the forecasts produced using only the data in the training set. The two most commonly used scale-dependent measures are based on the absolute errors or squared errors: The smaller the difference, the better the model is. Several criteria such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Weighted MAPE can be used to compare different forecasting models. In this study, two different error metrics are considered for the evaluation of the forecasting models: MAE and MAPE. MAPE usually expresses the prediction accuracy of a forecasting model as a percentage and it is defined in (2) and MAE is given in (3).

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3}$$

where $\hat{y}_i$ = estimated value of y_i, y_i = actual value, n = number of observations.

3. Results and Discussion

Results under the two approaches; deterministic decomposition and Seasonal ARIMA model are clearly explained in here.

3.1. Year wise evaluation of pepper export value

According to Figure 1, a clear seasonal pattern in export income in Sri Lanka can be observed in each year. The export value of pepper decreases in April in each year while it increases to a maximum value in June – July period. Another significant feature is that export value has gradually increased in each year. This proves that there is a significant pattern of export income of pepper in Sri Lanka in each year.

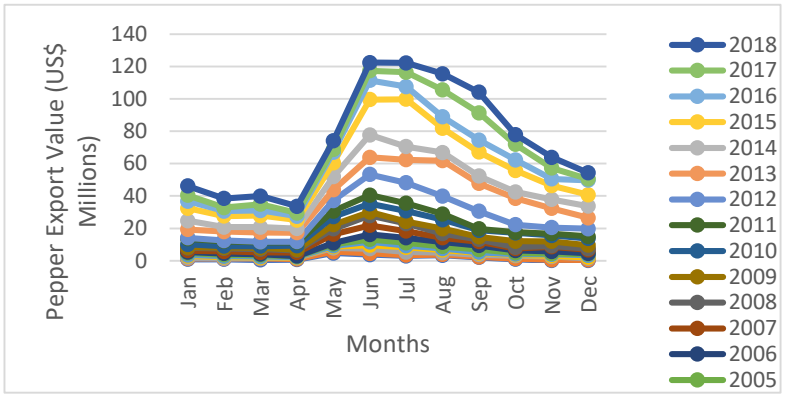


Figure 1: Pepper export income evaluation in Sri Lanka – Year-wise

3.2. SARIMA model

As the series does not contain missing values, the original series was used in model fitting. On plotting the export income data of pepper as shown in Figure 2, a slight trend with a seasonality can be observed. It is confirmed with the slow declines and oscillations given in the ACF plot of the series in Figure 3. Further seasonality WO test in R software which is a combined tests of QS-test, the QS-R test and the KW-R-test was also performed and all *p* values of the tests were less than 0.05 indicating the series exhibits seasonality.

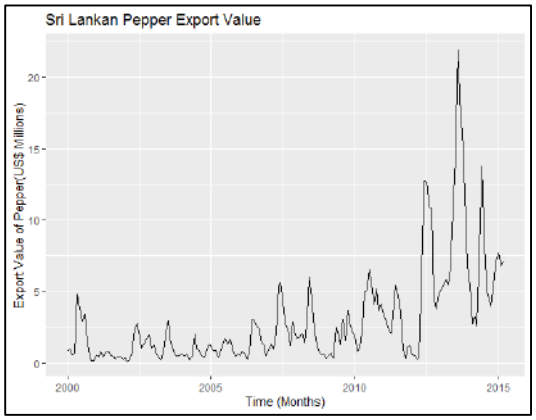


Figure 3. ACF plot of the series of pepper in Sri Lanka

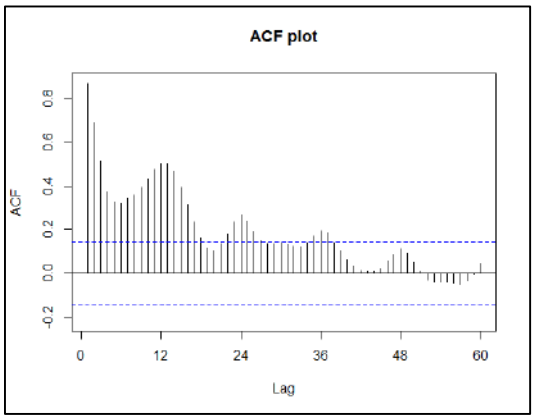


Figure 2: Time series plot of export income

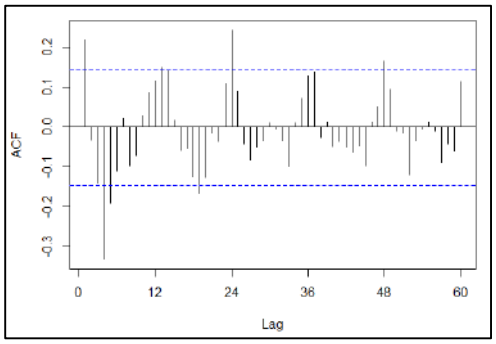


Figure 4: ACF of the first differenced series

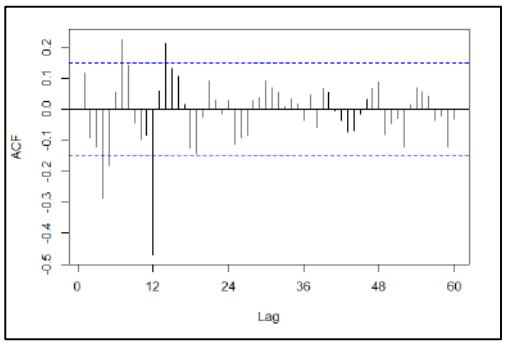


Figure 5: ACF of the first and Seasonal differenced series

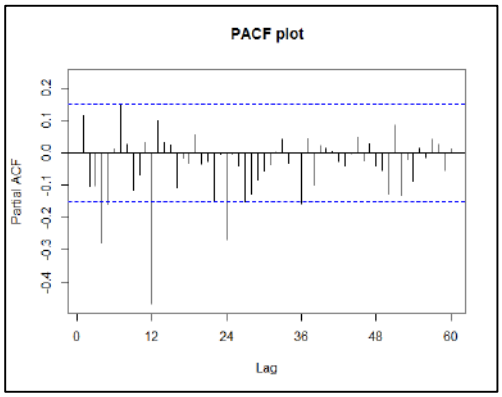


Figure 6: PACF of the first differenced and seasonal differenced series

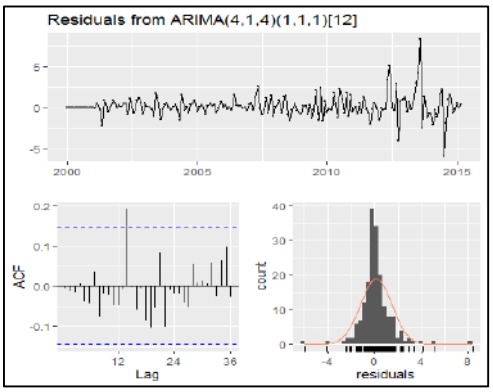


Figure 7: Model adequacy of ARIMA (4,1,4) (1,1,1) [12] model

Since KPSS (p value=0.1) test suggest that the series is not stationary at 5% level of significance, first difference of the series was taken. The ACF plot of the first differenced series is shown in Figure 4 which shows significant spikes at lags 12, 24 which identifies the seasonality period as 12.

A seasonal difference at lag 12 was taken then for better identification of the model parameters. Through the investigation of the ACF and PACF of the first and seasonal differenced series shown in Figure 5 and Figure 6, the seasonal and non-seasonal lags were identified as; Seasonal AR lags: 12, 24, 36, Seasonal MA lags: 12, Non – seasonal AR lags: 4, 5 and Non – seasonal MA lags: 4, 5. Then, several SARIMA models were tested and ARIMA (4, 1, 4) (1, 1, 1) [12] was identified as the best model with lowest AIC value (637.01) for modelling the income of pepper exports in Sri Lanka.

The ARIMA (4, 1, 4) (1, 1, 1) [12] model is represented as in (4).

$$\phi(B^{12})\varphi(B)(x_t - \mu) = \theta(B^{12})\theta(B)w_t \quad (4)$$

where,

Non- seasonal component:

$AR: \varphi(B) = 1 - 0.3229B - 0.0529B^2 + 0.2058B^2 - 0.1188B^3$

$MA: \theta(B) = 1 + 0.2922B + 0.1463B^2 + 0.1228B^2 + 0.5227B^3$

Seasonal component:

$Seasonl\ AR: \phi(B^{12}) = 1 + 0.3291B^{12}$

$Seasonl\ MA: \theta(B^{12}) = 1 + 0.7016B^{12}$

Residuals of the fitted model was evaluated as in Figure 7. Ljung Box test returns a large p value (0.1612) indicating that the residuals are random and independent variance at 5% level of significance. Further ARCH test gives large p value (0.9336) indicating the residuals have a constant variance at 5% level of significance.

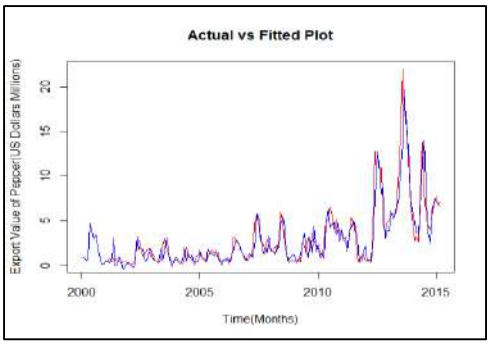


Figure 8: Actual and Fitted values

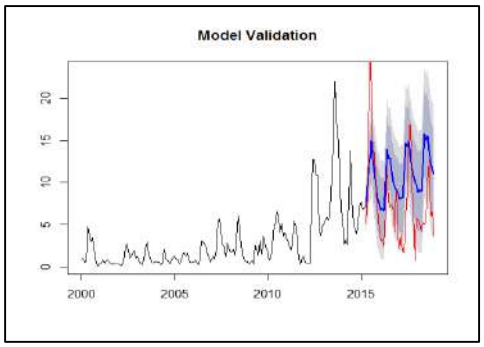


Figure 9: Model validation

In Figure 8, red lines shows the actual data and the blue line shows the fitted data points from (4) where it suggests that the identified model is suitable in forecasting the pepper export income beyond the year 2015 hence the gap between two lines is minimum. 20% of the data was used in model validation and it is shown in Figure 9.

It gives a considerable low MAE (4.76) and MAPE (43.81) which outlines ARIMA (4, 1, 4) (1, 1, 1) [12] is quite good in forecasting the income of pepper exports in Sri Lanka.

3.3. Deterministic decomposition method

As the series of pepper export income have seasonal variations which increase or decrease with the level of the series, multiplicative decomposition is used as shown in Figure 10. It shows the observed series, trend line, seasonal pattern and random part of the series. Hence the decomposition model looks as shown in (5). As the logarithms turn a multiplicative relationship into an additive relationship, taking the logarithms of both sides gives the additive model as shown in (6). Hence the logarithm of the series was taken as shown in Figure 11.

$$Y_t = T_t . S_t . \varepsilon_t \tag{5}$$

$$log(Y_t) = log(T_t) + log(S_t) + log(\varepsilon_t) \tag{6}$$

Both trend and the seasonal effects were estimated specifying a regression equation. The number of seasonal factors is equal to the frequency of the series: there are 12 seasonal factors as monthly data is used in here.

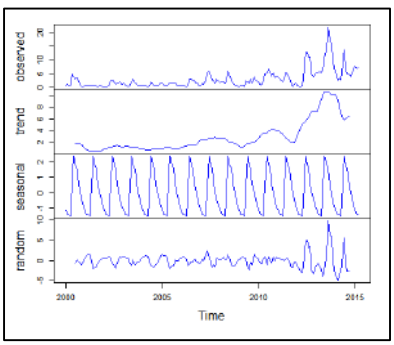


Figure 10: Decomposition of the time series plot of pepper export income of Sri Lanka

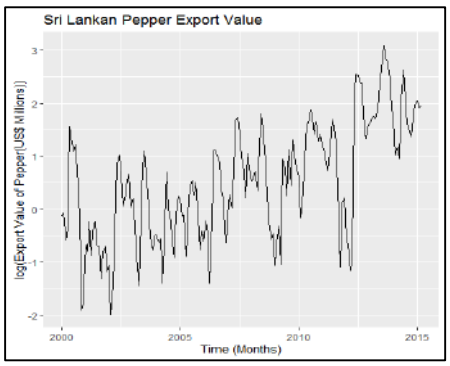


Figure 11: Log transformed series

The regression equation results are shown in Figure 12, where t = 1,2,3…, is for trend, dm1 is the dummy variable for the 1st month, dm2 for the second month etc. In order to avoid the dummy variable trap, one dummy variable is excluded from the regression model.

```
Call:
lm(formula = log(exp_val) ~ t + ., data = AP)

Coefficients:
(Intercept)            t          dm1          dm2          dm3          dm4          dm5          dm6
 -321.65810      0.16034     -0.01112     -0.18049     -0.25171     -0.40714      0.69329      1.13516
        dm7          dm8          dm9         dm10         dm11
     0.95050      0.77802      0.42310      0.12388      0.05692
```

Figure 12: Estimated trend and seasonal effects from regression equation

The error component can be obtained by subtracting trend and seasonal components as $\varepsilon_t = Y_t - T_t - S_t$ and the plot of that error component is shown in Figure 13.

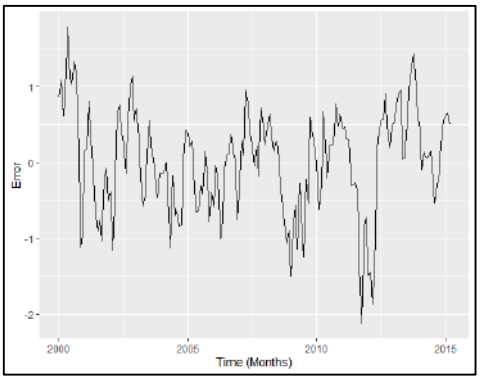


Figure 13: Time series plot of error component

As the series seems to be stationary, for further evidence stationary tests were carried out. Stationary tests confirm that the series of error component is stationary at 5% level of significance. As the residuals are stationary, next step is to fit a model for stationary residuals. ARMA (2, 1) model was selected as the best model with lowest AIC value (227.55) among many tentative models that were tested based on the ACF and PACF of the stationary residuals. The fitted ARMA (2, 1) model for the stationary residuals is given in (7).

$$x_t = 0.0179 + 0.0440x_{t-1} + 0.4757x_{t-2} + \varepsilon_t + 0.8487\varepsilon_{t-1} \quad (7)$$

Model adequacy of the fitted model for residuals was evaluated as in Figure 14. According to Figure 14, ACF plot of the residuals of the ARMA (2, 1) model shows that all autocorrelations are within the threshold limits indicating the residuals are behaving random. Further Ljung Box test on squared residuals gives large p value (0.0909) indicating the residuals have a constant variance at 5% level of significance.

Forecasting of the export income of pepper beyond year 2015 was achieved by forecasting the residuals and combining with the forecasts of the regression model that contains trend and seasonal components. Figure 15 represents the model validation of the time series decomposition model where black color series represents the actual data and the red color represents the forecasted data. Forecasting accuracy of the 20 % of the test data is measured with MAE and MAPE with values 4.3567 and 41.8718 respectively.

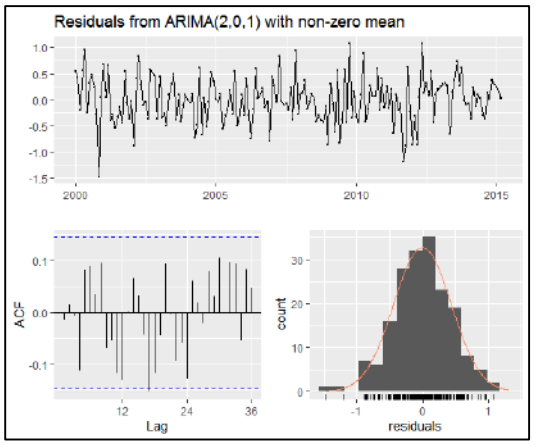


Figure 14: Model adequacy of ARMA (2,1) model

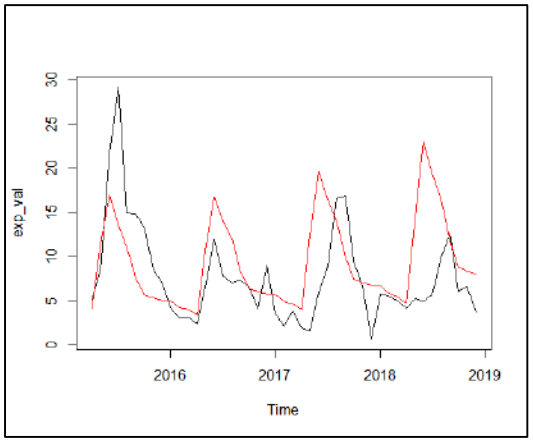


Figure 15: Model validation of the time series decomposition model

3.4. Model comparison

In order to identify the best model from SARIMA and the time series deterministic decomposition models, the forecasting accuracy of both models were compared with error metrics MAE and MAPE. The model comparison results are shown in Table 1.

Table 1: Model Comparison Results

Model Type	MAE	MAPE
SARIMA (4, 1, 4) (1, 1, 1) [12]	4.76	43.81
Deterministic decomposition model	4.36	41.87

According to Table 1, it is clear that deterministic decomposition model is the best model for forecasting the export income of pepper in Sri Lanka because it has minimum MAE and MAPE.

4. Results and Discussion

Forecasting the export income of pepper in Sri Lanka with time series modelling approach was carried out through this study. Monthly data on total pepper export value (US Dollar Millions) in Sri Lanka were used for the analysis. Two main approaches have been implemented to solve the research problem. First model is implemented as a stochastic class of models; SARIMA model. Second model is a Deterministic Decomposition model which assumes that the export income of pepper in Sri Lanka is a composition of three parts: a trend component, a seasonal component and an irregular component. The chosen decomposition model is multiplicative. After investigating time series approaches, the deterministic decomposition model was evidently selected as the best approach in forecasting the export income of pepper in Sri Lanka. ARIMA (4, 1, 4) (1, 1, 1) [12] which was selected as the best model under stochastic models approach also do a quite good job in forecasting export income of pepper in Sri Lanka but the best results can be achieved by the deterministic decomposition model which demonstrates a good performance in terms of both explained variability and forecasting. Most forecasted values are similar to actual values. Forecasts from the model shows a significant positive trend which can be considered as a growth of

pepper export market in Sri Lanka in the future. Therefore, this study helps Sri Lankan pepper exporters to consider about long term investment decisions by identifying trends in the export value. As well as this may be useful for the government policy makers to rethink about the policy agenda of the country. As the forecasts shows an increasing pepper export market which will need a higher production of pepper, the government can improve the awareness of farmers about the requirements of pepper in export market by providing infra-structure facilities such as establishing central collecting, processing, storage centres to improve quality of pepper at the intermediary stages of the value chain, by providing land and loans for the cultivation purposes. Forecasts also depicts an important piece of information for potential investors in the pepper export market. The researchers who are interested in this field also can conduct time series regression approaches by incorporating the factors which are affecting to the pepper export income in Sri Lanka. This forecasting method can be generalized in analysing the export income of other commodities with necessary alterations.

Acknowledgement

Authors wish to thank Department of Statistics and Computer Science, Sri Lanka Exports Development Board and Mr. S.B.M.S.S. Gunarathne for their enormous contribution.

References

Institute of Policy Studies in Sri Lanka and Department of National Planning (2017). *Analysis of Cinnamon Pepper and Cardamom Value Chains in Sri Lanka*. Ministry of National Policies and Economic Affairs, Govt. of Sri Lanka.

Ministry of Development Strategies and International Trade and Sri Lanka Export Development Board. (2018). *National Export Strategy of Sri Lanka*. Govt. of Sri Lanka.

Mokshapathy, S. (2017). Production and export performance of black pepper. *International Journal of Humanities and Social Science*, **2**(**4**), 36-44.

Naz, F. (2012). A univariate times series modelling of dates exports in Pakistan. *Journal of Contemporary Issues in Business Research* ©, **1**(**2**), 57–68.

Nwogu, E. C., Iwueze, I. S. and Nlebedim, V. U. (2016). Some tests for seasonality in time series data. *Journal of Modern Applied Statistical Methods*, **15**(**2**), 382–399. https://doi.org/10.22237/jmasm/1478002920.

Ollech, D. (2019). Package 'seastests'. R package version 0.14.2. https://cran.r-project.org/web/packages/seastests/seastests.pdf

Paul, R. K., Panwar, S., Sarkar, S. K., Kumar, A., Singh, K. N., Farooqi, S. and Choudhary, V. K. (2013). Modelling and forecasting of meat exports from India. *Agricultural Economics Research Review*, **26**(**2**), 249–255.

Plywood, I. (2015). Modelling and forecasting export and import of Indian wood based panel using ARIMA models. *Elixir Statistics*, **63**, 18145-18148.

Sivarajah, P. (2018). Impact of land size on productivity, income and profits from pepper cultivation in Sri Lanka. *AGROFOR International Journal*, **1**(**3**), 127-132. https://doi.org/10.7251/AGRENG1603127S.

Sri Lanka (LKA) Exports, Imports, and Trade Partners | OEC - The Observatory of Economic Complexity. (n.d.). Retrieved on June 11, 2020, from the following link https://oec.world/en/profile/country/lka/.

Urrutia, J. D., Abdul, A. M. and Atienza, J. B. E. (2019). Forecasting Philippines imports and exports using Bayesian artificial neural network and autoregressive integrated moving average. *AIP Conference Proceedings*, **2192**, 090015-1-090015-11.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021(New Series), pp 101 –114

Forecasting Infant Mortality Rate of India Using ARIMA Model: A Comparison of Bayesian and Classical Approaches

Manika Agarwal, Praveen Kumar Tripathi and Sarla Pareek
Department of Mathematics and Statistics
Banasthali Vidyapith, Rajasthan-304 022, India

Received: 09 July 2020; Revised: 13 August 2020; Accepted: 09 October 2020

Abstract

This paper attempts to analyze the general autoregressive integrated moving average model under the classical and Bayesian paradigms. The paper aims to forecast the infant mortality rate of India under the two setups. A real data set is first examined for the presence of stationarity and is achieved by performing logarithmic scale transformation and then differencing it twice. After achieving stationarity, the most appropriate model is selected among the various competing models by using Akaike's information criterion and Bayesian information criterion. The selected model is analysed and the results in classical framework are obtained on the basis of maximum likelihood estimators. A complete Bayesian analysis is performed by using vague priors for the parameters and posterior inferences are drawn using Markov chain Monte Carlo simulation technique. The retrospective as well as prospective predictions are obtained, under the two paradigms, for infant mortality rate data and the results are, in general, found to be satisfactory.

Key words: Autoregressive integrated moving average model; Infant mortality rate; Stationarity; Akaike's information criterion; Bayesian information criterion; Markov chain Monte Carlo simulation.

AMS Subject Classifications: 37M10, 62F15, 65C05

1. Introduction

Forecasting demographic characteristics like fertility, morbidity, mortality, etc., is an important facet for the socio-economic planners as it facilitates them to analyze and regulate policies for the betterment of the human population. To forecast such characteristics require an appropriate model building so that a reliable result can be obtained. In this paper, we attempted to predict the infant mortality rate (IMR) of India in classical as well as in Bayesian paradigms. Truly speaking, IMR represents the number of deaths of children under one year of age per thousand live births. Being a vital demographic characteristic, IMR affects the population structure of a country and the projection of human population as well. Cruciality of the IMR data enforce us to model and forecast such a salient characteristic with utmost care.

Corresponding Author: Praveen Kumar Tripathi
Email: pkt2088@gmail.com

In the past few decades, a deep review of literature shows a remarkable contribution of the researchers to model and forecast mortality (see, for example, Keyfitz (1982), Pollard (1987), McNown and Rogers (1989), Lee and Carter (1992), etc.). More specifically, McNown and Rogers (1989) used a kind of parametrization of time series model to forecast mortality. Later on, McNown and Rogers (1992) employed the use of time series methods to forecast cause specific mortality. In their pioneering work Lee and Carter (1992) have proposed a probabilistic approach to model the age-specific mortality and made a long term forecast using time series methodology. Interestingly, Carter (1996) examined the stability of the 'Lee-Carter method' for structural change in a time series and made a comparison with the 'Box-Jenkins methodology' of autoregressive integrated moving average (ARIMA) process. Tuljapurkar and Boe (1998) have critically examined the change of mortality pattern and its forecasting. One should refer to Booth (2006) for a deep review of the methodologies to model and forecast the demographic components.

The use of Bayesian methodology is no more exception in time series analysis, specifically with demographic characteristics. Some of the recent works include Pedroza (2006), Reichmuth and Sarferaz (2008), Alkema and Ann (2011), Tripathi *et al.* (2018) among others. Pedroza (2006) applied a Bayesian approach in Lee-Carter model to forecast the mortality rates. Reichmuth and Sarferaz (2008) have reanalyzed the Lee-Carter model in Bayesian paradigm using the latent variable approach. Alkema and Ann (2011) used a hierarchical time series model, in Bayesian paradigm, to estimate the under-five mortality rate. Recently, Tripathi *et al.* (2018) used ARIMA model to predict the total fertility rate (TFR) of India using classical and the Bayesian approaches.

In this paper, we have applied the methodology discussed by Tripathi *et al.* (2018) for the time series based on IMR of India. In his classical work on IMR time series, Bishai (1995) has explained the issues of non-stationarity and co-integration of IMR data with the other socio-economic variables. In another study on IMR, Kurniasih *et al.* (2018) has discussed about the different methods of forecasting and their relative comparison. To forecast IMR time series data is always crucial for the view point of demographic planning and, hence, for the strategic development of the nations like India. With the same very spirit, we attempted to forecast IMR of India using ARIMA model.

Let $\{y_t\}$; $t = 1, 2, ..., T$, be a sequence of time series observations and $\{\epsilon_t\}$ is a sequence of independently and identically distributed (iid) error terms following normal distribution each with mean zero and a constant variance, say, σ^2, then the general form of the autoregressive moving average (ARMA) model of order (p,q) is given by:

$$y_t = \theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i} + \sum_{j=1}^{q} \phi_j \epsilon_{t-j} + \epsilon_t, \tag{1}$$

where θ_0 represents the intercept term and θ_i's, ϕ_j's are the autoregressive (AR) and moving average (MA) coefficients respectively.

One can further introduce a generalization of ARMA models by taking the difference of a suitable order, say, d, of the original series y_t. This new generalization is known as the integrated form of ARMA model and is given by:

$$w_t = \theta_0 + \sum_{i=1}^{p} \theta_i w_{t-i} + \sum_{j=1}^{q} \phi_j \epsilon_{t-j} + \epsilon_t, \tag{2}$$

where w_t represents the d^{th} difference of y_t. Particularly, in ARIMA model, the order of differencing decides the level of stationarity of the time series. For more details one may refer to Box *et al.* (2015).

Although, there are a number of methods available in time series literature for the model assessment, we have adopted the techniques of autocorrelation function (ACF) and partial autocorrelation function (PACF) plots in our case. This technique is proposed by Box and Jenkins (1970) in their pioneering work on ARIMA model. Truly speaking, ACF can be defined as the correlation between the two observations in a time series. It measures the linear relationship between an observation at time t and the observation at some k (say) distance apart. Slightly different from ACF, the PACF measures the degree of association between the current and a previous observation, at a distance k (say), of a time series only after removing the effects of other intermediate observations in between.

This paper proceeds as follows. The next section elaborates the data structure those based on IMR of India and the model identification on the basis of ACF and PACF. The two model selection criteria, that is, Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are also being discussed for choosing an appropriate model. The section finally ends with the numerical illustration of selected ARIMA model including both retrospective as well as prospective predictions. Section 3 explains the necessary priors setup and algorithm to draw the inferences under Bayesian paradigm. A detailed implementation of Markov chain Monte Carlo (MCMC) procedure, using Gibbs sampler and Metropolis algorithm, is also being discussed in a separate subsection. This section ends with the retrospective and prospective predictions for IMR data. The last section provides with a brief summary of the work done that concludes the whole paper.

2. Data Structure and Model Selection Criterion

We considered a real data set on IMR of India over the period of 48 years from 1971 to 2018. The data set, in the form of time series, has been collected from the SRS bulletin, Registrar General of India (see, `https://data.gov.in/resources/time-series-data-crude-death-rate-cdr-and-infant-mortality-rate-imr-srs-1971-2016`) and is framed in Table 1. To see the movement of time series, we have plotted the original data set in Figure 1. It is quite evident from the plot that the observed data on IMR shows a non-stationary pattern as it shows a consistent decline over the years. One has to ensure that the time series must achieve the stationarity to get a reliable result and for the further analyses.

To achieve the stationarity, we exercised the logarithmic scale transformation of the original series and then differenced the transformed series twice. The resultant time series, then, plotted for the different years as shown in Figure 2. The plot (Figure 2) confirms the stationarity of the time series as the mean level is constant over the years. To further strengthened our conclusion we shall provide some numerical evidences, for the differenced data set, those based on Augmented Dickey-Fuller (ADF) test and Kwaitkwoski-Phillips-Schmidt-Shin (KPSS) test. In case of ADF test the p-value is found to be 0.01 (less than 0.05)

that rejects the null hypothesis. Truly speaking, the ADF test assumes the null hypothesis that a unit root is present in the AR process. KPSS test, on the other hand, assumes the null hypothesis that the process is stationary. The p-value, in KPSS test, is found to be 0.1 (greater than 0.05) that accepts the null hypothesis. On the basis of p-values, in the two tests, one may conclude that the data is stationary in nature and the values of the test statistics are not so significant which are calculated as –5.07 for ADF test and 0.06 for KPSS test, at 5% level of significance. It is important to mention here that an AR(p) model is said to be stationary if there is no unit root present in the process, that is, if all the roots of the characteristic polynomial lie outside the unit circle (see, for example, Tripathi *et al.* (2017)).

Table 1: IMR of India from 1971 to 2018 (from left to right)

129	139	134	126	140	129	130	127	120
114	110	105	105	104	97	96	95	94
91	80	80	79	74	74	74	72	71
72	70	68	66	63	60	58	58	57
55	53	50	47	44	42	40	39	37
34	33	32						

Source: SRS bulletin, Registrar General of India

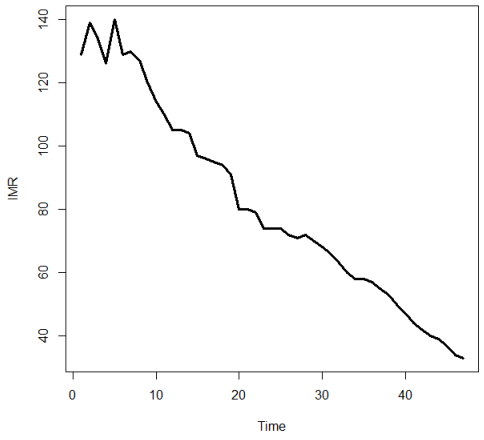


Figure 1: Time series plot for IMR data of India from 1971 to 2018

Once stationarity is achieved, we shall look forward for an appropriate model to get a reliable forecast. For this purpose, we start with the popular 'Box-Jenkins methodology' for the identification of order of ARIMA model, that is, p, d and q (see, for example, Box *et al.* (2015)). To estimate p and q we have plotted the values of ACF and PACF, respectively, against the lag values (see Figure 3). Following Box *et al.* (2015), we can observe from the Figure 3 that AR(2) and MA(1) are the most suitable choices for the components of ARIMA model. Since we have applied the second difference of the data set, therefore, the order of

differencing d is fixed as 2. Consequently, a conclusive ARIMA model can be easily assessed as ARIMA(2,2,1). For more details of the procedure one may refer to Tripathi *et al.* (2018).

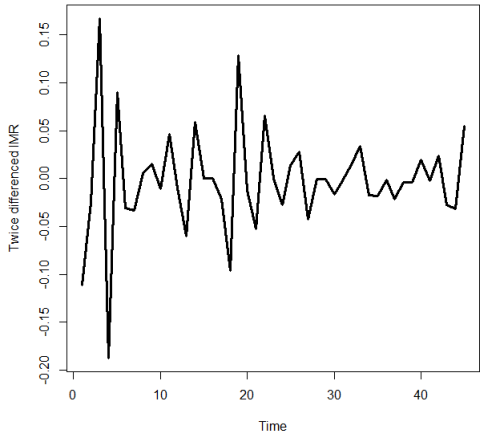


Figure 2: Time series plot for twice differenced transformed IMR data of India

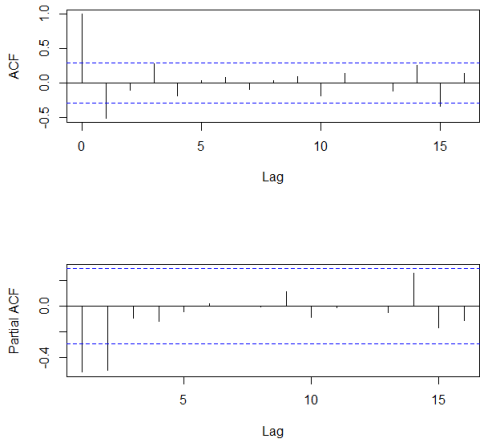


Figure 3: ACF and PACF plots for twice differenced IMR data

In order to avoid any fallacious conclusion, just on the basis of a tentative model assessed graphically, we shall consider some other nearby candidate models and a suitable model will be selected on the basis of some numerical findings. For this purpose, we used the two well known model selection criteria in statistics, namely; AIC and BIC (see, for example, Akaike (1974) and Schwarz (1978)). The two model selection criteria can be defined as below;

$$\text{AIC} = -2\log\hat{L} + 2k, \tag{3}$$

and

$$\text{BIC} = -2\log\hat{L} + k\log(T-p), \tag{4}$$

where $\hat{L}$ is the maximized likelihood function and k represents the number of parameters in the concerned model. These two criteria possess a common characteristic that they more

disagree with the model that contains large number of parameters and, hence, increase the complexity of a model. The model corresponding to the least value of AIC (BIC) is considered to be a good model and can be chosen for the purpose of further analysis. We will not go into the details of these criteria due to space restriction, but the interested candidate may refer to Ghosh *et al.* (2007) for more information. Moreover, to proceed further we have to formulate the likelihood function of the candidate models which can be accomplished by the general form of the likelihood of the ARIMA model.

Likelihood function of ARIMA model (2) can be approximately written by using the conditional density of the differenced observations, $\underline{w} : w_1, w_2, ..., w_{T-d}$ (see, for example, Tripathi *et al.* (2018)), which is given by

$$f(w_t|w_{t-1}, w_{t-2}, ..., w_{t-p}; \theta_0, \Theta, \Phi, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}(w_t - \theta_0 - \sum_{i=1}^{p} \theta_i w_{t-i} - \sum_{j=1}^{q} \phi_j \epsilon_{t-j})^2\right) \tag{5}$$

Now, we can write the approximate likelihood function of model (2), using (5), as;

$$L(\theta_0, \Theta, \Phi, \sigma^2|\underline{w}) \propto \left(\frac{1}{\sigma^2}\right)^{(T-d-p)/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=p+1}^{T-d} (w_t - \theta_0 - \sum_{i=1}^{p} \theta_i w_{t-i} - \sum_{j=1}^{q} \phi_j \epsilon_{t-j})^2\right), \tag{6}$$

where $\Theta = (\theta_1, \theta_2, ..., \theta_p)$ and $\Phi = (\phi_1, \phi_2, ..., \phi_q)$.

2.1. Model selection and prediction

As discussed in Section 2, we considered some nearby competing ARIMA models to validate our procedure. The competing models are; ARIMA(0, 2, 1), ARIMA(0, 2, 2), ARIMA(1, 2, 0), ARIMA(2, 2, 0), ARIMA(1, 2, 1), ARIMA(1, 2, 2) and ARIMA(2, 2, 2). By doing some small mathematical corrections in the expression (6), one can easily obtain the approximate maximum likelihood (ML) estimators for the parameters involved in the concerned model. All the competing models along with the ML estimates of their parameters and the values of log-likelihoods are reported in the Table 2. Since, the likelihood functions were not easily tractable, therefore, we have used numerical based approximation to obtain the ML estimates of the parameters.

Although the interpretation of results obtained in Table 2 is quite obvious still we shall highlight a few of them for the flow of analysis. The impact of intercept term on IMR data is not so considerable in all the considered models. Also, one may observe that the stationarized form of the data set shows less variability for error terms. Therefore, the fluctuations in the random component can be assumed to be constant over a period of time. We shall next proceed with the model selection for the observed stationarized form of the data.

The values of AIC and BIC, for each competing model, can be easily calculated by using the formulae (3) and (4) and the same are reported in Table 3. It is quite evident that both the criteria give their least value corresponding to ARIMA(2, 2, 0) that can be considered as the most appropriate model among others. Undoubtedly, a graphical assessment is always striking and reliable source of information still, we can not ignore the possibilities of numerical evidences which are more appealing as they consider any kind of loss due to fitting as well as complexity both. Certainly, we shall consider ARIMA(2, 2, 0) model for the further analyses

and prediction of IMR of India.

Table 2: Classical estimates of the competing ARIMA models

Model	Parameter	ML estimate	$\log \hat{L}$	Model	Parameter	ML estimate	$\log \hat{L}$
	θ_0	0.0005			θ_0	-0.0010	
ARIMA$(0,2,1)$	ϕ_1	-0.9900	-89.4400	ARIMA$(1,2,1)$	θ_1	-0.2200	-96.1500
	σ^2	0.0012			ϕ_1	-1.2200	
					σ^2	0.0008	
	θ_0	-0.0001			θ_0	-0.0005	
ARIMA$(0,2,2)$	ϕ_1	-1.0100	-86.3700	ARIMA$(2,2,1)$	θ_1	-0.5100	-90.3700
	ϕ_2	0.1400			θ_2	-0.5900	
	σ^2	0.0013			ϕ_1	-1.1800	
					σ^2	0.0009	
	θ_0	-0.0010			θ_0	0.0001	
ARIMA$(1,2,0)$	θ_1	-0.5100	-75.5600	ARIMA$(1,2,2)$	θ_1	-1.0900	-88.2300
	σ^2	0.0020			ϕ_1	0.2100	
					ϕ_2	-1.1100	
					σ^2	0.0011	
	θ_0	-0.0012			θ_0	0.0019	
ARIMA$(2,2,0)$	θ_1	-0.9100	-71.8000	ARIMA$(2,2,2)$	θ_1	-1.1500	-79.4000
	θ_2	-0.5800			θ_2	-0.5300	
	σ^2	0.0011			ϕ_1	-0.3400	
					ϕ_2	-1.0000	
					σ^2	0.0015	

Table 3: Values of AIC and BIC for the competing ARIMA models

Model	AIC	BIC
ARIMA$(0,2,1)$	184.89	190.37
ARIMA$(0,2,2)$	180.75	188.07
ARIMA$(1,2,0)$	157.13	162.55
ARIMA$(2,2,0)$	**151.60**	**158.73**
ARIMA$(1,2,1)$	200.31	207.54
ARIMA$(2,2,1)$	190.74	199.66
ARIMA$(1,2,2)$	186.46	195.49
ARIMA$(2,2,2)$	170.80	181.51

Before we extend our study to the Bayesian analysis, let us perform the classical prediction of IMR based on the selected ARIMA(2, 2, 0) model. For this purpose we took only

43 observations out of 48 observations (see Table 1) and left rest of the values as the test sample. We have obtained the necessary classical estimates for the parameters of ARIMA(2, 2, 0) model as discussed in Section 2 and predicted for the next (44^{th}) observation using the ML estimates of the parameters. This predicted observation then forms a sample of size 44 and again the next (45^{th}) observation is predicted by obtaining the corresponding ML estimates using these 44 observations in a similar way. Proceeding in this way, we can predict for all the corresponding values in the test sample.

Theoretically, to predict the future values of the original time series of size T, we have predicted the very next value for the differenced time series (of size $T-2$), that is, w_{T-1}^{th} observation. Next, we shall obtain the future value corresponding to the scaled transformed time series, that is, $\log(\hat{y}_{T+1})$ which can be calculated using the recurrence relationship given below;

$$\log(\hat{y}_{T+1}) = \hat{w}_{T-1} + 2\log(y_T) - \log(y_{T-1}), \tag{7}$$

where $\hat{w}_{T-1}$ is the estimated predictive value corresponding to w_{T-1}^{th} observation obtained by using ARIMA(2, 2, 0) process that can comfortably be obtained from (2) for the error term having the distribution $N(0, \hat{\sigma}^2)$. Hence, the estimated predictive value for the original series of IMR data set can be obtained by performing the inverse logarithmic transformation on $\log(\hat{y}_{T+1})$. To get the estimated predictive intervals, for the corresponding future values of the original series, let us represent the ML estimates of the parameters as $\hat{\sigma}^2, \hat{\theta}_1$ and $\hat{\theta}_2$. We have calculated the predictive intervals for the differenced series $\{w_t\}$ by means of the formula;

$$\hat{w}_{T-1} \pm z_{1-\alpha/2}\sqrt{Var(\hat{w}_{T-1})}, \tag{8}$$

where $z_{1-\alpha/2}$ is the standard normal percentile and $Var(\hat{w}_{T-1}) = \dfrac{\hat{\sigma}^2}{(1-\rho_1\hat{\theta}_1 - \rho_2\hat{\theta}_2)}$ with $\rho_1 = \frac{\hat{\theta}_1}{(1-\hat{\theta}_2)}$ and $\rho_2 = \hat{\theta}_2 + \frac{\hat{\theta}_1^2}{(1-\hat{\theta}_2)}$. It is to be mentioned here that the confidence interval for the scaled transformed series can be obtained by using the expression (7), just by replacing $\hat{w}_{T-1}$ with $L\hat{w}_{T-1}$ and $U\hat{w}_{T-1}$ to get lower and upper limits, respectively. Here $L\hat{w}_{T-1}$ and $U\hat{w}_{T-1}$ are the lower and upper limits of predicted intervals respectively for the differenced data. Finally, the confidence interval for the original series is obtained by using the similar inverse logarithmic transformation. Truly speaking, the expression (8) gives a $100(1-\alpha)\%$ confidence interval. To get 95% confidence interval, one may use the critical value at 0.05 level of significance from the standard normal table.

The 95% predictive intervals for the estimated predictive values, from 2014 to 2018, are given retrospectively in Table 4. We have also calculated the width of predictive intervals (ω) to observe the consistency of the prediction. It can be easily visualized that the estimated predictive values of IMR data are not too far from the true values, also the true values are well within the range of corresponding predictive intervals. The retrospective prediction is quite satisfactory and, hence, we can predict the future values prospectively. For the prospective prediction we have applied the same strategy on the whole data set and have predicted for the next five years. The future predictions, for IMR of India, are reported in the Table 5.

Table 4: Retrospective predictions of IMR from 2014 to 2018 using the ML estimates

Year	True value	Estimated predictive value	95% Estimated predictive interval		ω
2014	39.00	38.08	34.33	42.24	7.91
2015	37.00	38.08	34.33	42.24	7.91
2016	34.00	38.14	32.39	42.31	9.92
2017	33.00	36.72	32.11	40.73	8.62
2018	32.00	37.71	31.00	41.83	10.83

Table 5: Future predictions of IMR for the next 5 years using ML estimates

Year	Estimated predictive value	95% Estimated predictive interval		ω
2019	30.31	27.30	33.65	6.35
2020	30.32	27.30	33.66	6.36
2021	30.36	27.35	33.71	6.36
2022	29.25	26.34	32.47	6.13
2023	30.03	27.05	33.34	6.29

It is to be noted that the future values of IMR go down, with a good consistency, which is a good sign for a developing nation like India. Before we set up a concrete opinion about these predicted values, let us extend this study to the advanced level and perform the Bayesian analysis in the next section.

3. Bayesian Inference

The conditional likelihood function of the selected ARIMA(2,2,0) model for the differenced data is given by

$$f(\underline{w}|\theta_0,\theta_1,\theta_2,\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{(T-4)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t-\theta_0-\theta_1 w_{t-1}-\theta_2 w_{t-2})^2\right), \qquad (9)$$

where $w_t = \Delta^2 log(y_t)$. To perform the Bayesian analysis, a suitable choice of prior distribution is essential. Choosing the prior distributions is a vital aspect in Bayesian paradigm and one can use the information, if any, while selecting a prior distribution. Such priors are called informative priors. In case we have no information, the non-informative priors come into existence. For the present study, we shall consider the non-informative (vague) priors for σ^2, θ_0, θ_1 and θ_2 as we do not have any concrete information (see, for example, Tripathi *et al.* (2017)). The following prior distributions have been considered for the completion of Bayesian analysis;

$$g_1(\sigma^2) \propto \frac{1}{\sigma^2}; \qquad \sigma^2 \geq 0, \qquad (10)$$

$$g_2(\theta_0) \propto U[-N_1, N_1]; \qquad N_1 > 0, \qquad (11)$$

$$g_3(\theta_1) \propto U[-N_2, N_2]; \qquad N_2 > 0, \tag{12}$$

and

$$g_4(\theta_2) \propto U[-N_3, N_3]; \qquad N_3 > 0, \tag{13}$$

where N_1, N_2 and N_3 are the hyperparameters. One may choose the hyperparameters (large enough) in such a way that the priors remain vague over the range of parameters. We have considered the same set of values with opposite signs for the uniform range though, one can choose the different values which permit the vagueness of prior distributions. Moreover, the prior distribution for the σ^2, in (10), is a type of prior suggested by Jeffrey and is widely used by the researchers (see, for example, Marriott *et al.* (1996), Kleibergen and Hoek (2000) and Tripathi *et al.* (2018) among others).

Next, we shall obtain the joint posterior distribution by updating the prior distributions (from (10) to (13)) with the help of likelihood function (9) and it can be written up to proportionality as;

$$p(\theta_0, \theta_1, \theta_2, \sigma^2|\underline{w}) \propto \left(\frac{1}{\sigma^2}\right)^{(T-2)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t - \theta_0 - \theta_1 w_{t-1} - \theta_2 w_{t-2})^2\right) I_{[-N_1,N_1]}(\theta_0)$$
$$I_{[-N_2,N_2]}(\theta_1) I_{[-N_3,N_3]}(\theta_2), \tag{14}$$

where $I(.)$ denotes the usual indicator function that can take only two values; either zero or one. Truly speaking, if a value of the parameter falls in the interval, it will take the value one and zero otherwise.

It has been seen that the form of joint posterior (14) is analytically intractable, one has to go for the sample based approaches to get the required posterior samples from this. We, however, considered an MCMC approach and apply the Gibbs sampler with intermediate Metropolis steps. It has been seen that after a large number of iterations the sequence of parametric values converges in distribution to a random sample taken from the actual posterior distribution. For more details of the procedure one may refer to Gelfand and Smith (1991) and Upadhyay *et al.* (2001) among others. Once the posterior samples of desired size are obtained, the unobserved future value ($\hat{w}_{T-1}$) can be simulated, for each of the posterior samples, from the parent sampling distribution $p(w_{T-1}|\theta_0, \theta_1, \theta_2, \sigma^2, \underline{w})$. It can be easily verified that the predicted observation $\hat{w}_{T-1}$ follows an univariate normal distribution (see, for example, Tripathi *et al.* (2018)).

3.1. Full conditional distributions and MCMC implementation

To proceed for the MCMC implementation, let us calculate the full conditional distribution of each parameter, from the joint posterior (14), up to proportionality as below;

$$p(\theta_0|\theta_1, \theta_2, \sigma^2, \underline{w}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t - \theta_0 - \theta_1 w_{t-1} - \theta_2 w_{t-2})^2\right), \tag{15}$$

$$p(\theta_1|\theta_0, \theta_2, \sigma^2, \underline{w}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t - \theta_0 - \theta_1 w_{t-1} - \theta_2 w_{t-2})^2\right), \tag{16}$$

$$p(\theta_2|\theta_0,\theta_1,\sigma^2,\underline{w}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t-\theta_0-\theta_1 w_{t-1}-\theta_2 w_{t-2})^2\right), \tag{17}$$

$$p(\sigma^2|\theta_0,\theta_1,\theta_2,\underline{w}) \propto \left(\frac{1}{\sigma^2}\right)^{(T-2)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{T-2}(w_t-\theta_0-\theta_1 w_{t-1}-\theta_2 w_{t-2})^2\right). \tag{18}$$

Among all, the full conditional of σ^2 can be transformed into a gamma distribution by means of the transformation $\tau = 1/\sigma^2$. A simple mathematics shows that τ follows a gamma distribution with shape parameter $(T-4)/2$ and scale parameter $\frac{1}{2}\sum_{t=p+1}^{T-2}(w_t-\theta_0-\theta_1 w_{t-1}-\theta_2 w_{t-2})^2$. It is also to be noted that the full conditionals (15), (16) and (17) are not easily available in close form and direct simulation is not possible. We, therefore, adopted the Metropolis algorithm to simulate from these full conditionals. To employ the Metropolis algorithm, a univariate normal density is proposed in each case with mean value corresponding to the ML estimate of the respected parameter and standard deviation is taken to be c times the Hessian based approximation at the value of ML estimate. The constant c behaves like a tuning parameter whose value often suggested to lie between 0.5 and 1.0 (see, for example, Tripathi *et al.* (2018)). Thus, we created a single long run of the simulated values from the posterior distribution (14) via the simple implementation of the Gibbs sampler. Posterior estimates were, then, obtained by choosing the posterior samples at a regular gap, after avoiding the initial transient behavior, so that the correlation between them is close to zero. For further reading of the algorithm one may refer to Smith and Roberts (1993), Upadhyay *et al.* (2001) among others.

3.2. Numerical illustration for Bayesian analysis

To illustrate the Bayesian methodology, discussed above, we took the same set of data reported in Table 1. As discussed, we have calculated the ML estimates of the selected ARIMA(2, 2, 0) model to initialize the Markov chain. The values of hyperparameters, N_1, N_2 and N_3, were set to be 100 in each case to maintain the vagueness of the prior distributions. An iterative procedure indicated us to choose the value of tunning parameter c as 0.7. It is important to mention here that these values are chosen to get a good acceptance probability in Metropolis algorithm. For successful implementation of the Gibbs sampler, we have considered a single long run of the chain up to 50K iterations. After avoiding the initial transient behavior of the chain at about 10K iterations, we took a sample of size 1K by maintaining a gap of 40 so that the serial correlation is negligibly small. We have provided the posterior summary, for the differenced data, on the basis of these 1K posterior samples in Table 6.

The results obtained in Table 6 are self explanatory and it reveals the fact that the estimated marginal posterior densities, for all the parameters, exhibit a normal trend, that is, almost symmetrical in nature. Also, as obvious, the posterior modes are close enough to the corresponding ML estimates which might be because of the vague consideration of priors. Although, we are not giving the densities plots for the estimated parameters due to

space restriction still most of the inferences can be easily guessed from the table. Moreover, the length of highest posterior density (HPD) intervals tell the accuracy of the posterior estimates.

Table 6: Posterior summary for the parameters of ARIMA(2, 2, 0) model

Parameter	MLE	Posterior Mean	Posterior Median	Posterior Mode	0.95 HPD interval	
θ_0	−0.0012	−0.0034	−0.0035	−0.0035	−0.0143	0.0102
θ_1	−0.9110	−0.7719	−0.7710	−0.7622	−1.0235	−0.5528
θ_2	−0.5814	−0.5163	−0.5185	−0.5344	−0.7667	−0.2674
σ^2	0.0011	0.0016	0.0016	0.0016	0.0009	0.0023

Like the classical analysis, let us now work on the retrospective prediction of IMR in Bayesian framework. Again, we have considered only first 43 observations, out of 48 observations, as the informative data set (see Table 1) and rests are left to see the forecasting performance. We have applied the same strategy, as in classical prediction, to predict in Bayesian context. It is to be noted that the whole Bayesian analysis is performed repeatedly in each step of prediction until the last value is predicted. Moreover, to predict the next future value, we simulated 1K predictive samples based on 1K posterior samples and predictive summaries are drawn, for the next five values, in Table 7. It is important to know that the estimated predictive values are corresponding to the modal values of the predictive samples.

Table 7: Bayesian retrospective predictions of IMR from 2014 to 2018

Year	True value	Estimated Bayes predictive value	Estimated HPrD interval		ω
2014	39.00	38.83	35.01	41.08	6.07
2015	37.00	37.12	33.08	39.65	6.57
2016	34.00	34.46	30.61	36.80	6.19
2017	33.00	35.39	31.84	37.33	5.49
2018	32.00	38.18	31.32	39.74	8.42

It is nice to interpret that the predicted values are pretty close to the true values. Also, the 95% highest predictive density (HPrD) intervals are covering the corresponding true values nicely. Referring to Table 4, it can be inferred that the Bayes predictions, in general, appear to be more closer to the corresponding true values as compared to that on the basis of likelihood only. Also, the estimated predictive intervals in Bayesian paradigm appear to be more narrower than the classical paradigm (see the values of ω), that shows the accuracy of Bayesian analysis over the classical approach. Moreover, the widths of estimated predictive intervals ω in Table 7 look more consistent than those in Table 4.

Since the retrospective predictions (see Table 7) are found to be satisfactory, therefore, we did the prospective prediction of IMR of India using the same Bayesian methodology. For this prospective prediction, we considered the whole series (containing 48 observations) and apply the same strategy to forecast the next five observations. Table 8 provides the

future values of IMR for the next five years. It can be seen that values of IMR (Table 8) will remain close, on an average, to 30.20 which, in fact, is a good sign and showing a decreasing trend of IMR values in upcoming years. Although, in this study, we did not consider any other demographic component which effects the IMR still, our findings are very hopeful and realistic for the developing countries like India.

Table 8: Bayesian prospective predictions of IMR for the next 5 years

Year	Estimated Bayes predictive value	Estimated HPrD interval		ω
2019	30.91	27.69	32.49	4.80
2020	29.84	25.90	30.37	4.47
2021	30.23	28.40	33.57	5.17
2022	30.26	28.22	33.88	5.66
2023	29.76	27.06	32.73	5.67

4. Conclusion

This paper has successfully modelled and analysed the ARIMA model under classical and the Bayesian paradigms. The analyses resulted in retrospective as well as prospective (for the next 5 years) predictions of IMR data of India. Stationarity of the data set has been examined carefully using ADF and KPSS tests. The likelihood based estimates have been used for the classical predictions whereas, for Bayesian predictions the corresponding modal values of the parameters have been used. It is found that the latter paradigm provided us with more accurate and reliable results as compared to the former. It is expected that such an analysis will be helpful for the policy makers and researchers to come across an appropriate planning.

Acknowledgements

The authors express their thankfulness to the Editor and the Reviewers for their valuable suggestions and comments that helped to improve the earlier version of the manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19(6)**, 716–723.

Alkema, L. and Ann, W. L. (2011). Estimating the under-five mortality rate using a Bayesian hierarchical time series model. *PloS One*, **6(9)**, e23954.

Bishai, D. (1995). Infant mortality time series are random walks with drift: Are they cointegrated with socioeconomic variables? *Econometrics and Health Economics*, **4(3)**, 157–167.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, **22(3)**, 547–581.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control, Revised ed.* Holden-Day.

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

Carter, L. R. (1996). Forecasting US mortality: A comparison of Box-Jenkins ARIMA and structural time series models. *The Sociological Quarterly*, **37(1)**, 127–144.

Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations. *Communications in Statistics-Theory and Methods*, **20(5-6)**, 1747–1766.

Ghosh, J., Delampady, M., and Samanta, T. (2007). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer Science and Business Media.

Keyfitz, N. (1982). Choice of function for mortality analysis: Effective forecasting depends on a minimum parameter representation. *Theoretical Population Biology*, **21(3)**, 329–352.

Kleibergen, F. R. and Hoek, H. (2000). Bayesian analysis of ARMA models. *Tinbergen Institute Discussion Paper* Amsterdam, the Netherlands.

Kurniasih, N., Ahmar, A. S., Hidayat, D. R., Agustin, H., and Rizal, E. (2018). Forecasting infant mortality rate for China: A comparison between α-sutte indicator, ARIMA and Holt-Winters. *Journal of Physics: Conference Series*, **1028(1)**, 012195.

Lee, R. D. and Carter, L. R. (1992). Modelling and forecasting US mortality. *Journal of the American Statistical Association*, **87(419)**, 659–671.

Marriott, J., Ravishanker, N., Gelfand, A., and Pai, J. (1996). Bayesian analysis of ARMA processes: Complete sampling-based inference under exact likelihoods. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, 243–256.

McNown, R. and Rogers, A. (1989). Forecasting mortality: A parametrized time series approach. *Demography*, **26(4)**, 645–660.

McNown, R. and Rogers, A. (1992). Forecasting cause-specific mortality using time series methods. *International Journal of Forecasting*, **8(3)**, 413–432.

Pedroza, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics*, **7(4)**, 530–550.

Pollard, J. H. (1987). Projection of age-specific mortality rates. *Population Bulletin of the United Nations*, **No. 21/22**, 55–69.

Reichmuth, W. H. and Sarferaz, S. (2008). Modelling and forecasting age-specific mortality: A Bayesian approach. *SFB 649 Discussion Paper*, **(No. 2008, 052a)**.

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

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, **B 55**(**1**), 3–23.

Tripathi, P. K., Mishra, R. K., and Upadhyay, S. K. (2018). Bayes and classical prediction of total fertility rate of India using autoregressive integrated moving average model. *Journal of Statistics Applications and Probability*, **7(2)**, 233–244.

Tripathi, P. K., Ranjan, R., Pant, R., and Upadhyay, S. K. (2017). An approximate Bayes analysis of ARMA model for Indian GDP growth rate data. *Journal of Statistics and Management Systems*, **20(3)**, 399–419.

Tuljapurkar, S. and Boe, C. (1998). Mortality change and forecasting. *North American Actuarial Journal*, **2(4)**, 13–47.

Upadhyay, S. K., Vasishta, N., and Smith, A. F. M. (2001). Bayes inference in life testing and reliability via Markov chain Monte Carlo simulation. *Sankhya: The Indian Journal of Statistics*, **A 63(1)**, 15–40.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021(New Series), pp 115–125

On Association in Time of Markov Process With Application to Reliability and Survival Analysis

M. Manoharan[1] and V.M. Chacko[2]
[1]*Department of Statistics,University of Calicut, Kerala.*
[2]*Department of Statistics, St.Thomas College(Autonomous), Thrissur, Kerala.*

Received: 17 June 2020; Revised: 08 October 2020; Accepted: 14 October 2020

Abstract

In this paper, the association in time of a Markov process is considered. A measure based on transition probability function is proposed to obtain and compare the degree of association in time of two processes. A real data is analyzed.

Key words: Association; Degree of association; Transition probability function; Reliability; Markov process; Multistate system; Right tail increasing

AMS Subject Classifications: 60J25, 60J28

1. Introduction

For a multistate system it is difficult to calculate system reliability. The calculation of system reliability can become even more difficult, perhaps impossible, if the components of the system are maintained, or are interdependent. Hence, reliability bounds for reliability of multistate systems (MSSs) are useful. In reliability and survival analysis, the lifetime random variables are not independent but are associated. In structures in which the components share load, the failure of one component results in increased load on each of the remaining components. Associated random variables and time associated stochastic processes are useful for obtaining the reliability bounds for MSSs. Association of random variables is mathematically described by Esary *et al.* (1967). Esary and Proschan (1970) obtained a minimal cut lower bound for a non-maintained system, if the joint performance process of the components is associated in time. A repairable system modeled by semi-Markov process is considered by Dharmadhikari and Kuber (2006) and derived a sufficient condition for the association in time of the process governing the system. Hjort *et al.* (1985) introduced a sufficient condition for association when the marginal processes are Markovian. Bound for reliability of maintained systems without imposing conditions of association in time of marginal process of components is given by Natvig (1993). Minimal path structures of a coherent system having components in common behave in a similar manner, so that failure of a component will adversely affect the performance of all the minimal path structures.

A sufficient condition for association in time of the Markov performance process of a binary system, in terms of its transition probability functions, is given by Esary and Proschan (1970). Reliability analysis of MSSs can be seen in Barlow and Wu (1978).

Corresponding Author: V. M. Chacko
Email: chackovm@gmail.com

Relative degree (or strength) of association for two sets of random variables is described by Karlin (1983). In order to assess the degree of association of a Markov process or of comparing the relative strength of association of two Markov processes, we need measures in terms of transition probability functions. In this paper, in order to find degree of association of the Markov process or to compare the relative degree of association of two Markov processes, we consider a measure of association based on transition probability function. In MSS modeling, the information regarding exact state visited by components before the present state may not be available. At the same time, information regarding either the state is above or below a certain state of performance may be available. The conditions for association in time of the stochastic process which governs the MSS is still worthwhile. A weaker sufficient condition for association in time of the finite Markov process is given.

In section 2, we discuss the measure of degree of association in time of the Markov process in terms of the transition probability function. We examine the correlation in terms of transition probability functions to asses the relative degree (or strength) of association when comparing two Markov processes. In section 3, we present the weaker sufficient condition in terms of transition functions and intensities for the Markov process to be associated in time. An illustrative example is provided in section 4.

2. Measure of Degree of Association

An approach for assessing the level and form of dependence for multivariate observations is provided by Karlin (1983). It provides a fine tuning in evaluating relationships of pair of random variable by transforming the data in natural manifold ways and then computing the associated correlations whose totality reflects on the nature of dependence between array of transformed variables. The degree of dependence between two random variables X and Y can be computated by a single statistics.

The following definition gives the measure for ordering bivariate distributions by the strength of their association.

Definition 1: For two bivariate distributions corresponding to the random variables (X, Y) and (Z, W) we say that dependence of (X, Y) is stronger than the dependence of (Z, W) with respect to classes of non-decreasing functions F and G if $\rho[h(X), g(Y)] \geq \rho[h(Z), g(W)]$ for all $h \in \mathrm{F} and g \in \mathrm{G}$.

The comparisons are made with respect to the same transformations on the variables (X, Y) and (Z, W) for all functions $h \in \mathrm{F} and g \in \mathrm{G}$.

A measure which can be used to measure the degree of association of the Markov process is proposed below. We first discuss the measure of degree of association in discrete time stochastic process $\{X_k, k \geq 0\}$ with state space $E = \{1, 2, ..., M\}$. We have,
$\mathrm{Cov}(\mathrm{X}_k, X_{k-1}) = E(X_k.X_{k-1}) - E(X_k)E(X_{k-1})$
$= \sum_{i,j \in E} P(X_k \geq j, X_{k-1} \geq i) - \sum_{j \in E} P(X_k \geq j) \sum_{i \in E} P(X_{k-1} \geq i)$.

But, X_k and X_{k-1}, associated if, $Cov(X_k, X_{k-1}) \geq 0$.
$\Rightarrow \left(\sum_{i,j\in E} P(X_k \geq j, X_{k-1} \geq i) - \sum_{i,j\in E} P(X_k \geq j)P(X_{k-1} \geq i)\right) \geq 0$
$\Rightarrow \sum_{i,j\in E} [P(X_k \geq j, X_{k-1} \geq i) - P(X_k \geq j)P(X_{k-1} \geq i)] \geq 0$
or

$$\sum_{i,j\in E} [P(X_k \geq j|X_{k-1} \geq i) - P(X_k \geq j)]P(X_{k-1} \geq i) \geq 0. \quad (1)$$

Using one step transition probability, we get,

$$\sum_{i,j\in E} \sum_{i,j\in E} [P(X_k = j|X_{k-1} = i) - P(X_k = j)]P(X_{k-1} = i) \geq 0 \quad (2)$$

We can use the measure, $\text{Cov}(X_k, X_{k-1}) = \sum_{i,j\in E} \sum_{i,j\in E} [P(X_k = j|X_{k-1} = i) - P(X_k = j)]P(X_{k-1} = i)$, for assessing the association of the discrete time stochastic process.

Standardization of the covariance may be desired to achieve scale invariance and enable meaningful comparisons between different data sets. The condition of association,

$$Cov(h(X), g(Y)) \geq 0$$

for all functions $h \in F$ and $g \in G$, of two random variables with respect to the classes F and G is replaced by an equivalent requirement $\rho(X, Y) = \frac{Cov(h(X),g(Y))}{\sqrt{Var(h(X)).Var(g(Y))}} \geq 0$. For two stochastic processes $\{X_k, k \geq 0\}$ and $\{Y_k, k \geq 0\}$, the following measure of association is used for comparing two processes in terms of their strength of association. $\rho_{(X_k,X_{k-1})} = \frac{Cov(X_k,X_{k-1})}{\sqrt{Var(X_k).Var(X_{k-1})}}$ where $Var(X_k) = \sum_{i,j\in E} P(X_k \geq max(i, j)) - P(X_k \geq j)P(X_k \geq i)$.

If

$$\rho_{(X_k,X_{k-1})} \geq \rho_{(Y_k,Y_{k-1})} \quad (3)$$

the association between X_k and X_{k-1} is larger than association between Y_k and Y_{k-1}. If (3) is true for every k, then the stochastic process $\{X_k, k \geq 0\}$ is more associated than $\{Y_k, k \geq 0\}$.

Here we also consider a continuous time Markov process $\{X(t), t \geq 0\}$. Consider the random variables $X(t)$, $X(s)$, $s < t$ in the Markov process. It is clear that if $X(t)$ and $X(s)$, $s < t$ are associated if

$$Cov(X(t), X(s)) = \int_R \int_R P(X(t) > x, X(s) > y) - P(X(t) > x)P(X(s) > y)dxdy \geq 0. \quad (4)$$

Using transition probability function, $P(X(t) = j|X(s) = i)$ of the Markov process, we write (4) as,
$Cov(X(t), X(s)) =$

$$\int_R \int_R \sum_{\{i,j:X(s)=i>y,X(t)=j>x\}} [P(X(t) = j|X(s) = i) - P(X(t) = j)]P(X(s) = i)dxdy \geq 0 \quad (5)$$

Comparison of two Markov processes, $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$, only in terms of transition probabilities is not possible but comparison between covariances in terms of transition

probabilities and state probabilities is more reasonable. The degree of association of two Markov processes can be compared using the following correlation function, $\rho_X(t,s)$.

In the case of the Markov process we have, $\rho_X(t,s) = \frac{Cov(X(t),X(s))}{\sqrt{Var(X(t)).Var(X(s))}} \geq 0$ implies association between $X(s)$ and $X(t)$. We use the correlation $\rho_X(t,s)$ as a function of transition probability function and state probabilities to measure the degree of association in time of the Markov process.

We compare the degree of association of two Markov processes using $\rho_X(t,s)$. This gives a stochastic ordering of two Markov processes based on strength of their association. Denote, $\mathrm{C}_X(t,s) = \int_R \int_R [P(X(t) \geq x | X(s) \geq y) - P(X(t) \geq x)] P(X(s) \geq y) dxdy$,
$C_X(t,t) = \int_R \int_R [P(X(t) \geq max(x,y)) - P(X(t) \geq x) P(X(t) \geq y)] dxdy$
$\rho_X(t,s) = \frac{C_X(t,s)}{\sqrt{C_X(t,t).C_X(s,s)}}$. We propose the following definitions.

Definition 2: For two different Markov processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$, we say that association of $(X(t), X(s))$, $s < t$ is stronger than the association of $(Y(t), Y(s))$, $s < t$ if $\rho_X(t,s) \geq \rho_Y(t,s)$.

Definition 3: For two different Markov processes $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$, we say that association of X process is stronger than the association of Y process if $\forall s, t \in R$, $s < t$, $\rho_X(t,s) \geq \rho_Y(t,s)$. Some conditions of association in terms of the nondecreasing functions of the classes F and G and its distributional properties are given below. It provide a measure for the comparison of the degree of association of two system each consists of n associated components, see Prakash Rao and Dewan (2001).

Definition 4: A collection of random variables $\{X_n, n \geq 1\}$ is said to be associated if for every n and for every choice of coordinate-wise non-decreasing functions $h(\underline{x})$ and $g(\underline{x})$ from R^n to R,

$$Cov(h(\underline{X}), g(\underline{X})) \geq 0 \tag{6}$$

whenever it exist, where $\underline{X} = (X_1, ..., X_n)$.

Definition 5: The performance process of the ith component is a stochastic process $\{X_i(t), t \in \tau\}$ where for each fixed $t \in \tau, X_i(t)$ denotes the state of component i at time t. The joint performance process of the components is given by $\{\underline{X}(t), t \in \tau\} = \{(X_1(t), ..., X_n(t)), t \in \tau\}$.

Let $I = [t_A, t_B] \subset [0, \infty), \tau(I) = \tau \cap I$.

Definition 6: The joint performance process $\{\underline{X}(t),\ t \in \tau\}$ of the components is said to be associated in time interval I if and only if, for any integer m and $\{t_1, ..., t_m\} \subset \tau(I)$, the random variables in the array

$$\begin{matrix} X_1(t_1) & ... & X_1(t_m) \\ ... & ... & ... \\ X_n(t_1) & ... & X_n(t_m) \end{matrix}$$

are associated.

For the component performance process $\{X_i(t), t \in \tau\}$, $i \in \{1, 2, ..., n\}$ and fixed $t_1 < ... < t_m$, let $h_i(\underline{X}_i) \in F, g_i(\underline{X}_i) \in G$ are nondecreasing function of random variables from R^m to R, where $\underline{X}_i = (X_i(t_1), X_i(t_2), ..., X_i(t_m))$. $X_i(t_1), X_i(t_2), ..., X_i(t_m)$ are associated if for every $h_i(\underline{x}_i), g_i(\underline{x}_i)$, $Cov(h_i(\underline{X}_i), g_i(\underline{X}_i)) \geq 0$ where $\mathrm{Cov}(\mathrm{h}_i(\underline{X}_i), g_i(\underline{X}_i)) = \int_R \int_R [P(h_i(\underline{X}_i) > x, g_i(\underline{X}_i) > y) - P(h_i(\underline{X}_i) > x)P(g_i(\underline{X}_i) > y)]dxdy$, $i \in \{1, 2, ..., n\}$. This implies that if

$$\int_R \int_R [P(h_i(\underline{X}_i) > x | g_i(\underline{X}_i) > y) - P(h_i(\underline{X}_i) > x)]P(g_i(\underline{X}_i) > y)dxdy \geq 0 \tag{7}$$

we have association of the component performance processes $\{X_i(t), t \in \tau\}$.

Definition 7: A Markov performance process $\{X_i(t), t \in \tau\}$ of component i is associated if $\int_R \int_R [P(h_i(\underline{X}_i) > x | g_i(\underline{X}_i) > y) - P(h_i(\underline{X}_i) > x)]P(g_i(\underline{X}_i) > y)dxdy \geq 0$ for every collection of random variables $\underline{X}_i = (X_i(t_1), ..., X_i(t_m))$ and every choice of coordinate wise nondecreasing function $h_i(\underline{x}_i)$ and $g_i(\underline{x}_i)$ from R^m to R.

In a similar way, we can find a condition for association of joint performance process of components, in terms of non-decreasing functions, which is quite desirable. In the following definition, we consider the nondecreasing functions $H \in F$ and $G \in G$ from R^{nm} to R.

Definition 8: The joint performance process of the components $\{\underline{\mathrm{X}}(\mathrm{t}), \mathrm{t} \in \tau\}$ $= \{(X_1(t), ..., X_n(t)), t \in \tau\}$ is associated in time if

$$\int_R \int_R [P(H(\underline{X}) > x | G(\underline{X}) > y) - P(H(\underline{X}) > x)]P(G(\underline{X}) > y)dxdy \geq 0$$

for every collection of random variables,

$$\underline{X} = (X_1(t_1), X_2(t_1), ..., X_n(t_1), X_1(t_2), ..., X_n(t_2), ..., X_1(t_m), ..., X_n(t_m))$$

and every choice of coordinate wise nondecreasing function $H(\underline{x})$ and $G(\underline{x})$ from R^{nm} to R.

The measure of degree of association of the system which consists of n associated components governed by Markov processes is given below.

Denote

$$C_{\underline{X}}(H, G) = \int_R \int_R [P(H(\underline{X}) > x | G(\underline{X}) > y) - P(H(\underline{X}) > x)]P(G(\underline{X}) > y)dxdy$$

$$C_{\underline{X}}(H, H) = \int_R \int_R [P(H(\underline{X}) > max\{x, y\}) - P(H(\underline{X}) > x)P(H(\underline{X}) > y)]dxdy$$

and

$$\rho_{\underline{X}}(H,G) = \frac{C_{\underline{X}}(H,G)}{\sqrt{C_{\underline{X}}(H,H)C_{\underline{X}}(G,G)}}$$

The degree of association of two performance processes $\{\underline{X}(t), t \geq 0\}$ and $\{\underline{Y}(t),\ t \geq 0\}$ of two systems can be compared using the measures $\rho_{\underline{X}}(H,G)$ and $\rho_{\underline{Y}}(H,G)$.

Definition 9: For two performance process $\{\underline{X}(t), t \geq 0\}$ and $\{\underline{Y}(t), t \geq 0\}$, of two systems consists of n associated components governed by the Markov processes $\{\underline{X}_i(t), t \geq 0\}$ and $\{\underline{Y}_i(t), t \geq 0\}$, $i = 1, 2, ..., n$ respectively, we say that association of X-system is stronger than the association of Y-system if $\forall m, n \, and \, H \in F$, $G \in G$, from R^{mn} to R, $\rho_{\underline{X}}(H,G) \geq \rho_{\underline{Y}}(H,G)$.

The proposed measures may help us (i) to suggest whether a Markov process is associated in time; and (ii) to asses the relative degree (or strength) of association of two different Markov processes, and (iii) to asses the relative strength of association of two performance process of two systems consists of n associated components which are governed by Markov processes.

3. A Weaker Condition for Association in Time of a Markov Process

A sufficient condition using transition probability function for association with the marginal Markovian processes is given by Hjort *et al.* (1985). We consider much weaker conditions for the Markov process to be associated in time. Let $\mathrm{P}^*_{ij}(s,t) = P(X(t) = j | X(s) \geq i)$, $s < t$and $\mathbf{P}^*(s,t) = \{P^*_{ij}(s,t)\}_{i,j \in \{0,1,...,M\}}$, $s < t$ Assume the existence of the following intensities

$$\mu^*_{ij}(s) = \begin{cases} lim_{h \to 0+} \frac{P^*_{ij}(s,s+h)}{h}, & i \neq j \\ lim_{h \to 0+} \frac{P^*_{ij}(s,s+h)-1}{h}, & i = j \end{cases}$$

Let $\mathrm{P}^*_{i,\geq j}(s,t) = P(X(t) \geq j | X(s) \geq i)$
$P^*_{i,<j}(s,t) = P(X(t) < j | X(s) \geq i)$
$\mu^*_{i,\geq j}(s) = \sum_{\nu=j}^{M} \mu^*_{i\nu}(s), i < j$
$\mu^*_{i,<j}(s) = \sum_{\nu=0}^{j-1} \mu^*_{i\nu}(s), i \geq j$

Now we consider the following definitions, see Barlow and Proschan (1975).

Definition 10: A random variable T is stochastically right tail increasing (st. RTI) in random variables $S_1, ..., S_k$ if $P(T > t | S_1 \geq s_1, ..., S_k \geq s_k)$ is nondecreasing in $s_1, s_2, ..., s_k$.

Definition 11: Random variables $T_1, ..., T_n$ are conditionally RTI in sequence if T_i is st. RTI in $T_1, ..., T_{i-1}$ for $i = 2, 3, ..., n$.

Definition 12: A process $\{X(t), t \geq 0\}$ is conditionally RTI in time if $\mathrm{P(X(t)} \geq j | X(s_1) \geq i_1, ..., X(s_n) \geq i_n)$ is nondecreasing in $i_1, ..., i_n$ for each j and for each choices of $s_1 < ... < s_n < t, n \geq 1$.

Manoharan (1995) proved the following result.

Theorem 1: If the random variables $T_1, T_2, ..., T_n$ are conditionally RTI in sequence, then

they are associated.

Now using the Definition 12 and the above Theorem 1, we get the following result.

Theorem 2: If a stochastic process $\{X(t), t \geq 0\}$ is conditionally RTI in time, then it is associated in time.

A main result of this section which gives a weaker condition for the Markov process to be associated in time is given below.

Theorem 3: Let X be a continuous time Markov process with state space $\{0, 1, ..., M\}$. Assume $\mu^*_{ij}(s)$ to be continuous. Then each of the following three conditions are equivalent and implay that X is associated in time

(a) X is conditionally RTI in time.

(b) $P^*_{i,\geq j}(s,t)$ is nondecreasing in i for each j and for each choice of $s < t$.

(c) For each j and s

$\mu^*_{i,\geq j}(s)$ is nondecreasing in $i \in \{0, 1, ..., j-1\}$

$\mu^*_{i,<j}(s)$ is nonincreasing in $i \in \{j, j+1, ..., M\}$

Proof: In view of Theorem 2, it suffices to prove the equivalence of conditions (a), (b) and (c).

The equivalence of (a) and (b) follows from the Markov property of X.

Now to prove the equivalence of (b) and (c), note that statement (b) is equivalent to the following three conditions.

(i)$P(X(t) \geq j | X(s) \geq i) = P^*_{i,\geq j}(s,t)$ is nondecreasing in $i \in \{0, 1, 2..., j-1\}$

(ii)$P(X(t) < j | X(s) \geq i) = P^*_{i,<j}(s,t)$ is nonincreasing in $i \in \{j, j+1, ..., M\}$.

(iii)$P(X(t) \geq j | X(s) \geq j-1) \leq P(X(t) \geq j | X(s) \geq j)$

Thus if (b) holds then for $i < j$, $\mu^*_{i,\geq j}(s) = \sum_{\nu=j}^{M} \mu^*_{i\nu}(s)$
$= \sum_{\nu=j}^{M} lim_{h\to 0+} \frac{P^*_{ij}(s,s+h)}{h} = lim_{h\to 0^+} \frac{P^*_{i,\geq j}(s,s+h)}{h}$ is nondecreasing in $i \in \{0, 1, ..., j-1\}$, and for $i \geq j$

$$\mu^*_{i,<j}(s) = \sum_{\nu=0}^{j-1} \mu^*_{i\nu}(s) = lim_{h\to 0^+} \frac{P^*_{i,<j}(s, s+h)}{h}$$

is nonincreasing in $i \in \{j, j+1, ..., M\}$. Hence (b) implies (c).

To show that (c) implies (b), let $\mathbf{M}^*$ denote the class of all stochastic matrices $\mathbf{P}^* = (P^*_{ij})_{i,j\in\{0,1,...,M\}}$ such that $\sum_{\nu=j}^{M} P^*_{i\nu}$ is nondecreasing in i for each j. In order to prove that X has property (b) it is enough to show that $\mathbf{P}^*(s,t) \in M^*$ for each choice $s < t$.

Define

$$a_{ij}^*(u) = \begin{cases} \mu_{ij}^*(u), & i \neq j \\ 1 - \sum_{j \neq i} \mu_{ij}^*(u), & i = j \end{cases}$$

and let $\mathbf{A}^*(u) = (a_{ij}^*(u))_{i,j \in \{0,1,...,M\}}$. Also let $\mathbf{Q}^*(u) = \mathbf{A}^*(u) - \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Using the product integral representation (see Johansen (1977)) and the fact that the transition intensities are uniformly continuous on $[s,t]$, we have

$$\mathbf{P}^*(s,t) = lim_{n\to\infty} \prod_{j=0}^{n-1} [\mathbf{I} + \mathbf{Q}^*(s + (j/n)(t-s))(t-s)/n] \tag{8}$$

Note that $\mathbf{B}(u,h) = \mathbf{I} + \mathbf{Q}^*(u)h = (1-h)\mathbf{I} + \mathbf{A}^*(u)h$ is a stochastic matrix. Now if (c) is satisfied, $\sum_{\nu=j}^{M} b_{i\nu}(u,h) = \sum_{\nu=j}^{M} q_{i\nu}^*(u)h = \sum_{\nu=j}^{M} \mu_{i\nu}^*(u)h$ is nondecreasing in $i \in \{0,1,...,j-1\}$, and $\sum_{\nu=0}^{j-1} b_{i\nu}(u,h) = \sum_{\nu=0}^{j-1} \mu_{i\nu}^*(u)h$ is nonincreasing in $i \in \{j, j+1, ..., M\}$. Also $\sum_{\nu=j}^{M} b_{j\nu}(u,h) - \sum_{\nu=j}^{M} b_{j-1,\nu}(u,h) = (1-h) + h a_{jj}^*(u) + h \sum_{\nu=j+1}^{M} a_{j\nu}^*(u) - h \sum_{\nu=j}^{M} a_{j-1,\nu}^*(u)$
$= (1-h) + h \left[1 - \sum_{l \neq j} \mu_{jl}^*(u)\right] + h \sum_{\nu=j+1}^{M} \mu_{j\nu}^*(u) - h \sum_{\nu=j}^{M} \mu_{j-1,\nu}^*(u)$
$= 1 + h \left[\sum_{\nu=j+1}^{M} \mu_{j\nu}^*(u) - \sum_{\nu=j}^{M} \mu_{j-1,\nu}^*(u) - \sum_{l \neq j} \mu_{jl}^*(u)\right]$ Now choose h small enough so that $\sum_{\nu=j}^{M} b_{j\nu}(u,h) - \sum_{\nu=j}^{M} b_{j-1,\nu}(u,h) \geq 0 \, \forall j = 1, 2, ..., M$.Since μ_{ij}^*'s are bounded, we can choose h independent of u. Hence for sufficiently small h (independent of u),

$$\mathbf{B}(u,h) = \mathbf{I} + \mathbf{Q}^*(u)h$$

satisfies the conditions (i), (ii) and (iii), which means that $\mathbf{B}(u,h) \in M^*$. The class M^* being closed under multiplication and also under pointwise limits, we conclude from (8) that $P^*(s,t) \in M^*$. Hence (b) is true.

Remark 1: It can be easily seen that the conditions for the association of Markov process in Hjort *et al.* (1985) imply the conditions (a) to (c) of the Theorem 3 and hence the latter set gives a much weaker conditions for association in time of a Markov process.

Remark 2: For the binary reliability system (M=1) it is easily seen that condition (b) of the Theorem 3 is equivalent to $P_{11}(s,t) + P_{00}(s,t) \geq 1 \, for \, each \quad s < t$ which is the sufficient condition for the association in time of X given by Esary and Proschan (1970). Furthermore, when the transition intensities are continuous, the condition (c) of the Theorem 3 is always satisfied and hence the corresponding Markov process is always associated.

Remark 3: One may have further extension of the conditions (a)-(c) of the Theorem 3 for a semi-Markov process by augmenting the waiting time variable to the state variable as dealt in Kuber and Dharmadhikari (1996).

4. Application

We consider the data set from medical field for the illustration of concept of measure of degree of association in Markov processes.

Example 1: We re-examine the data on an oral hygiene study, discussed in Das and Chattopadhyay (2004)(*cf.* Dharmadhikari and Dewan (2006)) for the illustration of the association of a vector valued process. The reduction in the amount of plaque on teeth is

recorded. Each individual in the data was monitored for a couple of days. Two teeth were identified, one on the left lower canine which is in the left lower corner of a jaw, and one on molar at upper right jaw. The reduction in the thickness of plaque for subjects are usually recorded as belonging to four different categories, *viz*, no reduction, slight reduction, moderate reduction and vast reduction. To evaluate effectiveness of brushing, we use the proposed measures. To check whether it is possible to reduce the number of records per individual per day and there is some sort of dependence Das and Chattopadhyay (2004) developed a latent mixture regression model to study this categorical multivariate data. Table A.1 give a part of dental data analyzed. It gives stain on the same tooth at all four positions before and after brushing, respectively. Numbers under (P_1, P_2, P_3, P_4) indicate the amount of stain at each of the four positions on the selected tooth of an individual. The data in Table A.1 are conditionally increasing in its coordinates.

The state probabilities are given in Table A.2. The conditional probabilities $P(X(t) = j|X(s) = i)$ for $i, j \in \{0, 1, 2, 3\}$ for the four sets of data are calculated in Table A.3.

To get an ordering in terms of association we have to compute the measure of association. The values are obtained in Table A.4.

This shows that the data in the third (P_3) position is more associated. This information may be useful to medical practitioners.

5. Summary

The degree of association in time of a Markov process can be measures using proposed measures which are based on transition probability function. The measure can be used to compare two Markov process according to the degree of association. A weaker condition for association of a Markov process in time is derived. The proposed measure can be used in various areas such as engineering, medical, social science *etc.*

Acknowledgements

We are grateful to an anonymous referee who suggested many improvements and furthermore generously listed many useful references. Finally, we thank the Chair Editor for his encouragement, guidance and counsel.

References

Barlow, R. E., and Proschan, F., (1975). *Statistical Theory of Reliability and Life Testing.* New York: Holt, Rinehart and Winston.

Barlow, R. E., and Wu, A. S., (1978). Coherent system with multistate components. *Mathematics of Operations Research,* **3**, 275-281.

Dharmadhikari, A. D. and Dewan, I. (2006). Association in time of a vector valued process. *Statistics and Probability Letters,* **76**, 1147-1155.

Esary, J. D. and Proschan, F. (1970). A reliability bound for systems of maintained independent components. *Journal of the American Statistical Association*, **65**, 329-398.

Esary, J. D., Proschan, F. and Walkup, D. W. (1967). Association of random variables, with applications. *Annals of Mathematical Statistics*, **38**, 1466-1474.

Hjort, N. B., Natvig, B. and Funnemark, E.(1985). The association in time of a Markov process with applications to multistate reliability theory. *Journal of Applied Probability*, **22**, 473-479.

Johansen, S. (1977). *Product Integrals and Markov processes*, Preprint No. 3, University of Copenhagen, Institute of Mathematical Statistics.

Karlin, S. (1983). Association arrays in assessing forms of dependence between bivariate random variables. *Proceedings of the National Academy of Sciences*, USA, **80**, 647-651.

Kuber, M and Dharmadhikari, A. D. (1996). Association in time of a finite semi-Markov process. *Statistics and Probability Letters*, **26**, 125-133.

Manoharan, M. (1995). On a sufficient condition for association in time of a process, *Proceedings of the XV Indian Society of Probability and Statistics Conference* (M.S. University, Tirunelveli, India), 65-66.

Natvig, B. (1993). Strict and exact bounds for the availabilities in a fixed time interval for multistate monotone systems. *Scandanavian Journal of Statistics*, **20**, 171-175.

Prakasa Rao, B. L. S. and Dewan, I. (2001). *Associated sequences and related inference problems.* Elsevier, Shanbhag, D. N. and Rao, C. R. (Eds.), *Handbook of Statistics*, North-Holland, Amsterdam, **19**, 693-729.

Appendix A

Table A.1: Dental data stain before and after brushing

	Before brushing				After brushing					Before brushing				After brushing			
Individual	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4	Individual	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
1	1	1	1	2	0	0	0	0	14	2	1	2	2	0	0	1	1
2	1	1	2	2	0	0	0	1	15	2	2	2	2	0	0	1	1
3	1	1	2	2	0	0	0	1	16	2	2	2	2	0	0	1	1
4	1	1	2	2	0	0	0	1	17	2	2	2	2	0	0	1	1
5	1	1	2	2	0	0	0	1	18	2	2	2	2	0	0	1	1
6	1	2	2	2	0	0	0	1	19	2	2	2	2	0	0	1	1
7	1	2	2	2	0	0	0	1	20	2	2	2	2	0	0	1	1
8	1	2	2	2	0	0	0	1	21	2	2	2	2	0	1	1	1
9	1	2	2	2	0	0	0	1	22	2	2	2	2	0	1	1	1
10	1	2	2	2	0	0	0	1	23	2	2	2	2	0	1	1	1
11	1	2	2	2	0	0	0	1	24	2	2	2	3	0	1	1	1
12	1	2	2	2	0	0	0	2	25	2	2	2	3	1	1	1	2
13	1	2	2	3	0	0	0	2									

Table A.2: State probabilities

P_1	P_2	P_3	P_4
$P(X(s)=1)=13/25$	$P(X(s)=1)=6/25$	$P(X(s)=1)=1/25$	$P(X(s)=1)=0$
$P(X(s)=2)=12/25$	$P(X(s)=2)=19/25$	$P(X(s)=2)=24/25$	$P(X(s)=2)=22/25$
$P(X(s)=3)=0$	$P(X(s)=3)=0$	$P(X(s)=3)=0$	$P(X(s)=3)=3/25$
$P(X(t)=0)=24/25$	$P(X(t)=0)=20/25$	$P(X(t)=0)=13/25$	$P(X(t)=0)=1/25$
$P(X(t)=1)=1/25$	$P(X(t)=1)=5/25$	$P(X(t)=1)=22/25$	$P(X(t)=1)=21/25$
$P(X(t)=3)=0$	$P(X(t)=3)=0$	$P(X(t)=3)=0$	$P(X(t)=2)=3/25$

Table A.3: The conditional probabilities $P(X(t)=j|X(s)=i)$ for $i,j\in\{0,1,2,3\}$

$X(t)$ / $X(s)$	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
2	$\frac{11}{12}$	$\frac{1}{12}$	0	0	$\frac{14}{19}$	$\frac{5}{19}$	0	0	$\frac{12}{24}$	$\frac{12}{24}$	0	0	$\frac{1}{22}$	$\frac{20}{22}$	$\frac{1}{22}$	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{3}$	$\frac{2}{3}$	0

Table A.4: Covariance

Position	P_1	P_2	P_3	P_4
$Cov(X(t),X(s))$	0	30/625	12/25	2/25

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 127-138

Estimation of Cure Fraction and Misclassification Probabilities Using Continuous Time Hidden Markov Model

Gurprit Grover[1], Sangeeta Chakravarty[2], Arpan Kumar Thakur[1]
[1]*Department of Statistics, University of Delhi, Delhi- 110007, India*
[2]*Institute of Economic Growth, Delhi- 110007, India.*

Received: 19 June 2020; Revised: 14 November 2020; Accepted: 23 November 2020

Abstract

The central thrust of this paper is to accentuate the impact of Anti-Retroviral Therapy (ART) on cure rate of HIV/AIDS patients and on the transition intensities between the stages of disease using cure rate model and Hidden Markov model (HMM) respectively. Hidden Markov Model (HMM) is a captivating algorithm for temporal pattern recognition like automated speech, handwriting and gesture recognition in the signal processing field. Although it is based on Markov processes which are more widely used in estimating the transition rates between the different stages of a disease, but HMM is hardly being used in survival data modeling.

Key words: AIDS; CD4; Cure rate model; Hidden Markov model.

1. Introduction

Human Immunodeficiency Virus (HIV) is a kind of virus that ushers and leads to Acquired Immune Deficiency Syndrome (AIDS). HIV taints a particular type of white blood cells, known as T- cells (or CD4+ T-cells), that helps in fighting diseases. As time passes, HIV kills CD4+ T- cells and multiplying itself, that leads to weakening of the immune system. In due course of time, the infected person's immune system can no longer fight off diseases. So, proper measurement of CD4+ T cell count may be viewed as the snapshot of how good a patient's immune system is functioning.

Till date, there is no vaccine that can claim of curing HIV/AIDS. Although, a medication called antiretroviral (ARV) drug can steady the deteriorating immune system. The initiation of ARV drug is generally based on two clinical observations, one is CD4+ T cell count and another is viral load (HIV RNA concentrations) that measures HIV in the blood, lower is better. The purpose of the ARV drug is to make viral load undetectable and if it is able to do so, then infected person can't transmit HIV to partner [Veterans' Health Administration]. According to WHO guidelines also, initiation of ARV drug and for measuring disease progression, viral load should be preferred over the CD4+ cell count.

But, in India due to scarcity of resources, the decision about the commencement of treatment and disease progression is taken merely based on CD+ cell count. In spite of the fact that, national AIDS control organization (NACO) issued new guidelines that mandated to "*treat all persons living with HIV (PLHIV) with antiretroviral therapy regardless of CD4+ cell*

Corresponding author: Arpan Kumar Thakur
Email: arpankmr3@gmail.com

count, clinical stage, age or population" [NACO on May, 2017], CD4+ cell count play an indispensable role in entire treatment protocol.

To study the transmission of the virus to next-generation Bature *et al.* (2010) used a Markov chain model. The same model has been used for observing disease progression in liver cancer Kay *et al*. (1986), for Hepatitis C disease progression Sweeting *et al.* (2010), for tuberculosis (TB) progression Debanne *et al*. (2000), Alzheimer's disease Commenge *et al*. (2004), liver-cirrhosis progression Grover *et al*. (2014). Discretized Markov model has been developed and employed to AIDS prediction in England and Wales, Aalen *et al*. (2018) used Markov model to study disease progression among HIV/AIDS patients, Grover *et al*. (2019).

New and ameliorated statistical methods are always entailed for making decisions about initiation and switching treatment protocols. Nevertheless, antecedent studies have appropriately modeled disease progression using multistate Markov processes, very few have explored the aptness of the hidden Markov model.

The aftermath of lung transplantation is studied by Jackson and Sharples (2002), Guihenneuc-Jouyaux *et al*. (2000) used a Bayesian hierarchical model for hidden Markov processes by exemplifying HIV infected patient's data. On the contrary to the simple Markov model, where the state is directly observable, in HMM the true state is not directly visible (that's what name hidden symbolizes). Laake *et al*. (2014) used Hidden Markov Model to study dependent mark loss and for estimation of survivals of black bears. Johnson *et al*. (2016) employed multivariate Hidden Markov Models to study mark-recapture data of California sea lion vital rates. Dempsey *et al*. (2017) used this model to study mobile health (mHealth) data collected from sensor streams and self report. Discrete survival time data were studied in Bayesian framework by Kozumi (2000).

The HMM canvasses to recuperate the true sequence of states from the visible (observed) sequence of states It has a plethora of applications in speech recognition, in part of speech tagging, in object tracking, in computational molecular biology. HMM in one sense may be treated as an artefact in the sense that it has developed way back in late 1960's by Baum and Petrie (1966) but it's use is now ubiquitous in science including survival analysis.

In India, ART centers are compelled to use CD4+ T cell count instead of the viral load while staging the HIV patients. This may lead to a mismatch in staging, additionally measurement of CD4+ cell count itself is prone to error mainly due to intraindividual variability and to some extent due to measurement error. In this paper an attempt has been made to underline the mismatch using HMM.

The paper is organized as follows: in next section 2, a short explanation of material and method to be used is given. In section 3, results are provided followed by section 4 where discussions, limitations, future ambits and pipelined research is presented.

2. Material and Methods

2.1. Materials

It is a longitudinal retrospective follow-up study of 5300 HIV/AIDS patients undergoing treatment at ART center of Dr. Ram Manohar Lohia hospital in New Delhi, during the period April 2004 to December 2014. Exclusion criteria were the age at enrollment should be >= 18

years, should have baseline CD4+ cell count available, periodic CD4+ cell count available for at least two visits. By filtering using complete case analysis on variables like sex, smoking and alcohol consumption status, treatment (*virocomb-N* combination and *others*), we are left with only 1063 observations.

2.2. Methods

2.2.1. Cure Fraction model

Assume that C be the probability of an HIV patient being a long-term survivor and $(1 - C)$ be the probability of a patient being susceptible to death (Stage 5 of the disease). Then, Berkson *et al.* (1952) defined the survival function at any time t as:

$$S(t) = C + (1 - C) * S_u(t) \tag{1}$$

where, $S_u(t)$ is the survival function of the susceptible population which may be assumed to follow some life time distribution. Probability density function $f(t)$ of the overall population is written as

$$f(t) = (1 - C) * f_u(t) \tag{2}$$

where $f_u(t)$ is the probability density function of susceptible population.

Now let (t_i, δ_i) be the observed data of size n , where t_i is the survival time of the i^{th} patient and δ_i is censoring indicator variable which is defined as follows: $\delta_i = 0$ for right-censored observation and $\delta_i = 1$ for uncensored observation (*i = 1, 2, . . , n*).

Accordingly, the individual patient's contribution to the likelihood function can be written as

$$\begin{aligned} L_i &= [f(t_i)]^{\delta i} \ [S(t_i)]^{(1-\delta i)} \\ &= [(1 - C)f_u(t_i)]^{\delta i} \ [C + (1 - C)S_u(t_i)]^{(1-\delta i)} \end{aligned} \tag{3}$$

So, complete likelihood is given by

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} [(1 - C)f_u(t_i)]^{\delta_i} \ [C + (1 - C)S_u(t_i)]^{(1-\delta_i)} \tag{4}$$

Parameters are estimated by maximizing the complete data likelihood in equation (4) using WinBUGS software package using Gibbs sampling approach. Here we have used various lifetime distributions like exponential, Weibull, gamma, exponentiated Weibull *etc.*, based on least deviance information criteria (DIC) value we found exponentiated Weibull distribution to be the best model. For detailed review of the foregoing model one may refers to Farewell (1982), Yamaguchi (1992), Maller and Zhou (1995), Chen *et al.* (1999), Peng and Dear (2000), and Sy and Taylor (2000), Kannan *et al*. (2010), Achcar *et al*. (2012), Varshney *et al*. (2018).

2.2.2. Hidden Markov model

Before applying HMM, we have used a time-homogenous multistate Markov model to study disease progression among HIV/AIDS patients. For this purpose, stages of HIV/AIDS patients have been defined in terms of CD4+ cell count as:

Stage/State	1	2	3	4	5
CD4+ cell count range	>500	351-500	200-350	<200	Death

It is well established that ARV drugs improve the CD4+ cell count in most of the cases, but unfortunately for some patients, it might not do so, that results in deterioration of health. That is, the patients may move from a lower stage to higher stages of the disease, a significant proportion of patients move to end-stage, *i.e.* death stage too. So, backward progression / transition is also a possibility. Consequently we used reversible transition model that is depicted in Figure 1. Except for stage 5, which is absorbing stage all other stages are transient in nature.

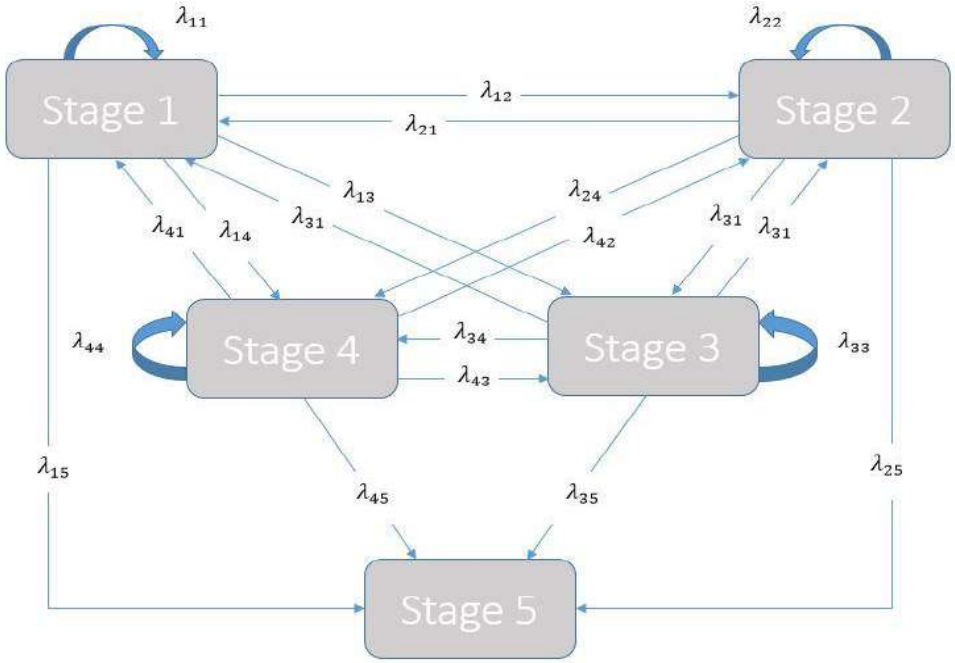


Figure 1: **Possible disease progression**

With the passage of time, a patient may move in possible state space *S={1,2,3,4,5}*. Let $X(t) = r$ be the current state of the patient, then the transition intensity λ_{rs} of advancing to state s in infinitesimal time δ_t is given by

$$\lambda_{rs} = \lim_{\delta_t \to 0} \frac{P(X(t + \delta_t) = s/X(t) = r)}{\delta_t}$$

Then the transition intensity matrix Q can be written as $Q = [\lambda_{rs}]_{r,s \in S}$ and possess the following two properties (a) $\sum_{s \in S} \lambda_{rs} = 0$ for all r and (b) $\lambda_{rs} = -\sum_{r \neq s} \lambda_{rs}$.

The maximum likelihood estimation technique developed by Kalbfleish and Lawless (1986) can be used to estimate the transition intensities, λ_{rs}. Estimated transition intensities in turn can be used to find the transition probability matrix $P(t) = [P_{rs}(t)]_{r,s \in S}$ and $P_{rs}(t)$ is defined as:

$$P_{rs}(t) = P(X(t + v) = s/X(t) = r)$$

Also, Cox and Miller (1965) defined transition probability matrix with the help of the intensity matrix as a Kolmogorov equation $P(t) = e^{tQ}$. Similarly, mean sojourn time, that is the time of stay in any transient state, is given by $-1/\lambda_{rr}$. Let us denote covariates vector as $\boldsymbol{Z}$, then the effect of covariates on transition intensity can be modeled by $q_{ij}(t)$, and defined in terms of Cox- proportional hazard regression as suggested by Marshall and Jones (1995):

$$q_{ij}(t) = q_{ij}(0)e^{z\beta_{ij}}$$

Here $q_{ij}(0)$, is the baseline intensity, β_{ij} is the coefficient of regression. Here it is assumed that covariates are time independent. Estimates can be obtained using the maximum likelihood procedure suggested by Kalbfleish and Lawless (1986).

A hidden Markov model is generally used for defining a probability distribution over a sequence of observations. For brief elucidation, consider the observation at time *t* by the variable X_{it}. It is presumed that *t* is an integer-valued index. Additionally, it is based on two assumptions: (i) the observations at time *t* is fostered by some process that is hidden from the observer and generated by misclassification matrix, (ii) it is also assumed that hidden state follows the Markov property with transition matrix *Q*, put in another way current state envelopes all information that is required to know about the historicity of the process to predict the subsequent future of the process, Ghahramani (2001), this intricate relationship for HMM is given in Figure 2. Generalized regressions can be used to model the covariates effect on transition intensity and misclassification probabilities.

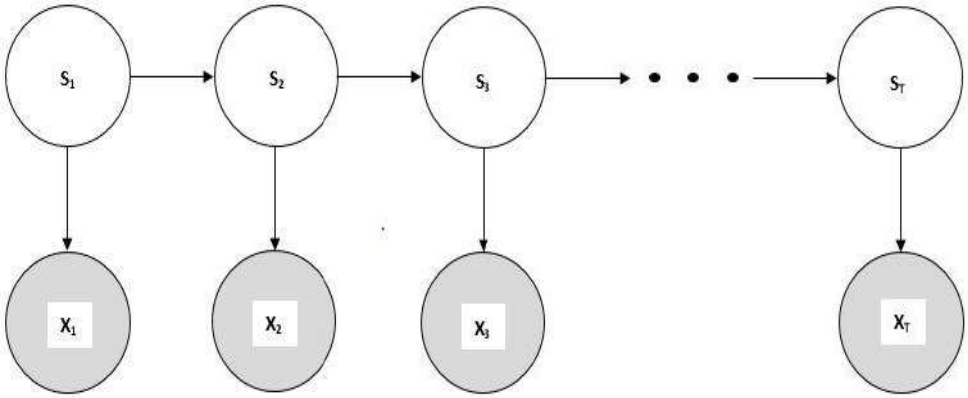


Figure 2: Hidden and observed stages

For mathematical formulation of the HMM, let $X_{iT} = [X_{i1}, \ldots, X_{iT_i}]$ denotes the observed state that triggered by the hidden state S_{it}. The observed states X_{it} are assumed to be conditionally independent of true hidden states. The likelihood contribution for patient *i* is given by

$$L_i = f(X_{i1}, \ldots, X_{iT_i})$$

$$= \sum_{S_i} f(X_{i1}, \ldots, X_{iT_i} / S_{i1}, \ldots, S_{iT_i}) f(S_{i1}, \ldots, S_{iT_i})$$

Given the values of the underlying hidden state, observed states are conditionally independent, using Markovian property of hidden states

$$P(S_{ij}/S_{i,j-1,\ldots,},S_{i1}) = P(S_{ij}/S_{i,j-1})$$

The resulting likelihood can be rewritten as,

$$\sum_{S_i}\prod_{t_i=t_1}^{T_i} f(X_{it_i}/S_{it_i})\left\{f(X_{i1})\prod_{t_i=t_2}^{T_i} f(X_{it_i}/S_{it_i-1})\right\}$$

In HMM, for the observable state X_t are conditionally emitted by hidden states S_t through misclassification matrix $M = [e_{rs}]_{r,s\in S}$, whose elements are defined by

$$ge_{rs} = P\{X_t = s/\ S_t = r\}, r, s \in S$$

An assumption about disease stages is that a stage can be misclassified only to the adjacent disease stage, it is reasonable to assume that misclassification due to random causes will give over/under estimation of the disease to immediate stage. By employing the Viterbi algorithm technique, we can recreate the optimal sequence in HMM using dynamic programming algorithm. It was disseminated by Viterbi (1967), but more elaborate elucidation was given by Bellman (1957).

3. Results and Discussions

The progression of disease stages in HIV/AIDS patients are given in Table 1. Diagonal entries in the table is the number of times a patient remains in the same stage. The number 19 signify that number of occasions where patient of stage 1 moves to stage 2. Likewise, there are 5,12, 22 and 35 number of cases of reaching end stage 5 from stage 1, stage 2, stage 3 and stage 4 respectively.

The estimated parameters of cure rate model have been presented in Table 2. Here stages are observed after one year of initiation of ARV drug. Following table shows that patients who are in stage 1 have 86% chance of being long-term survivors, and chances are shrinking with severity of the disease. Patients who are in stage 4 even after one year of treatment have comparatively less chance (only 58%) of being long-term survivors. This table also gives Monte Carlo (MC) standard error of the mean.

Table 3 presents the intensity of disease progression in the absence of prognostic factors. Patients of stage 3 are 1.82 times (0.841/0.462) more likely to move to less severe disease stage 1 than moving to severe stage 4. Similarly, the patients of stage 4 are 27.2 times (0.716/0.0263) more likely to move to stage 3 than moving to death stage 5. Confidence interval is calculated by simulating 1000 random vectors from the asymptotic multivariate normal distribution. From Table 4 it can be observed that on an average a patient elapsed 1.88 years in stage 1, and 0.517 years, 0.812 years, 0.769 years in stage 2, stage 3 and stage 4 respectively.

Table 5 presents the estimated transition intensities for misclassification model along with misclassification probabilities. Therein e_{rs}, r denotes true stage and s denotes observed stage. So, e_{12} signify that for true stage 1 misclassifying it to stage 2 has probability 0.106, in other words there is 10% chance that patient of stage 1 will be mistakenly treated as stage 2, similarly there is 0.06 probability of treating stage 2 patients as stage 3. Mean sojourn time for

misclassification model is given in Table 6. Even though prognostic factors effects have not been presented for simple Markov model, it is used for Hidden Markov model in Table 7. With sex (female) as reference, overestimation (e_{12}, e_{23}, e_{34}) of misclassification probability has odds ratio 1.46, 1.81 and 2.08 over male patients. Odds ratio for misclassification probability for age (>35) is 2.412, 1.477, 0.906 for over-estimation (e_{12}, e_{23}, e_{34}) with respect to age (<=35).

To decrypt the states that could have most pertinently generated the sequence of stages observed, we employ a Viterbi algorithm Table 8. The data set is divided into two parts namely training data and testing data. On training data which we have taken as around 80% of total 1043 data points (830 observations), we developed the model. On remaining 20% of the data (213 observations) that was kept for testing, the trained model is applied to check the model performance using Viterbi algorithms. Table 8 also presents the precision for each observed stage of disease. It is to be noted that at higher observed stages of disease precision is also higher *i.e.,* for patients having advanced stage of disease there are less chances of being misclassified.

4. Conclusion

The study shows that current ART treatment is successful and effective in making HIV/AIDS patients long-term survivors. Although, sticking to the treatment (adherence) is highly suggested but that isn't easy to comply. Sometimes antiretroviral drugs could cause such side effects that is severe enough to make patient stop taking them. Unfortunately, if a patient skips drugs the virus may start multiplying itself. This results in HIV to get resistant to drugs, the scenario relatively more prevalent in developing countries including India. That may be the reason of partially high morbidity and mortality due to HIV in India. This also showed by our *cure rate model* where stage 4 patients have less long-term survivors than the lower stages. We have demonstrated the alluring algorithm of pattern recognition, HMM in modeling the survival time data. This paper ventured to decipher the hidden Markov model in HIV/AIDS setup, where simple Markov model is effectively and predominantly being used to study disease progression. We obtained transition intensity for misclassification model and also the misclassification probabilities. Even though prognostic factor's effects were not considered in simple Markov model, it is contemplated whilst studying hidden Markov model. Notwithstanding the evidence that sex of the patient have no significant effect on the disease progression Jackson (2011), when it comes to misclassification of stages it do have effect on odds of misclassification probability. It can be observed that males have more odds of misclassification probability than the females (reference group) patients. In other words, males are more vulnerable to exaggeration of stages of disease than the females, it may be distantly attributable to the prejudices towards males with respect to debauchery in general and promiscuity in particular. This finding may be re-verified through large scale meta- analysis of HIV/AIDS data.

Patients with age more than 35 years at enrolment may be subject to overestimation of stages, which is partially understandable as older age is closely related with rapid progression of disease, Ghate *et al*. (2011), Touloumi *et al.* (1998). Thus our study solidify the point that person with relatively higher age with even higher CD4+ count should initiate ART. Likewise, smoking and alcohol consumption are associated with overestimation of stages of the disease.

Most significant and compelling finding is related with CD4+ count, whenever CD+ count is below 200 cells/μL, then odds of misclassification (overestimation) probability have

increased. We have to further study the subjectivity involved in this result. As we have filtered the data set, therefore out of 1063 patients, majority of patients (694) are those on whom *virocomb-N* treatment combination were administered and remaining were given Tenolam+ Efravinez-600 *etc*. Hence, we classify the treatment protocol as "*virocomb-N*" (reference group) and "*Tenolam+ Efravinez-600*" as target group. With *virocomb-N* in reference, the others treatment have more odds of misclassification (overestimation), *i.e.* if treatment combination administered is "*others*" then there is more chance that they will be misclassified to higher stages of the disease. At last, Viterbi algorithm is used to see the most probable sequence of disease progression stages that may have given rise to the stages that we perceive as observed stage. By employing the Viterbi algorithm, at one go we can get rid of glut of errors committed during staging of the disease.

Acknowledgements

We are immensely thankful to the esteemed reviewers and editor for their vital insights and suggestions.

References

Aalen, O. O., Farewell, V. T., De Angelis, D., Day, N. E. and Nöel Gill, O. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine*, **16(19)**, 2191-2210.

Achcar, J. A., Coelho-Barros, E. A. and Mazucheli, J. (2012). Cure fraction models using mixture and non-mixture models. *Tatra Mountains Mathematical Publications*, **51**(**1**), 1-9.

Bature, R. S., Obiniyi, A. A., Absalom, E. E. S. and Sule, O. O. (2010). Markov chain simulation of HIV/AIDS movement pattern. *International Journal of Computer Science and Information Security*, **8**(**2**), 156-167.

Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**(**6**), 1554-1563.

Bellmann, R. (1957). *Dynamic Programming*. Princeton University Press. Princeton, NJ.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(**259**), 501-515.

Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(**447**), 909-919.

Commenges, D., Joly, P., Letenneur, L. and Dartigues, J. F. (2004). Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*, **23**(**2**), 199-210.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. London: Chapman and Hall.

Debanne, S. M., Bielefeld, R. A., Cauthen, G. M., Daniel, T. M. and Rowland, D. Y. (2000). Multivariate Markovian modeling of tuberculosis: forecast for the United States. *Emerging Infectious Diseases*, **6**(**2**), 148.

Dempsey, W. H., Moreno, A., Scott, C. K., Dennis, M. L., Gustafson, D. H., Murphy, S. A. and Rehg, J. M. (2017). iSurvive: an interpretable, event-time prediction model for mHealth. *Proceedings of machine learning research*, **70**, 970.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 1041-1046.

Ghate, M., Deshpande, S., Tripathy, S., Godbole, S., Nene, M., Thakar, M., and Mehendale, S. (2011). Mortality in HIV infected individuals in Pune, India. *The Indian Journal of Medical Research*, **133**(**4**), 414.

Grover, G., Sabharwal, A., Kumar, S., and Thakur, A. K. (2019). A multi-state Markov model for the progression of chronic kidney disease. *Türkiye Klinikleri Biyoistatistik*, **11**(**1**), 1-14.

Grover, G., Sreenivas, V., Khanna, S., and Seth, D. (2014). Multi-state Markov model: An application to liver cirrhosis. *Statistics in Transition New Series*, **15**(**3**), 429-442.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, **15**(**01**), 9-42.

Guihenneuc-Jouyaux, C., Richardson, S., and Longini, I. M. (2000) Modelling markers of disease progression by a hidden Markov process: application to characterising CD4 cell decline. *Biometrics*, **56**, 733–741.

Jackson, C. H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, **38**(**8**), 1-29.

Jackson, C. H., and Sharples, L. D. (2002) Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, **21**, 113–128.

Johnson, D. S., Laake, J. L., Melin, S. R., and DeLong, R. L. (2016). Multivariate state hidden Markov models for mark-recapture data. *Statistical Science*, **31**, 233-244.

Kannan, N., Kundu, D., Nair, P., and Tripathi, R. C. (2010). The generalized exponential cure rate model with covariates. *Journal of Applied Statistics*, **37**(**10**), 1625-1636.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, **42**, 855–865.

Kozumi, H. (2000). Bayesian analysis of discrete survival data with a hidden Markov chain. *Biometrics*, **56**(**4**), 1002-1006.

Kuk, A. Y., and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 531-541.

Laake, J. L., Johnson, D. S., Diefenbach, D. R., and Ternent, M. A. (2014). Hidden Markov model for dependent mark loss and survival estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**(**4**), 522-538.

Lee, S., Ko, J., Tan, X., Patel, I., Balkrishnan, R., and Chang, J. (2014). Markov chain modelling analysis of HIV/AIDS progression: A race-based forecast in the United States. *Indian Journal of Pharmaceutical Sciences*, **76**(**2**), 107.

Maller, R. A., and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics*, **51**, 1197-1205.

Marshall, G., and Jones, R. H. (1995) Multi-state Markov models and diabetic retinopathy. *Statistics in Medicine*, **14**,1975–1983.

National AIDS Control Organization | MoHFW | GoI. (2018). *Naco.gov.in*. Retrieved 18 May 2020, from http://naco.gov.in/.

Peng, Y., and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**(**1**), 237-243.

Rabiner L. R, Fellow (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257-286.

Ross, J. M., Ying, R., Celum, C. L., Baeten, J. M., Thomas, K. K., Murnane, P. M., and Barnabas, R. V. (2018). Modeling HIV disease progression and transmission at population-level: The potential impact of modifying disease progression in HIV treatment programs. *Epidemics*, **23**, 34-41.

Sweeting, M. J., Farewell, V. T., and De Angelis, D. (2010). Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Statistics in Medicine*, **29**(**11**), 1161-1174.

Sy, J. P., and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56**(**1**), 227-236.

Touloumi, G., Hatzakis, A., Rosenberg, P. S., O'brien, T. R., and Goedert, J. J. (1998). Effects of age at seroconversion and baseline HIV RNA level on the loss of CD4+ cells among persons with hemophilia. *Aids*, **12**(**13**), 1691-1697.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**(**464**), 1063-1078.

Varshney, M. K., Grover, G., Ravi, V., and Thakur, A. K. (2018). Cure Fraction Model for the Estimation of Long-term Survivors of HIV/AIDS Patients under Antiretroviral Therapy. *Journal of Communicable Diseases*, **50**(**3**), 15-21.

Veterans Health Administration, a. (2018). HIV/AIDS Home. Hiv.va.gov. *Retrieved 2 May 2020, from https://www.hiv.va.gov/index.asp.*

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(**2**), 260-269.

WHO guidelines on the use of CD4, Viral load and EID test for initiation and monitoring of ART *WHO | World Health Organization*. (2018). *WHO International*, Retrieved 13 August 2019, from http://www.who.int/.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an applicationg to the analysis of "permanent employment" in Japan. *Journal of the American Statistical Association*, **87**(**418**), 284-292.

APPENDIX

Table 1: **Number of state transitions**

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Stage 1	130	19	7	1	5
Stage 2	131	128	65	7	12
Stage 3	75	251	314	64	22
Stage 4	28	133	484	363	35

Table 2: **Estimated cure rate model parameters**

		Mean	S.D.	MC- error
Stage 1	C	0.862	0.0587	0.05011
	α	4.85E-03	0.003741	6.57E-04
	β	0.06538	0.0995	0.00113
	γ	1.547	0.1095	0.0221
Stage 2	C	0.724	0.0418	0.00735
	α	5.74E-03	0.00411	2.51E-04
	β	0.00856	0.0997	0.001306
	γ	1.632	0.1014	0.0113
Stage 3	C	0.657	0.0156	0.00815
	α	6.85E-03	0.00412	5.27E-04
	β	0.006449	0.01317	0.001614
	γ	1.0546	0.2514	0.01822
Stage 4	C	0.587	0.0248	0.00139
	α	7.54E-03	0.00417	4.28E-04
	β	0.009324	0.0243	0.000908
	γ	0.693	0.168	0.099

Table 3: **Estimated transition intensities with 95% confidence interval**

From	To	Intensity	C.I.
Stage 1	Stage 1	–0.5306	(–0.759, –0.371)
Stage 1	Stage 2	0.303	(0.249, 0.730)
Stage 1	Stage 3	0.14	(0.023, 0.3621)
Stage 1	Stage 4	0.09	(0.01, 0.1625)
Stage 1	Stage 5	1.32E-06	(0, 8.035)
Stage 2	Stage 1	0.983	(0.734, 1.158)
Stage 2	Stage 2	–1.94	(–2.240, –1.371)
Stage 2	Stage 3	0.553	(0.335, 1.210)
Stage 2	Stage 4	0.405	(0.272, 1.116)
Stage 2	Stage 5	1.98E-05	(0, 2.920)
Stage 3	Stage 1	0.33	(0.234, 0.621)
Stage 3	Stage 2	0.841	(0.603, 1.331)
Stage 3	Stage 3	–1.64	(–1.837, –1.456)
Stage 3	Stage 4	0.462	(0.3604, 0.6001)
Stage 3	Stage 5	8.68E-03	(0.00067,0. 1115)
Stage 4	Stage 1	0.27	(0.13502, 0.3402)
Stage 4	Stage 2	0.7504	(0.613, 1.712)
Stage 4	Stage 3	0.716	(0.571, 1.966)
Stage 4	Stage 4	–1.76	(–1.966, –1.571)
Stage 4	Stage 5	2.63E-02	(0.0059, 0.118)

Table 4: **Mean sojourn times at different stages**

	Estimates (Std. error)	95 % C.I.
Stage 1	1.884 (0.343)	(1.318, 2.694)
Stage 2	0.517 (0.038)	(0.446, 0.598)
Stage 3	0.812 (0.036)	(0.544, 0.987)
Stage 4	0.769 (0.032)	(0.508, 0.963)

Table 5: **Estimated transition intensities and misclassification probabilities for misclassification model**

From	To	Intensity	Probability	
Stage 1	Stage 1	–0.517	e_{11}	0.894
Stage 1	Stage 2	0.233	e_{12}	0.106
Stage 1	Stage 3	0.15		
Stage 1	Stage 4	0.09		
Stage 1	Stage 5	0.046		
Stage 2	Stage 1	0.933	e_{21}	0.106
Stage 2	Stage 2	–1.845	e_{22}	0.834
Stage 2	Stage 3	0.514	e_{23}	0.06
Stage 2	Stage 4	0.382		
Stage 2	Stage 5	8.28E-03		
Stage 3	Stage 1	0.232		
Stage 3	Stage 2	0.625	e_{32}	0.152
Stage 3	Stage 3	–1.223	e_{33}	0.743
Stage 3	Stage 4	0.366	e_{34}	0.105
Stage 3	Stage 5	1.98E-05		
Stage 4	Stage 1	0.24		

Stage 4	Stage 2	0.783		
Stage 4	Stage 3	0.267	e_{43}	0.063
Stage 4	Stage 4	–1.29	e_{44}	0.937
Stage 4	Stage 5	1.27E-06		

Table 6: **Mean sojourn times for misclassification model**

	Estimates (Std. error)	**95 % C.I.**
Stage 1	1.934 (0.215)	(1.734,2.159)
Stage 2	0.542 (0.093)	(0.345.747)
Stage 3	0.817 (0.082)	(0.651,0.892)
Stage 4	0.775 (0.136)	(0.650,0.893)

Table 7: **Odds ratios for misclassification probabilities for prognostic factors**

	Misclassification					
	e_{12}	e_{21}	e_{23}	e_{32}	e_{34}	e_{43}
Sex	1.466	0.651	1.814	0.578	2.08	0.722
Age	2.412	0.855	1.477	0.881	0.906	0.763
Smoking	1.524	0.743	1.745	0.578	1.79	0.62
Alcohol	2.438	0.835	2.216	0.771	1.823	0.697
CD4 count	1.245	0.529	1.329	0.742	1.074	0.092
Treatment	1.586	0.784	1.157	0.635	5.428	0.083

Table 8: **Viterbi sequence**

		Generated by Viterbi Algorithm					Total	Precision
		Stage 1	Stage 2	Stage 3	Stage 4	Stage 5		
Observed	Stage 1	30	2	1	2	0	35	0.857143
	Stage 2	2	65	3	5	0	75	0.866667
	Stage 3	3	6	135	4	0	148	0.912162
	Stage 4	4	6	5	198	0	213	0.929577
	Stage 5	0	0	0	0	35	35	1

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 139-146

Joint Modeling of HIV and Tuberculosis through Copula-based Bivariate Binary Model

Atanu Pal[1] **and Arabinda Das**[2]
[1]*Department of Rheumatology and Nephrology, Institute of Post Graduate Medical Education and Research, Kolkata – 700 020, India*
[2]*Department of Statistics, Acharya Prafulla Chandra College, Kolkata – 700 131, India*

Received: 27 July 2020; Revised: 24 September 2020; Accepted: 08 December 2020

Abstract

Increasing spread of HIV facilitates the incidence of tuberculosis (TB) and HIV infected individuals co-infected with TB have much higher risk of developing active TB. In India, where TB infection is health burden, co-infection of TB with HIV increases the life-time risk by more than 5 times of developing TB. The objective of the study is to find the association between HIV and TB and the risk factors for infections of HIV and TB in Assam, north-east part of India. We used a joint bivariate binary model to accommodate the dependence between HIV and TB using a copula function. The maximum likelihood (ML) method has been used to estimate the model. We found a significant positive association between HIV and TB. The odds of developing TB in HIV infected person more than two-fold. We also found higher odds of HIV among younger people, who were men, military or paramilitary personnel, and heterosexual. These findings suggest that co-infection of HIV and TB is higher for this population and group interventions should be made to control the risk factors of the co-infection of HIV and TB in this part of India.

Key words: HIV; Tuberculosis; Bivariate binary outcomes; Copula function.

1. Introduction

While Human Immunodeficiency Virus (HIV) alone continues to be one of the deadliest diseases around the world, co-infection of tuberculosis (TB) with HIV is found to be the most leading cause of high mortality among people living with HIV (Corbett *et al.*, 2003; Lawn *et al.*, 2009). According to the UNAIDS report of 2019, about 37.9 million people globally were living with HIV at end of 2018 of which an estimated 10.0 million people developed TB disease, approximately 9% of all people living with HIV (UNAIDS fact sheet). Moreover, an estimated 49% of people living with HIV and TB are unaware of their co-infection and are therefore not receiving care. Increasing spread of HIV has become a major contributor in increasing the incidence of TB. Moreover, HIV infected individuals co-infected with TB have an annual risk of 5-15% of developing active TB, due to the reactivation of latent infection (Albalak *et al.*, 2007; Carvalho *et al.*,2001; Young, 2008; Mendelson, 2007). A HIV-positive person infected with TB has a 50 - 60% lifetime risk of developing TB as compared to an HIV-negative person who has only a 10% risk. Pulmonary tuberculosis showed no association with HIV and treatment failure. However co-infection of

Corresponding Author: Arabinda Das
Email: arabinda@apccollege.ac.in

pulmonary tuberculosis and HIV increases the probability of dying during treatment (Cabrera-Gaytán *et al*., 2016).

Thus, identification and treatment of TB as well as controlling the risk factors associated with TB are essential for successful HIV prevention in absence of effective vaccine. The objective of the present study was to find the association between HIV and TB in a survey among patients came to Assam Medical College and Hospital, Assam, north-east state of India and identify some of potential risk factors for HIV and TB infections. We used a joint response model with bivariate binary variables that measures the dependence between HIV and TB infections in assessing the socio-demographic, and biological risk factors of the two infections. Separate analysis of each infection ignores the dependence between the two infections and analyzing the data as if it were independent results in biased estimates.

The most common assumption for a bivariate binary mode to assume a bivariate normal distribution that guarantees normal marginals for the disturbances, and probit marginals for the binary dependent variables (Bhattacharya *et al*., 2006). Ghebremichael (2015) used a joint binary model to find the correlation between co-infection of HSV-2 and HIV-1. We consider a copula function based method to derive the joint model with the possibility of various marginal distributions and dependence structures for bivariate binary variables.

The copula approach of modelling gives wide flexibility in modelling for varied types of data. It has been widely used as a method to derive a joint bivariate distribution with non-normal marginal distributions and various dependence structures in recent years (see for details: Kolev and Paiva, 2009; Song, Li and Yuan, 2009 among others). Winkelmann (2012) proposed a bivariate probit model in modeling the effect of an endogenous binary regressor on a binary outcome variable with non-normal dependence among two variables using copula functions. There are number of copula families representing wide range of dependence structure among the random variables. A bivariate copula function, $C_\alpha(u_1, u_2)$ is a bivariate distribution function of the uniform [0,1] random variables and $\alpha \in \Omega$ measures the dependence between u_1 and u_2. Details of copula function can be found in Nelson (2006). The notion of statistical modeling by copula approach started with the pioneering work of Sklar (1959, 1973) where every multivariate distribution function can be uniquely constructed with a unique copula function which captures the dependence structure among the random variables and the marginal distribution functions of these random variables. Let $F(x_1, x_2; \lambda)$ be the joint distribution function of the random variables X_1 and X_2. Using Sklar's theorem, $F(x_1, x_2; \lambda)$ can be obtained with a unique copula function, $C_\alpha(u_1, u_2)$ such as $F(x_1, x_2; \lambda) = C_\alpha(F_1(x_1; \theta_1), F_2(x_2; \theta_2))$ where $F_i(x_i; \theta_i)$ is the distribution function of X_i, $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \alpha)'$.

The paper is organized as follows. We introduce a class of copula-based regression models for bivariate binary outcomes in Section 2 and discuss likelihood-based methods for model estimation. The model is estimated by maximum likelihood method. Section 3 introduces the data from the study population and summarizes the results from the proposed regression model applied to the data. Finally, Section 4, interprets the findings and concludes the paper.

2. Regression Model for Copula-based Bivariate Binary Outcomes

Consider two correlated binary outcomes X_i and Y_i obtained from each of *n* subjects, where X_i and Y_i are observed for the presence of HIV and TB respectively. The underlying

latent counterpart of X_i can be defined as $X_i^* = \boldsymbol{z}_{1i}^{'}\boldsymbol{\beta}_1 + \varepsilon_{1i}$ where $\boldsymbol{z}_{1i}$ denote explanatory variable, ε_{1i} is the random error and $\boldsymbol{\beta}_1$ is the corresponding parameter with

$$X_i = \begin{cases} 0 & if\ X_i^* \in (-\infty, \eta_1) \\ 1 & if\ X_i^* \in (\eta_1, \infty) \end{cases}$$

where η_1 is unknown threshold.

Similarly, the underlying latent counterpart of Y_i is presented as $Y_i^* = \boldsymbol{z}_{2i}^{'}\boldsymbol{\beta}_2 + \varepsilon_{2i}$ where $\boldsymbol{z}_{2i}$ denote explanatory variable, ε_{2i} is the random error and $\boldsymbol{\beta}_2$ is the corresponding parameter such that

$$Y_i = \begin{cases} 0 & if Y_i^* \in (-\infty, \gamma_1) \\ 1 & if Y_i^* \in (\gamma_1, \infty) \end{cases}$$

where γ_1 is unknown threshold.

Assuming $X_i^* \sim F_{X_i^*}$ and $Y_i^* \sim F_{Y_i^*}$ and the joint distribution function of (X_i^*, Y_i^*) as $F_{X_i^* Y_i^*}(x_i^*, y_i^*)$, the joint distribution of X_i and Y_i is then given by

$$\begin{aligned} &P(X_i = 0, Y_i = 0) = P(X_i^* \le \eta_1, Y_i^* \le \gamma_1) = F_{X_i^* Y_i^*}(\eta_1, \gamma_1) \\ &P(X_i = 0, Y_i = 1) = P(X_i^* \le \eta_1, \gamma_1 \le Y_i^*) = F_{X_i^*}(\eta_1) - F_{X_i^* Y_i^*}(\eta_1, \gamma_1) \\ &P(X_i = 1, Y_i = 0) = P(\eta_1 \le X_i^*, Y_i^* \le \gamma_1) \\ &= P(X_i^* \le \infty, Y_i^* \le \gamma_1) - P(X_i^* \le \eta_1, Y_i^* \le \gamma_1) \\ &= F_{Y_i^*}(\gamma_1) - F_{X_i^* Y_i^*}(\eta_1, \gamma_1) \end{aligned}$$

and

$$\begin{aligned} &P(X_i = 1, Y_i = 1) = P(\eta_1 \le X_i^*, \gamma_1 \le Y_i^*) \\ &= P(X_i^* \le \infty, Y_i^* \le \infty) - P(X_i^* \le \infty, Y_i^* \le \gamma_1) - P(X_i^* \le \eta_1, Y_i^* \le \infty) + P(X_i^* \le \eta_1, Y_i^* \le \gamma_1) \\ &= 1 - F_{Y_i^*}(\gamma_1) - F_{X_i^*}(\eta_1) - F_{X_i^* Y_i^*}(\eta_1, \gamma_1) \end{aligned}$$

The degree and type of dependence depends on the choice of copula. We consider the popular FGM copula, as this copula is comprehensive and allows for either negative or positive dependence. It also has a closed and simple analytic form. Moreover, the marginals are considered as standard logistic, i.e. logits for the binary dependent variables. The logit function is based on logistic distribution and its cdf turns out to be mathematically convenient. It is more popular in health sciences as the coefficients can be interpreted in terms of odds ratios. Therefore, the joint distribution function of X_i and Y_i with logit marginals and copula function is obtained as

$$F_{X_i^* Y_i^*}(x_i^*, y_i^*) = C_\alpha(\Lambda(x_i^*), \Lambda(y_i^*))$$

where $C_\alpha(.,.)$ is the copula function, $\Lambda(.)$ is the distribution function of a logistic random variable and is given by $\Lambda(z_i) = 1/(1 + \exp(-z_i))$.

Therefore, the joint density $f_{X_i, Y_i}(x, y) = P(X_i = x, Y_i = y)$ of X_i and Y_i under the copula is as follows:

$$P(X_i h = x, Y_i = y) = \begin{cases} C_\alpha\big(\Lambda(\zeta_1), \Lambda(\xi_1)\big) & if\ x = 0, y = 0 \\ C_\alpha(\Lambda(\zeta_1), 1) - C_\alpha\big(\Lambda(\zeta_1), \Lambda(\xi_1)\big) & if\ x = 0, y = 1 \\ C_\alpha\big(1, \Lambda(\xi_1)\big) - C_\alpha\big(\Lambda(\zeta_1), \Lambda(\xi_1)\big) & if\ x = 1, y = 0 \\ 1 - C_\alpha\big(1, \Lambda(\xi_1)\big) - C_\alpha(\Lambda(\zeta_1), 1) + C_\alpha\big(\Lambda(\zeta_1), \Lambda(\xi_1)\big) & if\ x = 1, y = 1 \end{cases}$$

where

$\zeta_1 = \eta_1 - \boldsymbol{z}_1^{'}\boldsymbol{\beta}_1$
$\xi_1 = \gamma_1 - \boldsymbol{z}_2^{'}\boldsymbol{\beta}_2$

Setting $\eta_1 = \gamma_1 = 0$ and under FGM copula, $C_\alpha(u_1, u_2) = u_1 u_2\{1 + \alpha(1 - u_1)(1 - u_2)\}$, the joint density $f_{X_i,Y_i}(x, y)$ becomes

$$P(X_i = 0, Y_i = 0) = \frac{1}{(1 + e^{-\zeta_1})}\frac{1}{(1 + e^{-\xi_1})}\left\{1 + \alpha\frac{1}{(1 + e^{\zeta_1})}\frac{1}{(1 + e^{\xi_1})}\right\}$$
$$P(X_i = 0, Y_i = 1) = \frac{1}{(1 + e^{-\zeta_1})}\left[1 - \frac{1}{(1 + e^{-\xi_1})}\left\{1 + \alpha\frac{1}{(1 + e^{\zeta_1})}\frac{1}{(1 + e^{\xi_1})}\right\}\right]$$
$$P(X_i = 1, Y_i = 0) = \frac{1}{(1 + e^{-\xi_1})}\left[1 - \frac{1}{(1 + e^{-\zeta_1})}\left\{1 + \alpha\frac{1}{(1 + e^{\zeta_1})}\frac{1}{(1 + e^{\xi_1})}\right\}\right]$$
$$P(X_i = 1, Y_i = 1) = 1 - \frac{1}{(1 + e^{-\zeta_1})} - \frac{1}{(1 + e^{-\xi_1})} + \frac{1}{(1 + e^{-\zeta_1})}\frac{1}{(1 + e^{-\xi_1})}\left\{1 + \alpha\frac{1}{(1 + e^{\zeta_1})}\frac{1}{(1 + e^{\xi_1})}\right\}$$

These joint probabilities described above depend on the selected copula as well as parameter vector $\boldsymbol{\lambda} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \alpha)'$. If the true copula is assumed to belong to a parametric family, consistent and asymptotically normally distributed estimates of the parameter $\boldsymbol{\lambda}$ can be obtained through maximum likelihood method.

Assuming an independent sample of *n* observations on $(x_i, y_i, \boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})$, the likelihood function is given by

$$L(\boldsymbol{\lambda}; x, y, \boldsymbol{z}_1, \boldsymbol{z}_2) = \prod_i\big\{P(X_i = 0, Y_i = 0|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})^{(1-x_i)(1-y_i)} \times P(X_i = 0, Y_i = 1|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})^{(1-x_i)y_i} \times P(X_i = 1, Y_i = 0|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})^{x_i(1-y_i)} \times P(X_i = 1, Y_i = 1|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})^{x_i y_i}\}$$

The corresponding log-likelihood function is then given by:

$$l(\boldsymbol{\lambda}; x, y, \boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{i=1}^{n}\{(1 - x_i)(1 - y_i)\log P(X_i = 0, Y_i = 0|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i}) + (1 - x_i)y_i \log P(X_i = 0, Y_i = 1|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i}) + x_i(1 - y_i)\log P(X_i = 1, Y_i = 0|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i}) + x_i y_i \log P(X_i = 1, Y_i = 1|\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})\}$$

The above presented log-likelihood function can be maximized using numerical optimization methods. These can use analytical first derivatives that have a relatively tractable form.

3. Data Analyses

This study was conducted at Department of TB & Chest Disease, Integrated Counselling and Testing Centre (ICTC) and Department of Medicine; Assam Medical College & Hospital, Dibrugarh from January 2007 to November 2008. Patients coming with suspected infection with HIV or Tuberculosis were sent to Department of TB and Chest Disease/RNTCP from ICTC. The study was approved by Assam Medical College & Hospital Ethics Committee and performed according to the Declaration of Helsinki, 1975. Written informed consent for participation in the study was obtained from participants. A total of 184 patients both male and female were screened for the study. Information was collected on different socio-demographic characteristics and sexual behaviors. Blood samples were drawn to test for HIV and Tuberculosis. HIV infection was determined using HIV enzyme-linked immunosorbent assay (ELISA), and reactive samples were confirmed using Wellcozyme HIV ELISA test. Western blot tests were used to confirm discordant ELISA test results. Presence of Tuberculosis were verified by acid fast bacilli (AFB) sputum smear, mycobacterial culture, histopathology, by clinical suspicion or radiological epidemiologic. Socio-demographic characteristics (age in years, gender, occupation), route of transmission, types of tuberculosis were considered as covariates in our analyses.

3.1. Results

One hundred and eighty four patients were recruited for the study, out of which 74% were male and 26% were female. The median age of the participants was 35 years (IQR = 25–54). Among the patients, 121 were HIV infected of which 35 were TB infected. The prevalence of TB among non-HIV infected patients was 11.11% compare to the prevalence of TB infection among HIV infected patients at 28.92% which is more than 2.6 times higher. The rate of HIV in TB positive was 83.33% compare to the rate of HIV among TB negative at 60.56%. The Pearson Chi-square test of interdependence between HIV and TB was highly significant indicating that patients who were infected by one of the infections were likely to be infected by the other as well (phi-coefficient = 0.201, p-value < 0.01). About 94% of the participants had heterosexual as route of transmission of HIV, and only 6% had homosexual as route of transmission of HIV. Eight-eight percent of HIV infected female were housewives, rest were commercial sexual workers. Among HIV infected male 74% were paramilitary & military personnel and private sector workers, 17% were drivers and rickshaw puller. Among TB infected patients 54% had pulmonary type of infection, and 46% had extra-pulmonary type of infection. Out of extra-pulmonary infection, lymphadenopathy were 26%, intestinal were 14% and meningitis were 6%. Table 1 showed male had more than twice odds (Odds Ratio (OR): 2.664, 95% Confidence Interval (CI): 0.845, 8.401) to have co-infection of HIV and TB. Urban patients had higher in-confection of HIV and TB (OR: 1.733, 95% CI: 0.672, 4.471) than their rural counterpart. Smokers showed almost 4 times odds (OR: 3.704, 95% CI: 1.513, 9.068) of developing co-infection. Drivers and paramilitary personnel showed seven times (OR: 7.200, 95% CI: 2.248, 23.063) and three times (OR: 2.965, 95% CI (0.964, 9.119) odds in favour of co-infection.

Table 1: Possible risk factors for tuberculosis (TB) in HIV and TB co-infected patients

		HIV + TB	HIV	OR (95% CI)
Gender				
	Female	4	22	1
	Male	31	64	2.664 (0.845, 8.401)
Age				
	<20	1	2	1
	20-40	29	57	0.983 (0.085, 11.295)
	40-60	4	24	3.000 (0.217, 41.353)
	>60	1	3	1.500 (0.055, 40.635)
Marital status				
	Married	11	25	1
	Single	24	61	1.118 (0.477, 2.622)
Residence				
	Rural	7	26	1
	Urban	28	60	1.733 (0.672, 4.471)
Smoking				
	No	8	45	1
	Yes	27	41	3.704 (1.513, 9.068)
Drug dependency				
	No	13	37	1
	Yes	22	49	1.278 (0.569, 2.865)
Occupation				
	Private	5	36	1
	Para-Military	14	34	2.965 (0.964, 9.119)
	Driver	16	16	7.200 (2.248, 23.063)
ROT				
Homosexual		1	10	1
Heterosexual		34	76	4.474 (0.551, 36.353)

Table 2: Outcome of Bivariate Binary Model: HIV and TB under FGM copula

Parameter	OR	SE	95% CI
HIV			
Age	1.372	0.135	(1.091, 2.114)
Gender	2.873	0.253	(1.138, 4.223)
ROT	5.611	1.574	(2.893, 26.432)
Occupation	4.627	1.282	(1.482, 7.324)
Residence	1.825	0.831	(1.121, 3.473)
TB			
Age	1.245	0.143	(0.913, 1.865
Smoking	3.122	0.986	(1.628, 7.421)
Residence	1.386	0.668	(0.739, 3.116)
Occupation	3.192	1.154	(1.433, 6.310)
Gender	2.521	0.431	(0.957, 3.264)
Correlation (α)	2.084	0.382	(0.891, 4.884)

The risk factors of HIV and tuberculosis were identified using the preliminary analysis. The variables that were found to be significantly associated with HIV and TB in the preliminary analysis were further considered for multivariate analysis. Age, gender, route of transmission (ROT), occupation were associated with both infections. We did a multivariate analysis using these variables as covariates in a joint response model. The results of our multivariate analysis are reported in Table 2. Odds ratios together with their corresponding standard errors and 95% confidence intervals are presented. Despite the relatively small sample, we found a significant association between HIV and TB. The results indicate that higher prevalence of HIV was associated among younger age patients. Male had significantly higher prevalence of HIV and TB than female. The odds of having HIV in younger people were 1.372 times higher than that of older people (OR = 1.372; 95% CI: 1.091–2.114). Men were associated with higher rates of HIV (OR = 2.873; 95% CI: 1.138–4.223) and TB (OR = 2.521; 95% CI: 0.957–3.264) infections. Heterosexual route of transmission was highly associated with higher prevalence of HIV (OR = 5.611; 95% CI: 2.893–26.432). The odds of having HIV and TB in urban patients were 1.825 times (OR = 1.825, 95% CI: 1.121-3.473) and 1.186 times (OR = 1.186, 95% CI: 0.439-3.116) higher. HIV was significantly associated with TB (OR = 2.084; CI: 0.891 – 4.884).

4. Conclusion

In this paper, we developed a joint response model for HIV and TB infections. Our study aimed to assess the association of HIV and TB infections with their determinants among peoples in Assam, north-eastern part of India. Socio-demographic characteristics such as age, gender and occupation as well as biological risk factors such as route of transmission of HIV and TB were considered. One hundred and eighty-four patients coming with suspected infection with HIV or TB were included in the study at Department of TB and Chest Disease/RNTCP, Assam Medical College and Hospital.

Assam has HIV prevalence of about 0.06%, compared to India's prevalence of 0.22% and TB prevalence of about 0.001%. The prevalence rates of HIV and TB among the study participants were 65.8% and 22.8%, respectively. The rate of TB in HIV infected patients was 2.6 times higher than rate of TB in non-HIV infected patients. Moreover, the rate of HIV in TB positive was 1.380 times higher the rate of HIV among TB negative. These indicate that the chance of infected from a disease increases for patients who have already infected from other disease. Thus, a joint response model is considered to accommodate the interdependence between the two infections of HIV and TB with the potential risk factors.
We found high prevalence of HIV among younger patients. The odds of having HIV in younger patients were 1.372 times higher than in older patients (OR = 1.372; 95% CI: 1.091–2.114). This finding may be result of demographical and social structure of India where most of the people are from young generation and they are more exposed to the western culture than their older counterpart. Men were associated with higher rates of HIV and tuberculosis infections. These findings are also associated with the social structure of India where women do not get much freedom and have restricted social life. Heterosexual route of transmission was highly associated with higher prevalence of HIV. Smoking played a pivotal role in developing TB. Study showed people from urban area were more prone to develop the co-infection of HIV and TB. The study also found that HIV was significantly associated with TB. Previous studies of HIV and TB co-infection modelled each infection separately ignoring the potential biological association between the two infections.

Acknowledgements

The authors express their gratefulness to the reviewers and the editor for their invaluable comments and suggestions.

References

Albalak, R., O'Brein, R. J., Kamemerer, S., *et al*. (2007). Trends in tuberculosis/Human Immunodeficiency Virus comorbidity, United States, 1993-2004. *Archives of Internal Medicine*, **167**(**22**), 2443-2452.

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in Medicine*, **25**(**3**), 389–413.

Cabrera-GaytaÂn, D. A., Niebla-Fuentes, Md. R., Padilla-VelaÂzquez, R., Valle-Alvarado, G., Arriaga-Nieto, L., Rojas-Mendoza, T., *et al*. (2016). Association of pulmonary tuberculosis and HIV in the Mexican Institute of Social Security, 2006-2014. *PLoS One*, **11**(**12**):e0168559. doi:10.1371/journal.pone.0168559.

Carvalho, A. C. C., DeRiemer, K., Nunes, Z. B., *et al*. (2001). Transmission of mycobacterium tuberculosis to contacts of HIV-infected tuberculosis patients. *American Journal of Respiratory and Critical Care Medicine*, **164**(**12**), 2166-2171.

Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C., *et al*. (2003). The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Archives of Internal Medecine*, **163**(**9**), 1009-1021.

Ghebremichael, M. (2015). Joint modeling of correlated binary outcomes: HIV-1 and HSV-2 co-infection. *Journal of Applied Statistics*, **42**(**10**), 2180-2191.

Kolev, N., and Paiva, D. (2009). Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, **139**(**11**), 3847-3856.

Lawn, S., and Churchyard, G. (2009). Epidemiology of HIV associated tuberculosis. *Current Opinion in HIV and AIDS*, **4**(**4**), 325-333.

Mendelson, M. (2007). Diagnosing tuberculosis in HIV-infected patients: challenges and future prospects. *British Medical Bulletin*, **81-82**, 149-165.

Nelson, R. B. (2006). *An Introduction to Copulas*. Springer, Berlin (ISBN 978-0-387-28678-5).

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, **8**, 229–231.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetika*, **9**, 449-460.

Song, P. X. K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copula. *Biometrics*, **65**, 60-68.

Winkelmann, R. (2012). Copula bivariate probit models: with an application to medical expenditures. *Health Economics*, **21**(**12**), 1444-1455.

Young, D. B., Perkins, M. D., Duncan, K., *et al*. (2008). Confronting the scientific obstacles to global control of tuberculosis. *The Journal of Clinical Investigation*, **118**(**4**), 1255-1265.

Statistics and Applications {ISSN 2454-7395(online)}
Volume 19, No. 2, 2021 (New Series), pp 147-159

Modeling of Mobile Telephone Subscribers Using Piecewise Nonlinear Growth Models

P.K. Das[1] **and Anita Pal**[2]
[1]*Department of Quantitative Techniques, Indian Institute of Foreign Trade, Kolkata-700107, India*
[2]*Department of Mathematics, National Institute of Technology, Durgapur-713209, India*

Received: 04 April 2020; Revised: 04 December 2020; Accepted: 09 December 2020

Abstract

The growth trajectory of the mobile telephone subscribers is highly nonlinear. The piecewise nonlinear growth model, comprising the well-known Gompertz and Bass models, was shown to be adequately describing the underlying data generating process of mobile telephone subscribers. This study used monthly time series data from March 1997 to December 2018 of telecom Circle A, representing the industrially advanced states like Maharashtra, Gujarat, Andhra Pradesh, Karnataka, and Tamil Nadu, was applied to develop the model. We partition the monthly data into the analysis sample (March 1997 to December 2017) and the test sample (January 2018 to December 2018). The parameters of the piecewise nonlinear model were estimated using Levenberg-Marquardt and sequential quadratic programming algorithms. The piecewise nonlinear model comprising Gompertz and Bass growth models was suitable for describing the monthly mobile subscribers' data in Circle A. The developed model was statistically validated using an appropriate *coefficient of determination* for the nonlinear models and *Root Mean Squared Error* (*RMSE*). We found the *RMSE* to be comparable for both training and the test sets. The forecasting capabilities of the piecewise nonlinear model, under mild violation of residual diagnostics, are compared to exponential smoothing (Holt's) and Gompertz models. We compared the performance of the model to the best fit Gompertz and Holt's models. In the test sample, we found the *RMSE* to be lower in the piecewise nonlinear model comprising Gompertz and Bass compared to the Holt's as well as the Gompertz model. We computed the forecast values of the subscribers during April-December 2020 using the developed model. As evident from the test sample and the published data of the Telecom Regulatory Authority of India (TRAI), the prediction from the developed model is lower than the actual values. The maximum potential number of subscribers in Circle A was 421.545 million, likely to be achieved in 2027. However, as the model predicted values are marginally smaller than the actual values, the maximum potential is expected to be completed before 2027.

*Key words***:** Piecewise nonlinear regression model; Gompertz model; Bass growth model; Sequential quadratic programming algorithm; Holt's model; Root mean squared error.

Corresponding Author: P.K. Das
Email: pkdas@iift.edu
Disclaimer: The views expressed in this paper are solely those of the authors and do not represent our employers' opinions.

1. Introduction

Indian telecommunication market is the second-largest in the world. The telecom sector showed remarkable growth and contributed substantially to creating new jobs and providing revenue to the Government. The industry will contribute ₹14.5 lakh crore to the economy and support 3 million direct jobs and 2 million indirect jobs by 2020 (GSMA, 2019). In 2015, the telecom sector accounted for 6.5% of India's Gross Domestic Product (GDP). With the leveraging of 5G technology, in 2020, it is estimated to reach 8.2% of India's GDP (ET, 2019, Wikipedia, 2020), if not getting delayed due to the COVID-19 pandemic. The bidding decision of the 5G spectrum and the telecom service providers' price in different circles would be based on the demand in that circle, among other parameters. All the stakeholders would be interested to know the future potential of the telecom markets in terms of subscriber base in the years to come. We can estimate the demand of a circle by predicting the number of mobile subscribers. There are four telecom circles in India, namely, metro, A, B, and C. In this study, we develop a model to forecast the total mobile subscribers of Circle A, which is comprising of industrially advanced states *viz*., Maharashtra, Gujarat, Andhra Pradesh, Karnataka, and Tamil Nadu. The number of subscribers (henceforth by subscribers, we would refer to customers subscribed to GSM services only) attained from 9698 in March 1997 to 3505.27 lakhs in December 2018 in Circle A – an impressive growth in subscriber base (COAI, 2018, TRAI, 2018).

In this study, our objectives are: (i) to develop an appropriate model to forecast the numbers of mobile subscribers in Circle A; (ii) to apply the developed model to forecast the numbers of mobile subscribers in Circle A. The modeling approach employed is that of piecewise nonlinear growth models. The results so obtained are compared to the exponential smoothing (Holt's) and Gompertz models. We compared the piecewise model's performance to the Holt's and Gompertz models using the test set. We divide the paper into five sections. Section 2 reviews the existing literature on piecewise nonlinear regression and exponential smoothing models and their applications. We discussed models, parameter estimation, fit statistics, and model selection metrics in Section 3. We discussed the results and salient findings in Section 4. Finally, in Section 5, we present the conclusion and the way forward.

2. Literature Review

The piecewise regression models are also known as segmented regression or broken stick regression models. In Econometrics, it is known as interrupted time series regression (Linden and Arbor, 2015). In this method, we partition the outcome (or the study) variable into regions, and a separate model is fit to each part. The piecewise regression is employed when the data is hypothesized to have been generated by more than one model (McZgee and Carleton, 1970). The piecewise linear regression models are pragmatic in the bio-physical (Vieth, 1989, Malash and El-Khaiary, 2010) and socio-economic domains (Birgit, 2006). The piecewise nonlinear regression models are also applied to biological and socio-economic studies (Oh and Kim, 2002, Maceina, 2007, Morrell *et al.*, 1995, Vanli and Kozat, 2014).

The application of the growth models is the most popular approach to study the growth trajectory of mobile subscribers in different markets, namely, Central and Eastern Europe (Gruber, 2001), Asia Pacific region (Wenrong *et al*., 2006), Greece (Michalakelis *et al*., 2008) and Taiwan (Wu and Chu, 2010). Several researchers (Sridhar, 2010; Hedau and Soni, 2016) also studied India's mobile telephone market. The effect of different techno-economic variables, among other things, on the growth of mobile services in different regions in India using panel data, was studied by Sridhar (2010). It was found that competition and network

were the two crucial variables to impact mobile services' growth. Though it provided significant insight into the impact of variables on mobile services' growth, it failed to make a time series forecast. The major lacuna in Hedau and Soni (2016) were in the model development and parameter estimation. The linearized model parameters were estimated; however, the original model was recommended as a forecast model. We can overcome this shortcoming by employing nonlinear estimation procedures. Though several studies were conducted to track the growth trajectory with limited success, no attempt has been made to engage piecewise nonlinear growth models to understand mobile subscribers' growth path. As growth models are mechanistic, they have advantages in understanding the data generating process and its future potential. We discuss the nonlinear growth models, piecewise nonlinear growth models, and related issues in the next section.

3. Models and Methodology

3.1. Nonlinear growth models

Let $n(t)$ and $N(t)$ denote the number of subscribers and the cumulative number of subscribers, respectively, to mobile service at time t in Circle A. If t_0 denotes the time at starting, *i.e.*, $t = 0$, then at time t, the cumulative number of subscribers can be expressed as:

$$N(t) = \int_{t_0}^{t} n(t)\,dt$$

where $n(t)$ is the non-cumulative number of adopters at time t. Further, let,

$$\frac{dN(t)}{dt} = \text{rate of growth at time } t, \text{ and}$$

$$\frac{1}{N(t)}\frac{dN(t)}{dt} = \text{relative rate of growth at time } t.$$

Let K denote the total number of potential subscribers in Circle A. It is also known as the carrying capacity or maximum potential of the system, *i.e.*, the markets of Circle A. Let us assume that:

(i) the rate of growth is proportional to the interaction of adopters and non-adopters. This can be expressed as:

$$\frac{dN(t)}{dt} \propto N(t)\left[1 - \frac{N(t)}{K}\right]. \tag{1}$$

In differential equation form, it becomes:

$$\frac{dN(t)}{dt} = rN(t)[\,1 - N(t)/K\,]. \tag{2}$$

Here, r is the intrinsic growth rate. The solution to this equation yields the following model:

$$N(t) = \frac{K}{1 + \frac{(K - N_0)}{N_0}\,exp(-rt)}, \tag{3}$$

where N_0 is the number of subscribers at $t = 0$.

Reparametrizing, the model can be written as:

where $B = (K - N_0)/N_0$.

$$N(t) = \frac{K}{1 + B * exp(-rt)}, \tag{4}$$

This model is known as the logistic growth model. The inflection point of this model is at $K/2$. The model described in Eq. (3) or the reparametrized model in Eq. (4) appeared deterministically as if data never deviates from the model. It is unrealistic. To make the model realistic, independently, identically, and normally distributed error term is added to the right-hand side (RHS) of the mathematical model. The resulting nonlinear regression model is:

$$N(t) = \frac{K}{1 + B * exp(-rt)} + \varepsilon_t \tag{5}$$

(ii) The relative rate of growth is proportional to the logarithm of the ratio of carrying capacity to subscribers' number at time t. Hence, this can be represented as:

$$\frac{1}{N(t)} \frac{dN(t)}{dt} \propto \ln\left(\frac{K}{N(t)}\right). \tag{6}$$

Therefore, the model in differential equation form can be expressed as:

$$\frac{dN(t)}{dt} = rN(t) \ln\left(\frac{K}{N(t)}\right) \tag{7}$$

with the boundary condition $N(t = t_0) = N_0$ = cumulative number of adopters at time t_0.

The parameter r is known as the intrinsic rate of growth. The solution to the above differential equation results in the following model:

$$N(t) = K * exp(-B * exp(-rt)) \tag{8}$$

where $B = ln(N_0 / K)$.

This model is known as the Gompertz model. By adding an error term to the RHS, we obtain the following statistical model:

$$N(t) = K * exp(-B * exp(-rt)) + \varepsilon_t . \tag{9}$$

The model is asymmetric, and the point of inflection is at K/e.

(iii) the rate of growth is proportional to the number of non-adopters, which can be expressed as:

$$\frac{dN(t)}{dt} \propto [K - N(t)]. \tag{10}$$

In the differential equation form, it becomes:

$$\frac{dN(t)}{dt} = r[K - N(t)]. \tag{11}$$

The solution to this equation yields the following model:

$$N(t) = K - (K - B) * exp(-rt), \tag{12}$$

where $r > 0$ and $K > B > 0$. Here, K is the maximum potential, and B is the number of subscribers at $t = 0$. The statistical model can be written as:

$$N(t) = K - (K - B) * exp(-rt) + \varepsilon_t \tag{13}$$

The model in Equation 13 is known as monomolecular (MM). The models in Equations 5, 9, and 13 are S-shaped growth models. These are nonlinear models in the statistical regression sense because at least one parameter of these models appears nonlinearly. The three parameters *viz*., *K*, *B*, and *r* of the models in Equations 5, 9, and 13, are estimated. For further details on S-shaped nonlinear growth models and the Richards model, readers can refer to Seber and Wild (2003).

(iv) The rate of growth is influenced by two types of subscribers, namely, innovators and imitators. We present the differential equation below:

$$\frac{dN(t)}{dt} = p\,[\,m - N(t)\,] + \left(\frac{q}{m}\right) N(t)[\,m - N(t)\,] \tag{14}$$

Here, *m* is the market potential, *p* and *q* are the coefficients of innovation and imitation. The solution of the above differential equation results in the following model:

$$N(t) = m\frac{1 - exp^{-(p+q)t}}{1 + \left(\frac{q}{p}\right) * exp^{-(p+q)t}} \tag{15}$$

By adding an error term to the RHS, we obtain the following statistical model:

$$N(t) = m\frac{1 - exp^{-(p+q)t}}{1 + \left(\frac{q}{p}\right) * exp^{-(p+q)t}} + \varepsilon_t \tag{16}$$

In this model, $m > 0$, $p > 0$, and $q > 0$. It is also a nonlinear model in the regression sense. The parameters of the model *viz*., *m*, *p*, and *q* are to be estimated. For further details on the Bass model, the readers can refer to Bass (1969) and Rogers (2003).

3.2. Piecewise nonlinear growth models

The piecewise nonlinear model of the following type is considered in this study:

$$(t < T^*) * N1(t) + (t >= T^*) * N2(t). \tag{17}$$

In Equation (17), T^* is the value of *t* at which the growth trajectory is found to be changing from one model to another model. It is also known as the knot, breaking point, change point, or joining point. The two nonlinear functions are denoted by $N1(t)$ and $N2(t)$. If $(t < T^*)$ is true, it returns one else zero. Similarly, if $(t >= T^*)$ is true, it returns 1 else zero. Here, $N1(t)$ can be any growth model *viz*., logistic, Gompertz, monomolecular, Bass, Richards, or any other growth model. The function, $N2(t)$, can also be any growth model *viz*., logistic, Gompertz, monomolecular, Bass, Richards, or any other growth model. As an example, let us consider the following combination:

$$N(t) = (t < T^*) * (K * exp(-B * exp^{(-rt)})) \tag{18}$$

$$+ (t \geq T^*) * m \frac{1 - exp^{-(p+q)t}}{1 + \left(\frac{q}{p}\right) * exp^{-(p+q)t}} + \varepsilon_t$$

In this segmented (piecewise) study, the first segment of the sample data is hypothesized to have been generated by a Gompertz model. The second segment of the sample data is hypothesized to have been caused by a Bass model. In general, when we consider '*n*' number of models for modeling a data set having two segments, there can be n^2 number of piecewise models.

3.3. Estimation of parameters

The nonlinear model differs in their estimation properties from linear regression models. Under the assumption of an independently and identically distributed normal error term, the linear model gives rise to unbiased, normally distributed minimum variance estimators. Nonlinear regression models tend to do so as the sample size becomes very large (asymptotically) (Ratkowsky, 1990, Ross, 1990, Intriligator, 1996). Like linear regression, in nonlinear regression also normal equations are obtained. However, these normal equations are nonlinear, and no explicit solutions can be obtained. Different algorithms are available in the literature to solve nonlinear normal equations. Three main algorithms are (i) Gauss-Newton, (ii) Sequential Quadratic Programming, and (iii) Levenberg-Marquardt. The sequential quadratic programming algorithm is appropriate for nonlinear constraint models. It is a combination of Lagrangian relaxation, active set strategy, and Newton-Raphson methods. The algorithm yields stable solutions in the majority of situations. The details of these methods and their merits and demerits are available in the literature (Draper and Smith,1998; Nocidal and Wright, 2006). These algorithms are iterative and require starting values of the parameters. A good starting value can ensure global convergence and can obtain a minimum value of the loss function. The sum of squared residuals can be considered a loss function in estimating piecewise nonlinear regression models' parameters. The choice of good starting values can influence the convergence of the algorithm in locating the fitted value or between rapid and slow convergence to the solution. However, there is no standard procedure for computing the starting values of the parameters. Sometimes a combination of two or three methods results in good starting values. In this study, a combination of techniques is used to obtain the starting values of the parameters. IBM SPSS Statistics version 26 software package is used to estimate the models' parameters and computation of goodness of fit measures (IBM, 2019). The goodness of fit of the nonlinear model is assessed by the *coefficient of determination* (R^2).

However, as Kvalseth (1985) pointed out, eight different expressions for R^2 appear in the literature. One of the most frequent mistakes occurs when the fits of a linear and a nonlinear model are compared by using the same R^2 expression. Thus, a logistic or a Gompertz model may first be linearized by using a logarithmic transformation and then fitted to data by using the ordinary least squares method. The R^2-value is then often calculated using the log of observed and log of predicted data points. The R^2 is, erroneously, interpreted as a measure of goodness of fit of even the original nonlinear model. Scott and Wild (1991) have given an example where two models are identical for all practical purposes and yet have very different values of R^2 calculated on the transformed scales. Kvalseth (1985) has emphasized that the following R^2

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{19}$$

where *RSS* is the residual sum of squares, and *TSS* is the total sum of squares, which is entirely appropriate even for nonlinear models. We present below the other necessary summary measures for nonlinear models:

$$Root\ Mean\ Squared\ Error\ (RMSE) = \sqrt{\frac{\sum_{t=1}^{T}(N(t) - \widehat{N(t)})^2}{T}}\ . \qquad (20)$$

Here, *T* is the total number of observed values. $N(t)$, and $\widehat{N(t)}$ are the number of actual and the predicted subscribers.

Mean Absolute Percentage Error (*MAPE*) is defined as:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{N(t)-\widehat{N(t)}}{N(t)}\right| * 100 \qquad (21)$$

$N(t)$, and $\widehat{N(t)}$ are the number of actual and the predicted subscribers, respectively.

4. Results

Monthly data on mobile subscribers in the Circle A, from March 1997 to December 2018, was collated from the Cellular Operators Association of India (COAI) (www.coai.in). Currently, in the COAI repository, monthly data is available from January 2005 - December 2018. The data before 2005 was collated from the same repository in 2013 when it was available in the database. Data is currently available on the TRAI website; however, there is a difference in the subscribers' numbers reported by COAI and TRAI. For example, in January 2018, COAI reported 340.41 million subscribers, whereas TRAI reported 402.81 million subscribers in Circle A. To avoid mixing data from two sources, only data from COAI, which is available until December 2018, has been analyzed. Moreover, cumulative data is required for the estimation of parameters of the growth models. The cumulative data is essentially increasing or equal to the previous observation where $N(t+1) >= N(t)$. However, in the reported data in some months, this essential criterion has been violated. Wherever $N(t+1)$ was reported to be smaller than the $N(t)$, it has been imputed by the $N(t)$. With this simple and essential imputation, the monthly data were pre-processed to estimate the model's parameters. Further, the complete data set was partitioned into an analysis sample and the test sample. The monthly data from March 1997 to December 2017 were used as the analysis sample, and January 2018 to December 2018 was retained as the test sample. The line plot of the data is presented in Figure 1. The line plots of Figure 1 and Figure 2 are generated in R studio (R Core Team, 2016, RStudio Team, 2015). Figure 1 depicts the growth trajectory of the corrected level data. The growth trajectory appears to be S-shaped until 2016 and follows a different path (Figure 1).

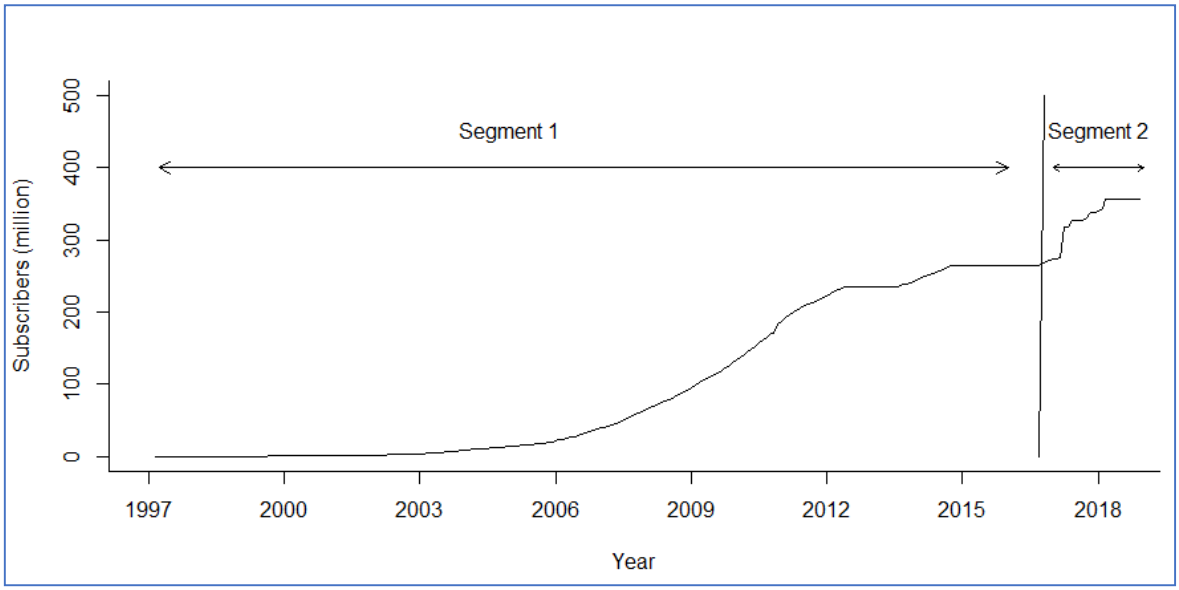


Figure 1: Partitioning of the data into Segment 1 and Segment 2 by the vertical line

It is hypothesized that the underlying data generating process follows an S-shaped model. The traditional growth models, namely, logistic, Gompertz, monomolecular, Bass, and Richards, were fitted to the data using a nonlinear estimation method. The results are presented in Table 1.

Table 1: Results of fitting nonlinear growth models to subscribers' data of Circle A

Fit Statistics/Model	Logistic	Gompertz	MM	Bass
R^2	0.98	0.99	0.952	0.98
RMSE	15.89	13.85	26.64	15.84
MAPE	1.38	0.36	49.94	0.55
K (millions)	318.79	359.29	IE	
m (millions)				319.74
IE: Inadmissible estimate				

The Gompertz model appears superior to other models in R^2, *RMSE*, and *MAPE*. The next appropriate model is the Bass model. The Richards model resulted in the non-convergence of the iterative algorithm. The maximum potential of the market is estimated using the parameter *K*. The actual maximum number of subscribers in the sample is 357.378 million. The monomolecular model resulted in an inadmissible estimate. The logistic and Bass models resulted in an estimate lower than the actual maximum value. The estimate of *K* given by the Gompertz model is only marginally higher than the actual maximum value. The Gompertz model is appropriate for describing the subscribers' data of Circle A based on the fit statistics. The estimated parameter value of *K* in the Gompertz model is admissible. We found that the results did not support the assumption of the normality of residuals in the Gompertz model using the Anderson-Darling test.

Moreover, the *RMSE* of 13.85, though minimum among all the models, is not small in an absolute sense. Hence, we look for alternative models for describing the sample data. To this end, we employ piecewise nonlinear growth models. The change point was identified visually by examining the graph (Figure 1) and scanning the values of the data during 2016 and found to be in

August 2016 ($t < 235$); t is the index of the time series representing month and year of the series. The change point divides the series into two segments, namely, segment one and segment two. We present this in Figure 1.

We hypothesize that the two segments can be modeled using one model applied to two segments separately or by two models. To evaluate this hypothesis, we fitted all 25 model combinations. To validate the piecewise nonlinear growth models' performance, we have partitioned the sample data to the training set (March 1997 to December 2017) and the test set (January 2018 to December 2018). Considering that the valuable information is present in the recent observations, we have retained only 12 observations (*i.e.*, one year's data) in the test set.

The piecewise nonlinear growth models are fitted to the training set, and the performance of the model was evaluated on the test set. Out of all 25 combinations, not all combinations converged or resulted in admissible parameter estimates. The combinations which converged and resulted in admissible parameter estimates are presented here. Two combinations, namely the Gompertz model for both the segments (let us name it Gompertz-Gompertz (*G-G*)), and the Gompertz model for Segment 1 and Bass model for Segment 2 (let us name it Gompertz-Bass (*G-B*)) were found to be comparable. We present here two sets of initial values, namely, $K = 200$, $B = 2$, $r = 0.05$, $m = 400$, $p = 0.00005$, and $q = 0.05$; and $K = 300$, $B = 2$, and $r = 0.05$, $m = 400$, $p = 0.00005$, $q = 0.05$. The first three parameters pertain to the Gompertz model, and the following three parameters pertain to the Bass model. These are obtained by combining linearization, intelligent guesses, and property of the model. We present the results in Table 2.

Table 2: Results of fitting piecewise nonlinear growth model to the analysis sample

Fit Statistics/Model	*G-G*	*G-B*
R^2	0.99	0.99
RMSE	6.94	7.14
MAPE	0.46	0.43
Parameter estimate (only maximum potential)		
K (millions)	384.07	
m (millions)		421.54

The results of the *G-G* model presented above are that of a local minimum. The algorithm failed to converge to a global minimum even when widely separated initial values were used. Therefore, it is prudent not to compare the results of *G-G* to that of *G-B*, which resulted in global convergence. However, for the sake of completeness, we presented the results here. The performance of the models (*G-G* and *G-B*) is compared using the test sample. The *RMSE* of the *G-G* model in the test sample was found to be 71.78, whereas the same value of the *G-B* model was 13.12. In the test sample, the value of the *RMSE* of the *G-B* model is much better than the *G-G* model. Furthermore, in the test sample, the value of the *RMSE* of the *G-B* model is marginally better than the only Gompertz model. Let us compare the test series results to that of the exponential smoothing model to decide the final model.

We describe below the exponential smoothing model (also known as Holt's model) (Gardner, 1985, Hanke and Wichern, 2013):

$$y_{t+1} = \alpha x_t + (1 - \alpha)\left(y_t + T_t\right) \tag{22}$$

$$T_{t+1} = \gamma(y_{t+1} - y_t) + (1 - \gamma)\, T_t \quad (23)$$

$$H_{t+h} = y_{t+1} + hT_{t+1}, \quad (24)$$

where, y_{t+1} = smoothed value for period $t+1$, α = smoothing constant for the level $(0 < \alpha < 1)$, x_t = observed value in period t, T_{t+1} = trend estimate, γ = Smoothing constant for the trend estimate $(0 < \gamma < 1)$, h = number of periods ahead to be forecast, and H_{t+h} = Holt's forecast value for period $t+h$.

The estimated values of alpha and gamma parameters of Holt's model were 0.933 and 0.134, respectively. The *RMSE* of Holt's model was found to be 2.883 and 14.97 for the training and the test samples, respectively. Therefore, the *RMSE* of the piecewise nonlinear growth model comprising Gompertz and Bass models was found to be superior to Holt's model. The residual diagnostics of the piecewise nonlinear model was found to deviate from the assumptions of normality and independence. However, both these deviations are mild and ignored because of the superior comparative performance of the model. Given the above, it can be concluded that the data generating process of the mobile subscribers' data of the Circle A was piecewise nonlinear, which can be modeled by Gompertz and the Bass models. We present the parameters of the final model in Table 3.

Table 3: Parameter estimates of piecewise nonlinear model

Parameters	*K*	*B*	*r*
Estimates	291.965	86.168	.031
Parameters	*m*	*p*	*q*
Estimates	421.545	7.328×10^{-9}	.07

In the final model, the estimated value of the parameter '*p*' is minimal. Such a small value, which is not exactly equal to zero but near zero, can occur and be meaningful in the present context as the data on which the models have been fitted are in millions. We present the final fitted model below:

$$N(t) = (t < T^*) * (291.965 * exp^{-86.168 * exp^{-.031 * t}})$$

$$+ (t >= T^*) * 421.545 \frac{1 - exp^{-(7.328 \times 10^{-9} + .07)\, t}}{1 + \left(\frac{.07}{7.328 \times 10^{-9}}\right) * exp^{-(7.328 \times 10^{-9} + .07)\, t}}$$

The actual (dotted line) and the predicted (solid line) trajectory of the mobile subscribers are depicted in Figure 2. We present the forecasted number of subscribers using the piecewise nonlinear growth model (Gompertz-Bass) for the last three quarters of 2020 in Table 4.

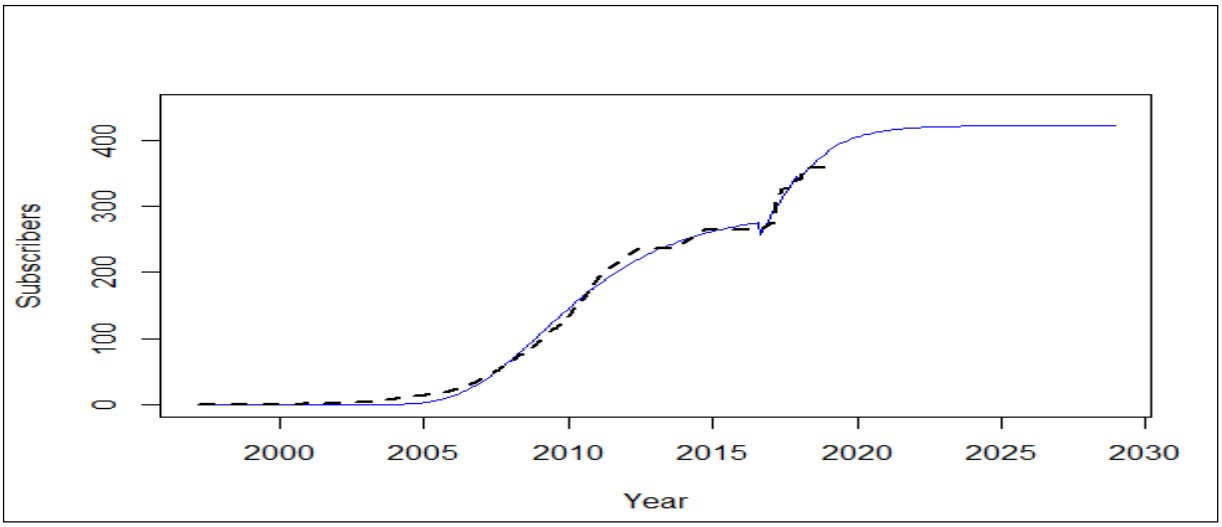


Figure 2: Actual (dotted) and predicted (solid line) mobile subscribers (in millions) in Circle A

Table 4: The forecasted subscribers (in millions) of Circle A in 2020

Month	Forecast		Month	Forecast
April	407.62		September	411.62
May	408.53		October	412.27
June	409.38		November	412.88
July	410.18		December	413.45
August	410.92			

It is evident from the predicted values and the graph (Figure 2) that the circle's maximum potential, which is 421.545 million, is predicted to be achieved in November 2027. However, it has been found earlier that the predictions from the growth models are conservative, and the maximum potential usually is attaining much before the model-predicted date (Das, 2013). Moreover, the data reported by TRAI is much higher than the data reported by COAI. As this study is based on COAI data, actual values are likely to be higher than the forecast provided in this paper.

5. Conclusions

We found the piecewise nonlinear growth model comprising Gompertz and Bass to be appropriate for describing the subscribers' data in Circle A. It confirms our assertion that the underlying data generating process can be divided into two segments, which shows strong evidence of ushering a new growth phase. Despite several issues in the telecom sector, Circle A comprises industrially advanced states like Maharashtra, Gujarat, Andhra Pradesh, Karnataka, and Tamil Nadu, which has entered into a new growth cycle. If this new cycle continues, it is likely to impact the Government in terms of revenue collection and the first- and second-degree stakeholders. As the post-COVID-19 economic scenario is different from the pre-COVID-19 economic scenario, the data

policy and pricing of the 5G spectrum are required to be such that it encourages more usages so that the current growth momentum continues.

Acknowledgments

The authors are grateful to the reviewers for diligent comments.

References

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, **15(5**), 215-227.

Birgit, Strikholm (2006). Determining the number of breaks in a piecewise linear regression model. *SSE/EFI Working Paper Series in Economics and Finance*, **648**, Stockholm School of Economics, The Economic Research Institute (EFI), Stockholm.

COAI (2018). *Annual Report 2018-19*. Cellular Operators Association of India. New Delhi, India.

Das, P. K. (2013). Nonlinear Statistical Model and its Applications to Diffusion of Mobile Telephony in India. *Annals of Pure and Applied Mathematics*, **5** (**1**), 64-70.

Draper, N., and Smith, H. (1998). *Applied Regression Analysis*, 3e. John Wiley, New York.

ET (2019). https://telecom.economictimes.indiatimes.com/news/telecom (Retrieved March 28, 2020).

Gardner, E. S. (1985). Exponential smoothing. The state of the art. *Journal of Forecasting*, **4**, 1-28.

Gruber, H. (2001). Competition and innovation: The diffusion of mobile telecommunications in Central and Eastern Europe. *Information Economics and Policy*, **13**, 19-34.

GSMA (2019). www.gsma.com (Retrieved March 28, 2020).

Hanke, John E., and Wichern, Dean W. (2013). *Business Forecasting*. PHI, New Delhi.

Hedau, Vandit, and Soni, Pragya, (2016). Increment of mobile subscriber in India: Gompertz curve. *Int. J. Res. Innov. Applied Science*, **1**(**1**), 1- 4.

IBM (2019). IBM Corp. Released 2019. *IBM SPSS Statistics for Windows*, Version 26.0. Armonk, NY: IBM Corp.

Intriligator, Michael D., Bodkin, Ronald, and Hsiao, Cheng (1996). *Econometric Models, Techniques, and Applications*. Prentice-Hall.

Kvalseth, T. O. (1985). Cautionary note about $R^{2.}$. *Amer. Statistician*, **39**, 279-85.

Linden, Ariel, (2015). Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal*, **15(2**), 480-500.

Maceina, M. J. (2007). Use of piecewise nonlinear models to estimate variable size-related mortality rates. *North American Journal of Fisheries Management*, **27**, 971-977.

Malash, G. F. and El-khaiary, M. I. (2010). Piecewise linear regression: A statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models. *Chemical Engineering Journal*, **163(3)**, 256-263.

McZgee, V. E. and Carleton, W. T. (1970). Piecewise regression. *Journal of the American Statistical Association*, **65**(**331**), 1109-1124.

Michalakelis, Christos, Varoutas, Dimitris and Sphicopoulos, Thomas (2008). Diffusion models of mobile telephony in Greece. *Telecommunications Policy*, **32**, 234-245.

Morrell, Christopher, H., Pearson, Jay, D., Carter Ballentine, H. and Brent, Larry, J. (1995). Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association*, **90**(**425**), 45-53.

Nocedal, J., and Wright, S. (2006). *Numerical Optimization*, 2nd *Edition*. Springer.

Oh, Kyong Joo and Kim, Kyoung-jae (2002). Analyzing stock market tick data using piecewise nonlinear model. *Expert Systems with Applications*, **22**, 249-255.

R Core Team (2016). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratkowsky, David A. (1990). *Handbook of Nonlinear Regression Models*. Marcel Dekker, New York.

Rogers, Everett M. (2003). *Diffusion of Innovations, 5th Edition*. Free Press, New York.

Ross, Gavin J. S. (1990). *Nonlinear Estimation*. Springer-Verlag, New York.

RStudio Team (2015). RStudio: *Integrated Development for R*. RStudio, Inc., Boston, MA.

Scott, A., and Wild, C. J. (1991). Transformations and R^2. *Amer. Statistician*, **45**, 127-28.

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression*. Wiley, New York.

Sridhar, Varadharajan (2010). An econometric analysis of mobile services growth across regions of India. *Netnomics*, **11**, 205-220.

Vanli, N. Denizcan and Serdar kozat, Suleyman (2014). A comprehensive approach to universal piecewise nonlinear regression based on trees. *IEEE Transactions on Signal Processing*, **62**(**20**), 5471-5486.

Vieth, E. (1989). Fitting piecewise linear regression functions to biological responses. *Journal of Applied Physiology*, **67**(**1**), 390-396.

TRAI (2018). *Press Release No. 86/2018*. Telecom Regulatory Authority of India, New Delhi, India.

Wenrong, W., Xie, M., and Tsui, K. (2006). Forecasting of mobile subscriptions in Asia Pacific using Bass diffusion model. *IEEE International Conference on Management of Innovation and Technology*, **1**, 300-303.

Wikipedia (2020). *Telecommunications in India*. In *Wikipedia*. Retrieved April 1, 2020, from https://en.wikipedia.org/wiki/Telecommunications-in-India.

Wu, Feng-Shang, and Chu, Wen Lin (2010). Diffusion models of mobile telephony. *Journal of Business Research*, **63**(**5**), 497-501.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 161-179

Estimating Sensitive Population Proportion Permitting Options for Various Respondents' Choices

Purnima Shaw[1] **and Sanghamitra Pal**[2]
[1]*Department of Statistics and Information Management, Reserve Bank of India*
[2]*Department of Statistics, West Bengal State University*

Received: 31 May 2020; Revised: 08 December 2020; Accepted: 11 December 2020

Abstract

Randomized Response (RR) Techniques (RRT's) are employed to reduce possible bias in gathering data related to sensitive characteristics. Alternatively, Item Count Techniques (ICT's) are also used for indirect questioning related to stigmatizing characteristics. Anticipating that a characteristic may be viewed as stigmatizing by some of the potential respondents but as innocuous by the rest who may not hesitate to give out direct responses (DR), literature is already developed as Optional Randomized Response (ORR) Techniques (ORRT's) permitting respondents to answer either an RR or a DR, exercising respective judgments. In this paper, two ORR techniques relating to qualitative sensitive characteristics are proposed allowing individuals selected by a general sampling scheme to choose DR, RR or ICT according to his/her own choice. Based on simulation results, estimates obtained from the proposed techniques are competitive to those obtained from an existing ORRT.

Key words: Item count technique; Optional randomized response technique; Stigmatizing characteristic; Unequal probability sampling.

AMS Subject Classification: 62 DO5

1. Introduction

Let $\boldsymbol{U} = (1,2,\dots,i,\dots,N)$ denote a finite population of a known number N of persons. Let y_i be a stigmatizing variate value such that $y_i = 1$ or 0, if the i^{th} $(i = 1,2,\dots,N)$ person bears a stigmatizing characteristic A or its complement A^C, respectively. Our objective is to estimate a finite population proportion of individuals bearing A, *i.e.*,

$$\theta = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1}$$

Warner's (1965) RRT is a well-known device for estimating θ. In this technique, a sampled person i is provided with a box containing similar cards marked A and A^C in proportions $p(\neq 0.5)$ and $(1-p)$, respectively. The individual's response is 1, if the card drawn randomly by him/her matches his/her characteristic and the response is 0, if there is no match. Warner's (1965) RR device was followed by several developments in which selection of sample was restricted to Simple Random Sampling with Replacement (SRSWR). Chaudhuri (2011) and Chaudhuri and Christofides (2013) recommended unequal probability sampling for selecting the units from the population and explained that RRT's are not conditioned by the sampling

Corresponding Author: Sanghamitra Pal
E-mail address: mitrapal2013@gmail.com

schemes. With this amendment, using RR's, y_i can be unbiasedly estimated for each i, followed by estimation of θ and estimate of its standard error.

As respondents may be suspicious of revelation of their privacy in RRT, the ICT, also known as the Block Total Response or the Unmatched Count Technique was introduced by Raghavarao and Federer (1979), Miller (1984) and Miller *et al*. (1986). Further developments in this area include those of Chaudhuri and Christofides (2007), in which two independent samples are required to be selected from U. A questionnaire is provided to the participants in the first sample in which there are G innocuous item statements and the $(G+1)^{th}$ item is —"I bear characteristic A or F", with F as an innocuous characteristic unrelated to A. Another questionnaire is provided to the participants in the second sample in which the same G innocuous item-statements along with the $(G+1)^{th}$ statement —"I do not bear characteristic A or I do not bear characteristic F", are present. Each of the participants responds the number of statements out of $(G+1)$ that are valid for him/her without revealing the answers to the individual statements. Using the responses from the two independent samples, θ and its standard error can be estimated. The developments in this area include those of Chaudhuri and Christofides (2013) and Shaw (2016) among others.

While some individuals may consider A as sensitive, others may prefer giving a direct response (DR). To tackle such situations, ORR devices were contributed by Arnab (2004), Chaudhuri and Saha (2005), Pal (2008), Mehta *et al.* (2012) and Sihm and Gupta (2015), among others. Several ORR devices are elaborated in Arnab and Rueda (2016). In the ORR device by Chaudhuri and Dihidar (2009) as explained in Chaudhuri (2011), each individual i, $i = 1,2,\dots,N$, in the population, bears an unknown probability C_i, $0 \le C_i \le 1$, to opt for giving a direct answer and a probability $(1-C_i)$ for preferring an RR. Respondents in a sample s drawn from U, are requested to either answer directly about A or provide an RR. However, they are instructed not to reveal the option chosen by them. Another similar response, independent of the first response, is collected from the same set of individuals. Pal (2007) developed an optional method in which, a sampled individual i in s is given the option to either provide an RR or answer to an ICT questionnaire, without revealing the choice of response to the investigator.

It is observed that all the ORR devices existing in the literature provide only two response options to the sampled individuals, *i.e.*, either DR and RR or RR and ICT. It is anticipated that while a few individuals in the population may prefer DR, some may opt for RR and the rest may be comfortable in answering to an ICT questionnaire. Motivated to fill up this gap in the literature, a generalized version of ORR device providing all the three modes of responses *viz.*, DR, RR and ICT is proposed in Section 2. This device mandates selection of 2 independent samples from the population by using a general sampling scheme. An alternative ORR device proposed in Section 3 requires selection of three independent samples chosen by a general sampling scheme. Section 4 provides a derivation of optimum allocation of sample sizes for a given cost of survey. In Pal's (2007) ORRT, there are two response options, *viz.*, RR and ICT; however, option for DR is not provided. So, we have compared our new ORRT's with Pal (2007). Hence in Section 5, performances of the two proposed devices have been compared with the performance of Pal (2007) ORR device, on the basis of a simulated data. The concluding remarks are presented in Section 6.

2. Proposed ORR Device Using Two Independent Samples

A respondent $i(i = 1,2,\dots,N)$, in the population, bears an unknown probability $C_{1i}(0 \le C_{1i} \le 1)$, with preference for opting a DR, a probability $C_{2i}(0 \le C_{1i} + C_{2i} \le 1)$, for an RR

and with the remaining probability $(1 - C_{1i} - C_{2i}), 0 \leq 1 - C_{1i} - C_{2i} \leq 1$, for an ICT. Consider a sample s_1 selected from $\boldsymbol{U}$ according to an unequal probability sampling design P admitting positive first order and second order inclusion probabilities $\pi_i = \sum_{s_1 \ni i} P(s_1)$, $\pi_{ij} = \sum_{s_1 \ni i,j} P(s_1), i \neq j, (i,j = 1,2,\dots,N)$. A respondent i is provided with options to either give a DR after multiplying with a constant or answer as per an RR device or answer to an ICT questionnaire, without divulging the chosen option.

If a respondent opts for DR, then, he/ she has to multiply the direct answer y_i with 2 and then give the resulting number in his/her response. In the option for RR, the respondent is requested to multiply his/ her value y_i with 2 and then add it with a number, say a_{11i}, randomly chosen from $(1,2,3,\dots,G-1)$. The questionnaire for ICT consists of G innocuous item statements, the $(G+1)^{th}$ statement being "I have characteristic A or F", *i.e.*, $(A \cup F)$, where F is an innocuous characteristic unrelated to A. A respondent opting for ICT, has to answer the total number of statements holding true for him/ her, say t_{1i}. Considering $f_i = 1$ or 0, if the i^{th} person bears innocuous characteristic F or its complement F^C, respectively, t_{1i} can be expressed as,

$$t_{1i} = \sum_{h=1}^{G} u_{ih} + y_i + f_i - y_i f_i \tag{2}$$

where, u_{ih} takes value 1 if the individual i bears the h^{th} innocuous characteristic, $h = 1,2,\dots,G$. Consider, the i^{th} respondent's answer as z_{11i}, where,

$$z_{11i} = \begin{cases} 2y_i \; with \; probability \; C_{1i}, & for \; DR \\ (2y_i + a_{11i}) \; with \; probability \; C_{2i}, & for \; RR \\ t_{1i} \; with \; probability \; (1 - C_{1i} - C_{2i}), & for \; ICT \end{cases} \tag{3}$$

The respondent i is requested to provide another response, say z_{12i}, independent of z_{11i}, following the same procedure. Let a_{12i} be the random number chosen from $(1,2,3,\dots,G-1)$ by the respondent opting RR, independent of the selection of a_{11i},

$$z_{12i} = \begin{cases} 2y_i \; with \; probability \; C_{1i}, & for \; DR \\ (2y_i + a_{12i}) \; with \; probability \; C_{2i}, & for \; RR \\ t_{1i} \; with \; probability \; (1 - C_{1i} - C_{2i}), & for \; ICT \end{cases} \tag{4}$$

The set of all possible answers for the DR is $\{0,2\}$, for the RR is $\{1,2,3,\dots,G+1\}$ and that for the ICT questionnaire is $\{0,1,2,\dots,G+1\}$. Hence, the investigator remains unaware of the respondent's choice. Taking E_R and V_R as the RR-based expectation and variance operators, respectively,

$$\begin{aligned} E_R(z_{11i}) &= E_R(z_{12i}) \\ &= 2C_{1i}y_i + C_{2i}\left(2y_i + \frac{G}{2}\right) + (1 - C_{1i} - C_{2i})\left(\sum_{h=1}^{G} u_{ih} + y_i + f_i - y_i f_i\right) \end{aligned} \tag{5}$$

Taking the concept of inter-penetrating network of sub-samples developed by Mahalanobis in 1936, consider,

$$r_{1i} = \frac{z_{11i} + z_{12i}}{2}, \quad v_{1i} = \frac{(z_{11i} - z_{12i})^2}{4} \tag{6}$$

Then,
$$E_R(r_{1i}) = E_R(z_{11i}) = E_R(z_{12i}) \tag{7}$$

and,
$$E_R(v_{1i}) = V_R(r_{1i}) \tag{8}$$

Consider a second sample s_2, selected from $\boldsymbol{U}$ (independent of the selection of s_1) according to an unequal probability sampling design P admitting positive first order and second order inclusion-probabilities $\pi_k = \sum_{s_2 \ni k} P(s_2)$, $\pi_{kl} = \sum_{s_2 \ni k,l} P(s_2)$, $k \neq l(k, l = 1,2, \dots, N)$. The selected individuals in s_2 are provided the options for DR, RR and ICT, which are slightly different from the device used for the first sample. A respondent k is provided with options to either give DR added with a constant or answer as per an RR device or answer according to an ICT questionnaire, without revealing the response medium opted by him/ her.

If DR is chosen, then, y_k is to be added with 1. If RR is chosen, then y_k is to be added with $(1 + a_{21k})$, where a_{21k} is randomly chosen from $(0,1,2, \dots, G)$. The ICT questionnaire contains the same G innocuous statements as in the questionnaire used for the first sample, with the $(G+1)^{th}$ statement being "I do not bear characteristic A or I do not bear characteristic F", *i.e.*, $(A^C \cup F^C)$ and the $(G+2)^{th}$ statement being "I bear characteristic F". A respondent opting for ICT, answers t_{2k}, where,

$$t_{2k} = \sum_{h=1}^{G} u_{kh} + 1 - y_k f_k + f_k \tag{9}$$

where, u_{kh} takes value 1 if the individual k bears the h^{th} innocuous characteristic, $h = 1,2, \dots, G$. Consider, the k^{th} respondent's answer as z_{21k}, where,

$$z_{21k} = \begin{cases} y_k + 1 \; with \; probability \; C_{1k}, & for \; DR \\ (y_k + 1 + a_{21k}) \; with \; probability \; C_{2k}, & for \; RR \\ t_{2k} \; with \; probability \; (1 - C_{1k} - C_{2k}), & for \; ICT \end{cases} \tag{10}$$

The respondent k is requested to provide another response, say z_{22k}, independent of z_{21k}, following the same procedure. Let a_{22k} be the number randomly chosen from $(0,1,2, \dots, G)$ by the respondent, independent of the selection of a_{21k},

$$z_{22k} = \begin{cases} y_k + 1 \; with \; probability \; C_{1k}, & for \; DR \\ (y_k + 1 + a_{22k}) \; with \; probability \; C_{2k}, & for \; RR \\ t_{2k} \; with \; probability \; (1 - C_{1k} - C_{2k}), & for \; ICT \end{cases} \tag{11}$$

The set of all possible answers for DR being $\{1,2\}$, for RR being $\{1,2,3, \dots G+2\}$ and for the ICT being $\{1,2,3, \dots G+2\}$, the medium of response chosen by the respondent is not revealed. Then, similar to (6), taking,

$$r_{2k} = \frac{z_{21k} + z_{22k}}{2}, \quad v_{2k} = \frac{(z_{21k} - z_{22k})^2}{4} \tag{12}$$

gives,
$$E_R(r_{2k}) = E_R(z_{21k}) = E_R(z_{22k}) \tag{13}$$

and, $$E_R(v_{2k}) = V_R(r_{2k}) \tag{14}$$

We consider the Horvitz Thompson (1952) estimator e to estimate θ, where,

$$e = 1 + \frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i} - \frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k} \tag{15}$$

Then, assuming E_P and V_P as the design-based expectation and variance operators respectively,

$$E(e) = E_R E_P(e) = E_P E_R(e) = \frac{1}{N}\sum_{i=1}^{N} y_i = \theta \tag{16}$$

Hence, e is an unbiased estimator of θ. Now, taking clue from Chaudhuri and Pal (2002), variance of e can be expressed as,

$$V(e) = V\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_P V_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V_P E_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + E_P V_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right) + V_P E_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_R V_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V_R E_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + E_R V_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right) + V_R E_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right)$$

$$= E_R\left[\frac{1}{N^2}\left\{\sum_{i}^{N}\sum_{<j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{r_{1i}}{\pi_i} - \frac{r_{1j}}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{\beta_i}{\pi_i}{r_{1i}}^2\right\}\right] + \frac{1}{N^2}\sum_{i=1}^{N} V_R(r_{1i})$$

$$+E_R\left[\frac{1}{N^2}\left\{\sum_{k}^{N}\sum_{<l}^{N}(\pi_k\pi_l - \pi_{kl})\left(\frac{r_{2k}}{\pi_k} - \frac{r_{2l}}{\pi_l}\right)^2 + \sum_{k=1}^{N}\frac{\beta_k}{\pi_k}{r_{2k}}^2\right\}\right] + \frac{1}{N^2}\sum_{k=1}^{N} V_R(r_{2k}) \tag{17}$$

writing,

$$\beta_i = 1 + \frac{1}{\pi_i}\sum\nolimits_{j\neq i}^{N}\pi_{ij} - \sum\nolimits_{i=1}^{N}\pi_i, \qquad \beta_k = 1 + \frac{1}{\pi_k}\sum\nolimits_{k\neq l}^{N}\pi_{kl} - \sum\nolimits_{k=1}^{N}\pi_k, \tag{18}$$

If every sample s_1 and s_2 contains a common number of distinct units in it, then, $\beta_i = 0\ \forall$ i and $\beta_k = 0\ \forall$ k throughout in $V(e)$ above.
Then, taking clue from Chaudhuri and Pal (2002), an unbiased estimator for $V(e)$ is,

$$v(e) = \frac{1}{N^2}\left\{\sum_{i<}\sum_{j\in s_1}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r_{1i}}{\pi_i} - \frac{r_{1j}}{\pi_j}\right)^2 + \sum_{i\in s_1}\frac{\beta_i}{{\pi_i}^2}{r_{1i}}^2\right\} + \frac{1}{N^2}\sum_{i\in s_1}\frac{v_{1i}}{\pi_i} \tag{19}$$

$$+\frac{1}{N^2}\left\{\sum_{k<}\sum_{l\in s_2}\left(\frac{\pi_k\pi_l-\pi_{kl}}{\pi_{kl}}\right)\left(\frac{r_{2k}}{\pi_k}-\frac{r_{2l}}{\pi_l}\right)^2+\sum_{k\in s_2}\frac{\beta_k}{{\pi_k}^2}{r_{2k}}^2\right\}+\frac{1}{N^2}\sum_{k\in s_2}\frac{v_{2k}}{\pi_k}$$

with $\beta_i = 0\ \forall$ i and $\beta_k = 0\ \forall$ k in $v(e)$ when applicable. Hence, $v(e)$ is an unbiased estimator of $V(e)$, such that $E\{v(e)\} = E_P E_R\{v(e)\} = E_R E_P\{v(e)\} = V(e)$. A $100(1-\alpha)\%$ Confidence Interval for θ is, $[L, U]$, where,

$$L = e - \left\{\tau_{\alpha/2}\sqrt{v(e)}\right\}, \qquad U = e + \left\{\tau_{\alpha/2}\sqrt{v(e)}\right\} \tag{20}$$

where, $\tau_{\alpha/2}$ is the upper $\alpha/2$ point of $N(0,1)$ distribution

This device, although provides three choices to the respondents, it has a limitation. A respondent may prefer to give direct value to the investigator, instead of multiplying or adding it with a constant. The authors have resolved this issue in the proposed device in Section 3.

3. Proposed ORR Device Using Three Independent Samples

In this device, a respondent i in the first sample s_1 opting DR, has to answer y_i directly. For giving an RR, y_i is to be multiplied with a number, say a'_{11i}, randomly chosen from $(0,1,2,\dots,G+1)$. The ICT questionnaire contains G innocuous item statements and the $(G+1)^{th}$ statement is "I bear characteristic A or F". Let t'_{1i} be the total number of statements valid for respondent i who has chosen ICT. Then,

$$t'_{1i} = \sum_{h=1}^{G} u_{ih} + y_i + f_i - y_i f_i \tag{21}$$

where, u_{ih} takes value 1 if the individual i bears the h^{th} innocuous characteristic, $h = 1,2,\dots,G$. Consider, the i^{th} respondent's answer as z'_{11i},

$$z'_{11i} = \begin{cases} y_i \text{ with probability } C_{1i}, & \text{for DR} \\ a'_{11i}y_i \text{ with probability } C_{2i}, & \text{for RR} \\ t'_{1i} \text{ with probability } (1-C_{1i}-C_{2i}), & \text{for ICT} \end{cases} \tag{22}$$

The respondent is requested to provide another response, say z'_{12i}, independent of z'_{11i}, following the same procedure. Let a'_{12i} denote the number randomly chosen from $(0,1,2,\dots,G+1)$ by the respondent, independent of the selection of a'_{11i}, then,

$$z'_{12i} = \begin{cases} y_i \text{ with probability } C_{1i}, & \text{for DR} \\ a'_{12i}y_i \text{ with probability } C_{2i}, & \text{for RR} \\ t'_{1i} \text{ with probability } (1-C_{1i}-C_{2i}), & \text{for ICT} \end{cases} \tag{23}$$

The set of all possible answers for DR is $\{0,1\}$, for RR is $\{0,1,2,\dots G+1\}$ and that for ICT questionnaire is $\{0,1,2,\dots G+1\}$. Hence, the response option chosen by the respondent remains unknown. Then, taking

$$r'_{1i} = \frac{z'_{11i} + z'_{12i}}{2}, \quad v'_{1i} = \frac{(z'_{11i} - z'_{12i})^2}{4} \tag{24}$$

gives,

$$E_R(r'_{1i}) = E_R(z'_{11i}) = E_R(z'_{12i}) \tag{25}$$

and,

$$E_R(v'_{1i}) = V_R(r'_{1i}) \tag{26}$$

Respondents in the second sample s_2 are also provided with all the three response options but the ORR device in this case is differently designed as compared to the device used for the first sample. The RR is obtained by adding a number, say a'_{21k}, randomly chosen from $(0,1,2,\dots,G+1)$ with another number, say b'_{21k}, randomly chosen from $(0,1,2,\dots,H)$ and then multiplying this sum with the $y-$value. The ICT questionnaire contains the same G innocuous item statements as in the questionnaire used for the first sample, with an additional set of H innocuous item statements and the $(G+H+1)^{th}$ statement being "I do not bear characteristic A or I do not bear characteristic F". A sampled individual choosing ICT, has to answer, say t'_{2k}, where,

$$t'_{2k} = \sum_{h=1}^{G} u_{kh} + \sum_{o=1}^{H} w_{ko} + 1 - y_k f_k \tag{27}$$

where, u_{kh} takes value 1 if the individual k bears the h^{th} innocuous characteristic, $h = 1,2,\dots,G$ and w_{ko} takes value 1 if the individual k bears the o^{th} innocuous characteristic, $o = 1,2,\dots,H$. Let the k^{th} sampled individual's answer be z'_{21k}, where

$$z'_{21k} = \begin{cases} y_k \; with \; probability \; C_{1k}, & for \; DR \\ (a'_{21k} + b'_{21k}) y_k \; with \; probability \; C_{2k}, & for \; RR \\ t'_{2k} \; with \; probability \; (1 - C_{1k} - C_{2k}), & for \; ICT \end{cases} \tag{28}$$

The respondent k is requested to provide another response, say z'_{22k}, independent of z'_{21k}, following the same procedure. Let a'_{22k} denote the number randomly chosen from $(0,1,2,\dots,G+1)$ and another number, say b'_{22k}, randomly chosen from $(0,1,2,\dots,H)$ by the respondent, independent of the selection of b'_{21k},

$$z'_{22k} = \begin{cases} y_k \; with \; probability \; C_{1k}, & for \; DR \\ (a'_{22k} + b'_{22k}) y_k \; with \; probability \; C_{2k}, & for \; RR \\ t'_{2k} with \; probability \; (1 - C_{1k} - C_{2k}), & for \; ICT \end{cases} \tag{29}$$

The sets of all possible answers for DR, RR and ICT questionnaire being $\{0,1\}$, $\{0,1,2,\dots G+H+1\}$ and $\{0,1,2,\dots G+H+1\}$, respectively, the respondent's choice remains hidden. Then, taking

$$r'_{2k} = \frac{z'_{21k} + z'_{22k}}{2}, \quad v'_{2k} = \frac{(z'_{21k} - z'_{22k})^2}{4} \tag{30}$$

gives,

$$E_R(r'_{2k}) = E_R(z'_{21k}) = E_R(z'_{22k}) \tag{31}$$

and,

$$E_R(v'_{2k}) = V_R(r'_{2k}) \tag{32}$$

A third sample s_3 is chosen (independent of the selection of s_1 and s_2) from $\boldsymbol{U}$ according to an unequal probability sampling design P with positive first and second order inclusion-probabilities $\pi_d = \sum_{s_3 \ni d} P(s_3)$, $\pi_{dq} = \sum_{s_3 \ni d,q} P(s_3), d \neq q(d, q = 1,2, \dots, N)$. The instructions for RR and ICT for the sampled individuals in this case differ from those in the first two samples. Respondents are free to choose any of the three response options and answer accordingly without revealing the chosen response option.

RR is generated by adding a number, say b'_{31d}, randomly chosen from $(0,1,2, \dots, H)$, with 1 and then multiplying the resulting number with the respondent's $y-$ value. All the H innocuous item statements in the second questionnaire are repeated in the current ICT questionnaire along with the $(H+1)^{th}$ statement "I do not bear characteristic F". Let the response for ICT, if chosen by the d^{th} respondent, be t'_{3d} where,

$$t'_{3d} = \sum_{o=1}^{H} w_{do} + 1 - f_d \tag{33}$$

where, w_{do} takes value 1 if the individual d bears the o^{th} innocuous characteristic, $o = 1,2, \dots, H$. Consider, the d^{th} respondent's answer as z'_{31d},

$$z'_{31d} = \begin{cases} y_d \; with \; probability \; C_{1d}, & for \; DR \\ (1 + b'_{31d}) y_d \; with \; probability \; C_{2d}, & for \; RR \\ t'_{3d} \; with \; probability \; (1 - C_{1d} - C_{2d}), & for \; ICT \end{cases} \tag{34}$$

The respondent d is requested to provide another response, say z'_{32d}, independent of z'_{31d}, following the same procedure. Then, taking b'_{32d} as the number randomly chosen from $(0,1,2, \dots, H)$ by the respondent, independent of the selection of b'_{31d},

$$z'_{32d} = \begin{cases} y_d \; with \; probability \; C_{1d}, & for \; DR \\ (1 + b'_{32d}) y_d \; with \; probability \; C_{2d}, & for \; RR \\ t'_{3d} \; with \; probability \; (1 - C_{1d} - C_{2d}), & for \; ICT \end{cases} \tag{35}$$

The sets of all possible responses for DR, RR and ICT are $\{0,1\}$, $\{0,1,2, \dots, H+1\}$ and $\{0,1,2, \dots, H+1\}$, respectively, thus indicating that the investigator is unaware of the choice of the respondent. Taking,

$$r'_{3d} = \frac{z'_{31d} + z'_{32d}}{2}, \quad v'_{3d} = \frac{(z'_{31d} - z'_{32d})^2}{4} \tag{36}$$

gives,

$$E_R(r'_{3d}) = E_R(z'_{31d}) = E_R(z'_{32d}) \tag{37}$$

and,

$$E_R(v'_{3d}) = V_R(r'_{3d}) \tag{38}$$

We consider the Horvitz Thompson (1952) estimator e', where,

$$e' = \frac{1}{N} \sum_{i \in s_1} \frac{r'_{1i}}{\pi_i} - \frac{1}{N} \sum_{k \in s_2} \frac{r'_{2k}}{\pi_k} + \frac{1}{N} \sum_{d \in s_3} \frac{r'_{3d}}{\pi_d} \tag{39}$$

Then,

$$E(e') = E_R E_P(e') = E_P E_R(e') = \frac{1}{N}\sum_{i=1}^{N} y_i = \theta \tag{40}$$

Hence, e' is an unbiased estimator of θ. Now, to find out the variance of e',

$$V(e') = V\left(\frac{1}{N}\sum_{i\in s_1}\frac{r'_{1i}}{\pi_i}\right) + V\left(\frac{1}{N}\sum_{k\in s_2}\frac{r'_{2k}}{\pi_k}\right) + V\left(\frac{1}{N}\sum_{d\in s_3}\frac{r'_{3d}}{\pi_d}\right)$$

$$= E_P V_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r'_{1i}}{\pi_i}\right) + V_P E_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r'_{1i}}{\pi_i}\right) + E_P V_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r'_{2k}}{\pi_k}\right) + V_P E_R\left(\frac{1}{N}\sum_{k\in s_2}\frac{r'_{2k}}{\pi_k}\right)$$
$$+ E_P V_R\left(\frac{1}{N}\sum_{d\in s_3}\frac{r'_{3d}}{\pi_d}\right) + V_P E_R\left(\frac{1}{N}\sum_{d\in s_3}\frac{r'_{3d}}{\pi_d}\right)$$

$$= E_R V_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r'_{1i}}{\pi_i}\right) + V_R E_P\left(\frac{1}{N}\sum_{i\in s_1}\frac{r'_{1i}}{\pi_i}\right) + E_R V_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r'_{2k}}{\pi_k}\right) + V_R E_P\left(\frac{1}{N}\sum_{k\in s_2}\frac{r'_{2k}}{\pi_k}\right)$$
$$+ E_R V_P\left(\frac{1}{N}\sum_{d\in s_3}\frac{r'_{3d}}{\pi_d}\right) + V_R E_P\left(\frac{1}{N}\sum_{d\in s_3}\frac{r'_{3d}}{\pi_d}\right)$$

$$= E_R\left[\frac{1}{N^2}\left\{\sum_{i}^{N}\sum_{<j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{r'_{1i}}{\pi_i} - \frac{r'_{1j}}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{\beta_i}{\pi_i}{r'_{1i}}^2\right\}\right] + \frac{1}{N^2}\sum_{i=1}^{N} V_R(r'_{1i})$$

$$+E_R\left[\frac{1}{N^2}\left\{\sum_{k}^{N}\sum_{<l}^{N}(\pi_k\pi_l - \pi_{kl})\left(\frac{r'_{2k}}{\pi_k} - \frac{r'_{2l}}{\pi_l}\right)^2 + \sum_{k=1}^{N}\frac{\beta_k}{\pi_k}{r'_{2k}}^2\right\}\right] + \frac{1}{N^2}\sum_{k=1}^{N} V_R(r'_{2k}) \tag{41}$$

$$+E_R\left[\frac{1}{N^2}\left\{\sum_{d}^{N}\sum_{<q}^{N}(\pi_d\pi_q - \pi_{dq})\left(\frac{r'_{3d}}{\pi_d} - \frac{r'_{3q}}{\pi_q}\right)^2 + \sum_{d=1}^{N}\frac{\beta_d}{\pi_d}{r'_{3d}}^2\right\}\right] + \frac{1}{N^2}\sum_{d=1}^{N} V_R(r'_{3d})$$

where,

$$\beta_i = 1 + \frac{1}{\pi_i}\sum_{j\neq i}^{N}\pi_{ij} - \sum_{i=1}^{N}\pi_i\,, \qquad \beta_k = 1 + \frac{1}{\pi_k}\sum_{k\neq l}^{N}\pi_{kl} - \sum_{k=1}^{N}\pi_k \tag{42}$$

and,

$$\beta_d = 1 + \frac{1}{\pi_d}\sum_{d\neq q}^{N}\pi_{dq} - \sum_{d=1}^{N}\pi_d$$

If every sample s_1, s_2 and s_3 contains a common number of distinct units in it, then, $\beta_i = 0\ \forall$ i, $\beta_k = 0\ \forall$ k and $\beta_d = 0\ \forall$ d throughout in $V(e)$ above, using Chaudhuri and Pal (2002). Then, taking clue from Chaudhuri and Pal (2002), an unbiased estimator of $V(e')$ is,

$$v(e') = \frac{1}{N^2}\left\{\sum_{i<}\sum_{j\in s_1}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r'_{1i}}{\pi_i} - \frac{r'_{1j}}{\pi_j}\right)^2 + \sum_{i\in s_1}\frac{\beta_i}{{\pi_i}^2}{r'_{1i}}^2\right\} + \frac{1}{N^2}\sum_{i\in s_1}\frac{v'_{1i}}{\pi_i} \tag{43}$$

$$+\frac{1}{N^2}\left\{\sum_{k<}\sum_{l\in s_2}\left(\frac{\pi_k\pi_l-\pi_{kl}}{\pi_{kl}}\right)\left(\frac{r'_{2k}}{\pi_k}-\frac{r'_{2l}}{\pi_l}\right)^2+\sum_{k\in s_2}\frac{\beta_k}{{\pi_k}^2}{r'_{2k}}^2\right\}+\frac{1}{N^2}\sum_{k\in s_2}\frac{v'_{2k}}{\pi_k}$$

$$+\frac{1}{N^2}\left\{\sum_{d<}\sum_{q\in s_3}\left(\frac{\pi_d\pi_q-\pi_{dq}}{\pi_{dq}}\right)\left(\frac{r'_{3d}}{\pi_d}-\frac{r'_{3q}}{\pi_q}\right)^2+\sum_{d\in s_3}\frac{\beta_d}{{\pi_d}^2}{r'_{3d}}^2\right\}+\frac{1}{N^2}\sum_{d\in s_3}\frac{v'_{3d}}{\pi_d}$$

with $\beta_i = 0\ \forall$ i, $\beta_k = 0\ \forall$ k and $\beta_d = 0\ \forall$ d in $v(e')$ when applicable. Hence, $v(e')$ is an unbiased estimator of $V(e')$, such that, $E\{v(e')\} = E_P E_R\{v(e')\} = E_R E_P\{v(e')\} = V(e')$. A $100(1-\alpha)\%$ Confidence Interval for θ is, $[L', U']$, where,

$$L' = e' - \left\{\tau_{\alpha/2}\sqrt{v(e')}\right\}, \qquad U' = e' + \left\{\tau_{\alpha/2}\sqrt{v(e')}\right\} \tag{44}$$

4. Optimum Sample Size Allocation for Fixed Survey Cost

As the two proposed ORR devices with options for DR, RR and ICT, mandate selection of multiple samples from the population, this section demonstrates a procedure to minimize the variance of the estimate of $\bar{Y}$ by assuming a fixed cost of the survey. Consider the ORR device in Section 2 and assume that both the samples (of sizes n_1 and n_2, say) are chosen independently from the population by following the Hartley and Rao (1962) sampling scheme and using variable x as the size measure for sample selection, with population total $X = \sum_{i=1}^{N} x_i$. Then, putting the expressions for first and second order inclusion probabilities and using (17), variance of the estimate $\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}$ obtained from the first sample can be written as,

$$V\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) = E_P V_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) + V_P E_R\left(\frac{1}{N}\sum_{i\in s_1}\frac{r_{1i}}{\pi_i}\right) = \frac{V_{11}}{n_1} + V_{12} \tag{45}$$

where,

$$V_{11} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{V_R(r_{1i})}{p_i} + \frac{1}{N^2}\left[\sum_{i}^{N}\sum_{<j}^{N} p_{ij}\left\{\frac{E_R(r_{1i})}{p_i} - \frac{E_R(r_{1j})}{p_j}\right\}^2\right], \tag{46}$$

$$V_{12} = \frac{1}{N^2}\left[\sum_{i}^{N}\sum_{<j}^{N}(p_i p_j - p_{ij})\left\{\frac{E_R(r_{1i})}{p_i} - \frac{E_R(r_{1j})}{p_j}\right\}^2\right], \tag{47}$$

$$p_i = \frac{x_i}{X}, \qquad p_j = \frac{x_j}{X}, \tag{48}$$

$$E_R(r_{1i}) = 2C_{1i}y_i + C_{2i}\left(2y_i + \frac{G}{2}\right) + (1 - C_{1i} - C_{2i})\left(\sum_{h=1}^{G} u_{ih} + y_i + f_i - y_i f_i\right), \tag{49}$$

$$V_R(r_{1i}) = \frac{E_R(r_{1i}^2) - E_R^2(r_{1i})}{2}, \tag{50}$$

with,

$$E_R(r_{1i}^2) = 4C_{1i}y_i + C_{2i}\left(4y_i + \frac{G^2}{3} - \frac{G}{6} + 4y_i\frac{G}{2}\right) + (1 - C_{1i} - C_{2i})\left(\sum_{h=1}^{G} u_{ih} + y_i + f_i - y_i f_i\right)^2 \quad (51)$$

Similarly, variance of the estimate $\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}$ obtained from the second sample can be expressed as,

$$V\left(\frac{1}{N}\sum_{k\in s_2}\frac{r_{2k}}{\pi_k}\right) = \frac{V_{21}}{n_2} + V_{22} \quad (52)$$

where,

$$V_{21} = \frac{1}{N^2}\sum_{k=1}^{N}\frac{V_R(r_{2k})}{p_k} + \frac{1}{N^2}\left[\sum_{k}^{N}\sum_{<l}^{N} p_{kl}\left\{\frac{E_R(r_{2k})}{p_k} - \frac{E_R(r_{2l})}{p_l}\right\}^2\right], \quad (53)$$

$$V_{22} = \frac{1}{N^2}\left[\sum_{k}^{N}\sum_{<l}^{N}(p_k p_l - p_{kl})\left\{\frac{E_R(r_{2k})}{p_k} - \frac{E_R(r_{2l})}{p_l}\right\}^2\right], \quad (54)$$

$$p_k = \frac{x_k}{X}, \qquad p_l = \frac{x_l}{X}, \quad (55)$$

$$E_R(r_{2k}) = C_{1k}(y_k + 1) + C_{2k}\left(y_k + 1 + \frac{G}{2}\right) + (1 - C_{1k} - C_{2k})\left(\sum_{h=1}^{G} u_{kh} + 1 - y_k f_k + f_k\right), \quad (56)$$

$$V_R(r_{2k}) = \frac{E_R(r_{2k}^2) - E_R^2(r_{2k})}{2}, \quad (57)$$

with,

$$E_R(r_{2k}^2) = C_{1k}(y_k + 1)^2 + C_{2k}\left\{(y_k + 1)^2 + \frac{G^2}{3} + \frac{G}{6} + 2(y_k + 1)\frac{G}{2}\right\} + (1 - C_{1k} - C_{2k})\left(\sum_{h=1}^{G} u_{kh} + 1 - y_k f_k + f_k\right)^2 \quad (58)$$

Hence,

$$V(e) = \frac{V_{11}}{n_1} + \frac{V_{21}}{n_2} + V_{12} + V_{22} \quad (59)$$

Let C be the total cost of the survey, C_0 be the overhead cost and C' be the cost per unit in the samples selected from the population. Then,

$$C = C_0 + C'(n_1 + n_2) \quad (60)$$

In order to find n_1 and n_2 under the above cost function, consider the following Lagrangian function with λ as the Lagrange multiplier,

$$\Phi = V(e) + \lambda(C - C_0) = \frac{V_{11}}{n_1} + \frac{V_{21}}{n_2} + \lambda C'(n_1 + n_2) + V_{12} + V_{22}$$
$$= \left(\sqrt{\frac{V_{11}}{n_1}} - \sqrt{\lambda C' n_1}\right)^2 + \left(\sqrt{\frac{V_{21}}{n_2}} - \sqrt{\lambda C' n_2}\right)^2 + 2\sqrt{\lambda C'}\left(\sqrt{V_{11}} + \sqrt{V_{21}}\right) + V_{12} + V_{22} \quad (61)$$

Thus, Φ is minimum when,

$$n_1 = \frac{\sqrt{V_{11}}}{\sqrt{\lambda C'}} \qquad n_2 = \frac{\sqrt{V_{21}}}{\sqrt{\lambda C'}} \tag{62}$$

Now, considering C as a pre-specified fixed survey cost, then, $C'(n_1 + n_2) = C - C_0$ gives,

$$\sqrt{\lambda} = \frac{\sqrt{C'}\left(\sqrt{V_{11}} + \sqrt{V_{21}}\right)}{(C - C_0)} \tag{63}$$

Hence,

$$n_1 = \frac{\sqrt{V_{11}}(C - C_0)}{C'\left(\sqrt{V_{11}} + \sqrt{V_{21}}\right)} \qquad n_2 = \frac{\sqrt{V_{21}}(C - C_0)}{C'\left(\sqrt{V_{11}} + \sqrt{V_{21}}\right)} \tag{64}$$

Thus, for fixed survey cost, the variance $V(e)$ is minimum if the two independent samples chosen from the population are of sizes as specified above.

For the ORR device proposed in Section 3, sizes of the three independent samples, say n_1, n_2 and n_3 for which variance of estimate e' is minimum for fixed survey cost, can be similarly worked out, as,

$$n_1 = \frac{\sqrt{V'_{11}}(C - C_0)}{C'\left(\sqrt{V'_{11}} + \sqrt{V'_{21}} + \sqrt{V'_{31}}\right)} \qquad n_2 = \frac{\sqrt{V'_{21}}(C - C_0)}{C'\left(\sqrt{V'_{11}} + \sqrt{V'_{21}} + \sqrt{V'_{31}}\right)}$$
$$n_3 = \frac{\sqrt{V'_{31}}(C - C_0)}{C'\left(\sqrt{V'_{11}} + \sqrt{V'_{21}} + \sqrt{V'_{31}}\right)} \tag{65}$$

where,

$$V'_{11} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{V_R(r'_{1i})}{p_i} + \frac{1}{N^2}\left[\sum_{i}^{N}\sum_{<j}^{N} p_{ij}\left\{\frac{E_R(r'_{1i})}{p_i} - \frac{E_R(r'_{1j})}{p_j}\right\}^2\right] \tag{66}$$

$$V'_{21} = \frac{1}{N^2}\sum_{k=1}^{N}\frac{V_R(r'_{2k})}{p_k} + \frac{1}{N^2}\left[\sum_{k}^{N}\sum_{<l}^{N} p_{kl}\left\{\frac{E_R(r'_{2k})}{p_k} - \frac{E_R(r'_{2l})}{p_l}\right\}^2\right] \tag{67}$$

$$V'_{31} = \frac{1}{N^2}\sum_{k=1}^{N}\frac{V_R(r'_{3d})}{p_k} + \frac{1}{N^2}\left[\sum_{k}^{N}\sum_{<l}^{N} p_{kl}\left\{\frac{E_R(r'_{3d})}{p_k} - \frac{E_R(r'_{3q})}{p_l}\right\}^2\right] \tag{68}$$

It may be noted that for both the proposed ORR devices, the probabilities of choosing DR, RR and ICT are different and unknown for each individual.

5. Numerical Illustration

To examine the performances of the Generalized ORR devices proposed in Sections 2 and 3, a simulated population of $N = 117$ individuals has been considered, wherein, A indicates consumption of alcohol, F indicates the individual's preference in playing football and z is the number of family members of the respondent. It is desired to estimate the population proportion of individuals consuming alcohol. A sample of size $n = 11$ individuals is drawn from the population by following Hartley and Rao's (1962) sampling scheme. In this scheme, a systematic sample is drawn by Probability Proportional to Size (PPS) method following a random arrangement of the population units. The size measure used for the sampling purpose is z.

For the sake of simplicity in notations, the device by Pal (2007), the generalized ORR device proposed with two independent samples (Section 2) and the alternative ORR device

with three independent samples (Section 3) are denoted as Device-I, Device-II and Device-III, respectively. For individuals opting to answer an RR, the specifications of the RR device are stated below:

(i) **Device-I:** An individual in the first or second sample, opting for RR has to choose a random number from $(0,1,2,\dots,G)$.

(ii) **Device-II:** For the first sample, RR in this device is generated by choosing a random number from $(1,2,3,\dots,G-1)$. For the second sample, a random number is to be chosen from $(0,1,2,\dots,G)$.

(iii) **Device-III:** This device for the first sample mandates selection of a random number from $(0,1,2,\dots,G+1)$. In the second sample, RR is computed by choosing two numbers randomly from $(0,1,2,\dots,G+1)$ and $(0,1,2,\dots,H)$, respectively. A random number is to be chosen from $(0,1,2,\dots,H)$ for generating RR in the third sample.

As per requirement of ICT, the two sets of $G=5$ and $H=4$ innocuous items statements denoted by B_1, B_2, B_3, B_4, B_5 and E_1, E_2, E_3, E_4, considered here are described below:

Set-1

B_1 : I like listening to music.
B_2 : I am diagnosed with liver disease.
B_3 : I am married.
B_4 : I am planning to buy a house.
B_5 : I love painting.

Set-2

E_1 : I like watching movies.
E_2 : I prefer cricket test matches over one day matches.
E_3 : I am currently employed.
E_4 : My birthday is in December.

The specifications of the ICT questionnaire followed are stated below:

(i) **Device-I:** For both the samples, G innocuous item statements in the questionnaire are as given in Set-1. The $(G+1)^{th}$ item in the questionnaire for the first sample is "I consume alcohol or I love playing football". On the other hand, the $(G+1)^{th}$ item in the questionnaire for the second sample is "I do not consume alcohol or I don't love playing football".

(ii) **Device-II:** G innocuous item statements from Set-1 are used for both the samples. In the questionnaire for the first sample, the $(G+1)^{th}$ statement is "I consume alcohol or I love playing football". In the questionnaire to be used for the second sample, the $(G+1)^{th}$ statement is "I do not consume alcohol or I don't love playing football" and the $(G+2)^{th}$ statement is "I love playing football".

(iii) **Device-III:** For the first sample, G innocuous item statements in the questionnaire are as given in Set-1. The $(G+1)^{th}$ statement is "I consume alcohol or I love playing football". For the second sample, $(G+H)$ innocuous item statements in the

questionnaire are provided in Set-1 and Set-2. The $(G+H+1)^{th}$ statement is "I do not consume alcohol or I don't love playing football". The H innocuous item statements in Set-2 are used in the questionnaire for the third sample. The $(H+1)^{th}$ statement is "I do not love playing football".

For each of the three devices, independent samples are drawn each of size $n = 11$. Various scenarios on different proportion of individuals in the sample opting for DR, RR and ICT for Devices I, II and III are identified. For each of these scenarios, e, $v(e)$, L and U for Device-II and e', $v(e')$, L' and U' for Device-III are calculated. Similarly, the proportion estimate e'', its variance estimate $v(e'')$ and confidence interval (L'', U'') are also computed for Device-I. The estimates are derived each time for $D = 1000$ re-samples drawn from the population and then to compare Devices II and III with Device-I, the following are calculated:

Average Estimates: $\frac{1}{D}\sum_{d=1}^{1000} e_d$, $\frac{1}{D}\sum_{d=1}^{1000} e'_d$ and and $\frac{1}{D}\sum_{d=1}^{1000} e''_d$,

Average Relative Efficiency (Device-II relative to Device-I): $\frac{\frac{1}{D}\sum_{d=1}^{1000} v(e''_d)}{\frac{1}{D}\sum_{d=1}^{1000} v(e_d)} 100$,

Average Relative Efficiency (Device-III relative to Device-I): $\frac{\frac{1}{D}\sum_{d=1}^{1000} v(e''_d)}{\frac{1}{D}\sum_{d=1}^{1000} v(e'_d)} 100$,

Average Relative Bias: $\left|\frac{\frac{1}{D}\sum_{d=1}^{1000} e_d - \theta}{\theta}\right|$, $\left|\frac{\frac{1}{D}\sum_{d=1}^{1000} e'_d - \theta}{\theta}\right|$ *and* $\left|\frac{\frac{1}{D}\sum_{d=1}^{1000} e''_d - \theta}{\theta}\right|$,

Actual Coverage Percentage for Devices I, II and III *viz.*, percentage of cases out of 1,000 re-samples, in which (L, U), (L', U') and (L'', U'') covers θ and Average Length of the 1,000 replicates of Confidence Intervals for θ for Devices I, II and III are also computed.

If Average Relative Efficiency of a proposed device relative to Device-I is more than 100, then the proposed device is more efficient than Device-I. On the other hand, lower the Average Relative Bias, better the device. Further, closer the Actual Coverage Percentage to 95% and smaller the Average Length, better is the performance of that device.

The Average Estimates (AE), Average Relative Bias (ARB), Actual Coverage Percentage (ACP) and Average Length (AL) obtained from Devices II and III are compared to those obtained from Device-I for various scenarios of individuals' choices on the medium of response. Further, the Average Relative Efficiency (ARE) of the estimates obtained from each of Device-II and Device-III relative to those calculated using Device-I are also derived. Few such comparisons are displayed in Tables 1 and 2. Table 3 provides a similar comparison of the performances of the proposed Devices II and III.

From Table 1, it is observed that Device-II outperforms Device-I marginally in all aspects *viz.*, ARE, ARB and ACP and AL. From Table 2, it is observed that Device-III is better than Device-I in respect of ARE, ARB, ACP as well as AL. Amongst Device-II and Device-III, Device-II shows better performance in terms of ARE, ARB and AL (Table 3). Hence, it can be safely concluded that the proposed Devices II and III are competitive with Device-I. The very purpose of proposing Devices II and III is to accommodate a variety of responses, *viz.*, DR, RR and ICT. The proposed devices not only fulfil this purpose but also perform efficiently in comparison to the existing device Pal (2007) with two response options.

Table 1: Comparison of performance of proposed ORR Device-II with Device-I by Pal (2007)

Sample proportion with a chosen response option					AE ($\theta = 0.67$)		ARE of Device II relative to Device I	ARB		ACP		AL	
Device-II			Device-I		Device			Device		Device		Device	
DR	RR	ICT	RR	ICT	II	I		II	I	II	I	II	I
0.4	0.4	0.2	0.8	0.2	0.67	0.58	103.6	0.005	0.131	97.9	94.8	4.57	4.57
0.6	0.1	0.3	0.7	0.3	0.64	0.60	131.0	0.052	0.098	97.7	95.4	3.80	4.42
0.7	0.1	0.2	0.8	0.2	0.73	0.76	152.4	0.096	0.130	97.8	95.4	3.51	4.47
0.8	0.1	0.1	0.9	0.1	0.68	0.70	210.6	0.021	0.050	96.7	95.4	3.11	4.57
0.6	0.2	0.2	0.8	0.2	0.66	0.71	148.2	0.008	0.055	98.0	96.3	3.76	4.64
0.7	0.2	0.1	0.9	0.1	0.67	0.68	166.4	0.003	0.016	97.8	96.3	3.56	4.57
0.4	0.4	0.2	0.4	0.6	0.67	0.75	113.4	0.005	0.125	97.9	94.8	4.57	4.66
0.4	0.2	0.4	0.2	0.8	0.71	0.62	112.9	0.062	0.071	98.3	95.8	4.32	4.58
0.6	0.2	0.2	0.2	0.8	0.66	0.62	162.1	0.008	0.069	98.0	96.6	3.76	4.71

Table 2: Comparison of performance of proposed ORR Device-III with Device-I by Pal (2007)

Sample proportion with a chosen response option					AE ($\theta = 0.67$)		ARE of Device III relative to Device I	ARB		ACP		AL	
Device-III			Device-I		Device			Device		Device		Device	
DR	RR	ICT	RR	ICT	III	I		III	I	III	I	III	I
0.7	0.1	0.2	0.8	0.2	0.66	0.76	115.4	0.008	0.130	100.0	95.4	4.16	4.47
0.6	0.2	0.2	0.8	0.2	0.67	0.71	111.7	0.000	0.055	100.0	96.3	4.49	4.64
0.7	0.2	0.1	0.9	0.1	0.67	0.68	158.5	0.003	0.016	99.8	96.3	3.74	4.57
0.6	0.3	0.1	0.9	0.1	0.67	0.64	117.8	0.007	0.047	99.5	94.9	4.27	4.62
0.8	0.1	0.1	0.9	0.1	0.70	0.63	181.2	0.052	0.059	99.8	94.7	3.30	4.41
0.6	0.1	0.3	0.7	0.3	0.70	0.60	120.5	0.041	0.104	100.0	95.0	4.63	4.76
0.7	0.1	0.2	0.1	0.9	0.66	0.66	127.0	0.008	0.015	100.0	95.5	4.16	4.50
0.8	0.1	0.1	0.1	0.9	0.65	0.72	173.0	0.025	0.076	99.7	92.5	3.38	4.33
0.6	0.2	0.2	0.2	0.8	0.67	0.62	122.2	0.000	0.069	100.0	96.6	4.49	4.71
0.7	0.2	0.1	0.2	0.8	0.67	0.65	160.8	0.003	0.026	99.8	93.6	3.74	4.46
0.6	0.3	0.1	0.3	0.7	0.67	0.64	113.5	0.007	0.050	99.5	94.8	4.27	4.40

Table 3: Comparison of performances of proposed ORR devices *viz.*, Device-II and Device-III

Sample proportion with a chosen response option			AE ($\theta = 0.67$)		ARE of Device II relative to Device III	ARB		ACP		AL	
DR	RR	ICT	Device			Device		Device		Device	
			II	III		II	III	II	III	II	III
0.1	0.3	0.6	0.68	0.73	139.9	0.018	0.088	96.0	99.4	5.00	6.26
0.1	0.8	0.1	0.63	0.74	104.6	0.060	0.109	95.1	98.9	5.25	5.74
0.2	0.3	0.5	0.69	0.61	155.9	0.028	0.082	97.1	99.5	4.82	6.38
0.2	0.6	0.2	0.72	0.73	125.0	0.075	0.094	96.9	99.2	5.18	5.89
0.3	0.5	0.2	0.65	0.64	128.8	0.035	0.046	97.5	99.3	4.91	5.71

0.4	0.4	0.2	0.67	0.58	134.4	0.005	0.140	97.9	99.5	4.57	5.38
0.4	0.2	0.4	0.68	0.64	159.9	0.021	0.045	97.9	99.8	4.46	5.87
0.6	0.1	0.3	0.64	0.73	140.8	0.052	0.096	97.7	99.9	3.80	4.77
0.8	0.1	0.1	0.68	0.71	103.5	0.012	0.055	97.0	100.0	3.16	3.31
0.6	0.3	0.1	0.65	0.60	114.7	0.023	0.099	97.9	99.9	3.91	4.27

Next, as the two proposed devices mandate selection of multiple samples from the population, an attempt has been made here to compute the optimum samples sizes, based on the discussion in Section 4. The probabilities of choosing DR, RR and ICT are different and unknown for each individual. However, for conducting the numerical computations, it is assumed that these probabilities are same for each individual in the population. Taking this assumption, the optimum sample sizes are calculated for various scenarios of proportion of individuals opting for DR, RR and ICT. Tables 4 and 5 illustrate the optimum sample sizes of independent samples required to be drawn from the population for the two proposed devices. Subsequently, the resulting variances of the estimated population proportions given a fixed survey cost are also displayed. For both the devices, it is observed that for increase in survey costs, gain in efficiency of estimates is achieved with increasing sample sizes.

Table 4: Population variance for fixed survey cost in ORR device with options for RR and ICT using two independent samples (Device-II)

Proportion of individuals with a chosen response option			C **(Rs.)**	C_0 **(Rs.)**	C' **(Rs.)**	n_1	n_2	$V(e)$
DR	**RR**	**ICT**						
0.1	0.2	0.7	300	12	22	6	7	4.65
			600	23	29	9	11	3.36
			900	32	35	12	13	2.87
			1200	47	43	13	14	2.72
			1900	73	64	13	15	2.61
0.2	0.4	0.4	300	12	22	6	7	4.30
			600	23	29	9	10	3.09
			900	32	35	12	13	2.63
			1200	47	43	13	14	2.49
			1900	73	64	14	15	2.39
0.2	0.3	0.5	300	12	22	6	7	4.25
			600	23	29	9	10	3.06
			900	32	35	12	13	2.61
			1200	47	43	13	14	2.47
			1900	73	64	14	15	2.37
0.2	0.6	0.2	300	12	22	6	7	4.52
			600	23	29	9	10	3.25
			900	32	35	12	13	2.76
			1200	47	43	13	14	2.62
			1900	73	64	14	15	2.50
0.4	0.2	0.4	300	12	22	6	7	3.42
			600	23	29	9	10	2.45
			900	32	35	12	13	2.09
			1200	47	43	13	14	1.97
			1900	73	64	14	15	1.89

0.6	0.2	0.2	300	12	22	6	7	2.38
			600	23	29	10	10	1.70
			900	32	35	12	13	1.44
			1200	47	43	13	14	1.36
			1900	73	64	14	15	1.30

Table 5: Population variance for fixed survey cost in ORR device with options for DR, RR and ICT using three independent samples (Device-III)

Proportion of individuals with a chosen response option			C (Rs.)	C_0 (Rs.)	C' (Rs.)	n_1	n_2	n_3	$V(e')$
DR	RR	ICT							
0.1	0.2	0.7	300	12	22	4	6	3	8.87
			600	23	29	6	9	5	6.22
			900	32	35	8	11	6	5.21
			1200	47	43	8	12	6	4.90
			1900	73	64	9	13	7	4.67
0.2	0.4	0.4	300	12	22	4	6	3	7.37
			600	23	29	6	9	5	5.09
			900	32	35	7	11	6	4.23
			1200	47	43	8	12	7	3.96
			1900	73	64	9	13	7	3.77
0.2	0.3	0.5	300	12	22	4	6	3	7.60
			600	23	29	6	9	5	5.27
			900	32	35	8	11	6	4.39
			1200	47	43	8	12	7	4.12
			1900	73	64	9	13	7	3.92
0.2	0.6	0.2	300	12	22	4	6	3	7.00
			600	23	29	6	9	5	4.85
			900	32	35	7	11	7	4.03
			1200	47	43	8	12	7	3.78
			1900	73	64	8	13	8	3.59
0.4	0.2	0.4	300	12	22	4	6	3	6.02
			600	23	29	6	9	5	4.15
			900	32	35	8	11	6	3.44
			1200	47	43	8	12	7	3.23
			1900	73	64	9	13	7	3.06
0.6	0.2	0.2	300	12	22	4	6	3	3.41
			600	23	29	6	9	5	2.32
			900	32	35	8	11	6	1.91
			1200	47	43	8	12	7	1.78
			1900	73	64	9	13	7	1.69

6. Conclusion

To estimate a proportion of individuals bearing a sensitive characteristic in the population, ORR devices present in the literature provide only two types of response options to the survey participants. It is anticipated that in reality, the population is heterogeneous enough to contain individuals out of which a few may opt DR, a few may choose RR and the rest may opt for answering an ICT questionnaire. In such a case, using any ORR device existing in the literature which provides only two response options (DR and RR or RR and ICT), would result in plausible non-responses. Hence, to avoid this issue, two ORR devices are proposed here, both of which provide all the three response options (DR, RR and ICT) to each sampled individual who may choose any one option at his/her discretion without disclosing the choice to the investigator. The first proposed device requires selection of two independent samples from the population and the second device mandates selection of three independent samples. Based on a simulation exercise with different scenarios of respondents' choices for DR, RR and ICT, it is concluded that both the proposed devices are competitive to the existing ORR device.

Acknowledgements

The authors are thankful to the referee and the editor for providing useful comments which have helped in improving the manuscript.

References:

Arnab, R. (2004). Optional randomized response techniques for complex designs. *Biometrical Journal*, **46**, 114-124.

Arnab, R., and Rueda, M. (2016). Optional Randomized Response: A Critical Review. *Handbook of Statistics*, **34**. On *Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, 253-271.

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton: Chapman and Hall, CRC Press, Taylor & Francis Group (ISBN: 9781138115422).

Chaudhuri, A., and Christofides, T. C. (2007). Item count technique in estimating the proportion of people with a sensitive feature. *Journal of Statistical Planning and Inference*, **137**, 589-593.

Chaudhuri, A., and Christofides, T. C. (2013). *Indirect Questioning in Sample Surveys*. Springer.

Chaudhuri, A., and Dihidar, K. (2009). Estimating means of stigmatizing qualitative and quantitative variables from discretionary responses randomized or direct. *Sankhya*, **B71**, 123-136.

Chaudhuri, A., and Pal, S. (2002). On certain alternative mean square error estimators in complex survey sampling. *Journal of Statistical Planning and Inference*, **104**, 363-375.

Chaudhuri, A., and Saha, A. (2005a). On relative efficiencies of optional versus compulsory randomization in responses: A simulation-based numerical study covering three RR schemes. *Pakistan Journal of Statistics*, **21**(**1**), 87-98.

Chaudhuri, A., and Saha, A. (2005b). Optional versus compulsory randomized response techniques in complex surveys. *Journal of Statistical Planning and Inference*, **135**, 516-527.

Hartley, H. O., and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, **33**(**2**), 350-374.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **77**, 89-96.

Mehta, S., Dass, B. K., Shabbir, J., and Gupta, S. N. (2012). A three-stage optional randomized response model. *Journal of Statistical Theory and Practice*, **6**(**3**), 417–427.

Miller, J. D. (1984). *A New Survey Technique for Studying Deviant Behaviour*. Ph.D. Thesis, The George Washington University.

Miller, J. D., Cisin, I. H., and Harrel, A. V. (1986). A new technique for surveying deviant behavior: item count estimates of marijuana, cocaine and heroin. Paper presented at the annual meeting of the *American Association for Public Opinion Research, St. Petersburg, Florida*.

Pal, S. (2007). Estimating the proportion of people bearing a sensitive issue with an option to item count lists and randomized response. *Statistics in Transition*, **8**(**2**), 301-310.

Pal, S. (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. *Metrika*, **49**, 157-164.

Raghavarao, D., and Federer, W. F. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society*, **B41**, 40-45.

Shaw, P. (2016). Estimating a finite population proportion bearing a sensitive attribute from a single probability sample by Item Count Technique. *Handbook of Statistics*, **34**. On *Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, 387-403.

Sihm, J. S., and Gupta, S. (2015). A two-stage binary optional response model. *Communications in Statistics – Simulation and Computation*, **44**(**9**), 2278–2296.

Warner, S. L. (1965). RR: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 181–198

Irrigation Practices and Soft Computing Applications: A Review

Pradeep H.K.[1], Jasma Balasangameshwara[2], M.S. Sheshshayee[3], K. Rajan[4] and Prabhudev Jagadeesh[5]

[1]Department of Computer Science & Engineering, JSS Academy of Technical Education, Bengaluru & Research Scholar, Visvesvaraya Technological University – RRC, Belgaum, Karnataka, India.

[2]Department of Computer Science & Technology, Dayananda Sagar University, Bengaluru, Karnataka, India.

[3]Department of Crop Physiology, University of Agricultural Sciences, Bengaluru, Karnataka, India.

[4]ICAR - Indian Institute of Soil and Water Conservation, Ooty, India.

[5]Department of Computer Science & Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India.

Received: 20 June 2020; Revised: 10 December 2020; Accepted: 15 December 2020

Abstract

Repeated droughts, population expansion and global warming force thorough limitations on irrigation practices. The low water usage efficiency is the universal problem encountered by most of the irrigation systems. A survey was carried out over irrigation practices, which comprises of conventional irrigation methods, micro-irrigation systems, intelligent irrigation approaches, estimation of reference evapotranspiration (ET_0) using soft computing models and performance indicator models. The outcome of the survey reveals that, the software techniques must be integrated with traditional irrigation practices to improve water productivity and economy.

Key words: Irrigation methods; Land suitability; Machine learning; Performance indicators.

1. Introduction

Economic progress and expanding worldwide populace extend the interest for innovative irrigation system. According to the expectation of food and agriculture organization (FAO), food necessity will increment about 60% by year 2050 (Alexandratos and Bruinsma, 2012). Internationally, evaluated that agriculture action devours around 70% of the gross water, grouped with 10% for civic use and remaining water is used by mechanical sector (Provenzano and Sinobas, 2014). Worldwide, inundated land represents 302Mha and possesses just 16% of the cultivatable region (Playan *et al.*, 2013). Presently, 36% of land by bone-dry and semi-parched locales and anticipated that drought risk will further increment (Safriel *et al*., 2006; Alcamo *et al*., 2007; Arnell *et al*., 2011). The water productivity (WP) is the proportion between crop yield and complete water use (Pereira *et al*., 2002). The water devoured by plants is under 65% of provided water and right volume of plants upon right time improves the WP (Chartzoulakis *et al*., 2015). The design of effective irrigation system is complex because of barometrical conditions, soil properties, crop species and irrigation strategy (Dabach *et al*., 2013; Soulis and Elmaloglou, 2018). The generally

Corresponding Author: Pradeep H.K.
Email: phk.contact@gmail.com

utilized irrigation system strategies are surface, pressurized sprinkler, low volume drip and micro-sprinkler. The subsurface irrigation is another water system wherein water is applied straightforwardly inside the soil (Orang *et al*., 2008). The deficit irrigation method was an efficient strategy for Mediterranean environment land considering drought tolerant crop (Galindo *et al*., 2018; Hargreaves and Samani, 1984). The surface irrigation strategy is most widely utilized method and this methodology is generally popular and prudent but the low water system proficiency is the key issue (Raghuwanshi *et al*., 2010). The sprinkler water system structure includes pipe network water streams with power through spouts and it mimics precipitation with of overhead splashing (Valipour, 2015). In trickle water system, water is provided through fixed model line organization and gradually discharged to plants (Tindula *et al*., 2013). The advancement of first generation water system innovation was begun with multi-customer electronic hydrants for usage at regulation organization. The second era water system innovation was variable recurrence siphons. The micro-irrigation system strategy was the third era in irrigation innovation wherein WP was expanded however hardly introduced because of high initial speculation. The sub surface trickle water system was the fourth era in irrigation innovation designed to address the difficulties of surface drip water system, wherein producer obstructing issue is killed. The fifth era in water system innovation was deficiency water system developed for ideal water application considering crop development stage without influencing the yield (Levidow *et al*., 2014; kang *et al*., 2017). Artificial intelligence (AI) based water system frameworks are likely ways to deal with affordable and effective models for agricultural water management (Torres-Rua *et al*., 2012; Niu *et al*., 2017; Chlingaryan *et al*., 2018; Behmann *et al*., 2015; Griffiths *et al*., 2011; Gutierrez *et al*., 2018; Haider *et al*., 2008; Kamilaris and Prenafeta-Bouldu, 2018).

2. Land Suitability for Different Irrigation Methods

The land suitability for surface and micro-irrigation system was dissected utilizing parametric assessment strategy to decide the possible technique. The dirt properties were utilized to decide the reasonable water system technique in Fakkeh area of West Iran. The investigation displayed that trickle water system technique improved land sufficiency over sprinkler and surface strategy. The dirt surface was restricting variable for surface and sprinkler strategy, calcium carbonate was central question for drip irrigation system (Landi *et al*., 2008). The dirt properties were utilized to decide the appropriate water system techniques in Abbas plain territory of West Iran. The dirt properties were utilized to decide the appropriate water system strategy in Dosalegh locale of Iran. The investigation displayed that drip water system technique improved land sufficiency over sprinkler and surface strategy. The dirt surface, saltiness and incline were restricting components for surface and sprinkler strategy, calcium carbonate, soil surface and saltiness were key restricting variable for drip water system (Albaji *et al*., 2010). The dirt properties were utilized to decide the appropriate water system strategies in Gotvand plain zone of Iran. The investigation showed that sprinkler water system strategy improved land sufficiency over trickle and surface technique. The calcium carbonate and seepage were restricting variables for all water system strategies (Albaji *et al*., 2014). The dirt properties were utilized to decide the appropriate water system strategy in Rasht area of Iran. The investigation showed that trickle water system strategy improved land ampleness over sprinkler and surface technique. The dirt surface and seepage were key restricting variables for all the water system techniques (Seyedmohammadi *et al*., 2016). The audit of soil properties and land appropriateness model shows that micro-irrigation system surpasses surface water system over expanding irrigation land inside the accessible water resources.

3. Irrigation Methods

The irrigation method adoption depends on soil and land characteristics, WP and Economic water productivity (EWP). In the following section the basin, tube sprinkler, pillow and drip irrigation strategies were compared over investment, electricity cost, water usage efficiency and crop yield. The furrow and deficit drip strategies were compared on water savings and yield. The surface drip and sub surface drip were compared over emitter clogging, water consumption and yield. The drip and sprinkler methods were analyzed over delivery efficiency.

3.1. Comparison of basin, pillow, drip, and tube sprinkler irrigation

To address water scarcity, a field study was carried out at North China Plain, the three micro-irrigation methods improved WP but EWP of basin irrigation method was higher compared to micro irrigation methods. The comparisons of drip, basin, pillow and tube sprinkler irrigation methods are represented in Table 1 and Figure 1.

Table 1: Comparison of basin, tube sprinkler pillow and drip irrigation methods

References	Irrigation method	Investment cost (Yuan/ha)	Electricity cost (Yuan/ha)	Irrigation depth applied (mm)	WP (kg/m^3)	Yield (kg/ha)	Crop species
Fang *et al*., (2018)	Basin	700	0.22	90	1.57	6217.5	Winter wheat
	Drip	4125	0.33	90	1.91	6937.8	
	Pillow	3225	0.35	90	1.73	6898.3	
	Tube Sprinkler	4443	0.26	90	1.63	6614.5	

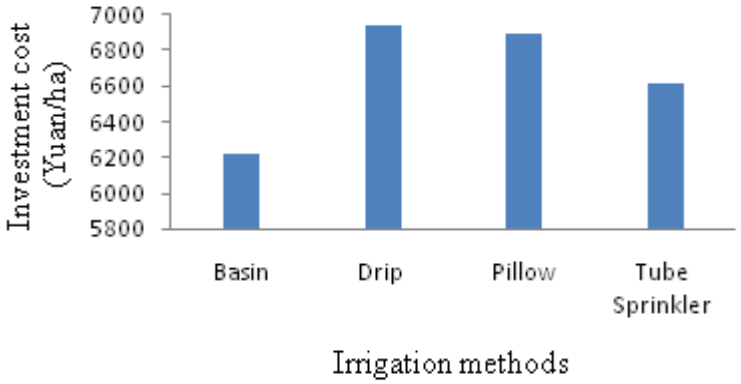


Figure 1: Comparison of different irrigation over investment cost

3.2. Comparison of furrow and drip irrigation

In a semiarid region a field study was conducted on drip and furrow irrigation for sugar beet to analyze WP. The drip tape irrigation method surpasses furrow method on sugar beet with higher WP. The details of water savings and yield are represented in Table 2 and Figure 2.

Table 2: Comparison of furrow and drip irrigation on WP

References	Irrigation method	Water savings compare to furrow irrigation		Average yield (t/ha)	Crop species
		with monitoring	without monitoring		
Ghamarnia *et al.*, (2011)	Drip (100% ET)	28.8%	38.4%	15.55	Sugar beets
	Drip (75% ET)	46.6%	53.8%	14.62	
	Drip (50% ET)	64.5%	69.2%	11.78	
	Drip (25% ET)	82.2%	84.6%	9.36	

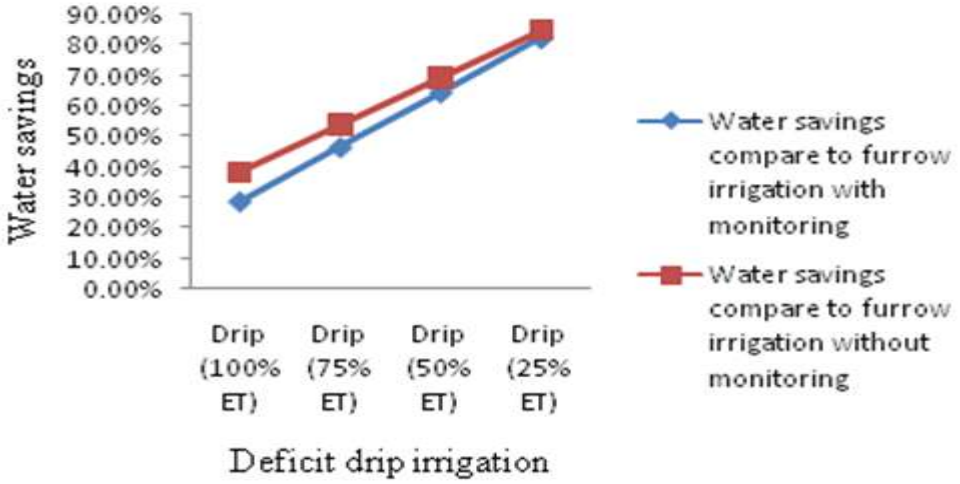


Figure 2: Comparison of furrow and drip irrigation over WP

3.3. Comparison of surface drip (DI) and subsurface drip irrigation (SDI)

Irrigation efficiency is an important issue in semiarid region due to water scarcity. Detecting leakages and repairing them is difficult task in surface drip irrigation though it is very efficient method. To overcome the drawbacks mentioned above an alternative subsurface irrigation system was introduced in southern Spain. The subsurface drip irrigation WP was high comparing to traditional drip irrigation method and easy to install. Comparison of DI and SDI based on WP are outlined in Table 3.

Table 3: Comparison of DI and SDI based on water WP and yield

References	Irrigation method	WP (kg/m^3)	Average yield (kg/tree)	Emitter Clogging issue	Crop species
Martinez and Reca, (2014)	DI	0.22	17.15	More exposure to emitter clogging and difficulty to detect clogged emitters and leakages.	Organic olive orchard
	SDI	0.24	19.24	Reduced exposure to emitter clogging and also easy to detect and replace clogged emitters.	

3.4. Analysis of sprinkler and drip irrigation

The drip and sprinkler irrigation strategies were compared on delivery efficiency (DE), maintenance cost and economy. The WP in drip irrigation system was lower than sprinkler irrigation system, in most of the plots water supply was higher than the actual requirement of water by crops. According to water users associations the sprinkler irrigation system has higher EWP than drip irrigation system (Corcoles *et al*., 2011). The comparison of sprinkler and drip irrigation performance are summarized in Table 4.

Table 4: Comparison of drip and sprinkler irrigation on economy and efficiency

References	Irrigation method	DE (%)	MOMId (€/m^3)	Energy Cost	OIa (€/ha)	Crop species
Corcoles *et al*., (2011)	Sprinkler	92.7	0.05	45% of MOM	4,408.16	Maize, Barley, Alfalfa, Onion, Carrot, Vineyard
	Drip	80	0.13	20% of MOM	2,388.16	Vineyards, Olive trees, Almond trees

MOMId = Management, Operation and Maintenance cost per unit irrigation delivery, OIa = Economic output per unit irrigation area.

4. Soft Computing (SC) Techniques for Irrigation System

SC is a space of software engineering that emulates marvel of human mind (Gocic *et al*., 2015). The perspectives, for example, cognizance and perception are key highlights of SC strategies. The SC techniques abuse obstruction for vulnerability and imprecision and also guarantee similarity and offers prudent arrangements. (Keskin and Terzi, 2006). To assemble smart and reasonable machines SC strategies have been utilized in numerous applications including ET_0. The ET_0 is a significant measurement to comprehend the harvest water prerequisites to acquire good yield (Temesgen *et al*., 2005). The ET_0 is crucial parameter for estimation of irrigation water requirements (Allen *et al*., 1991).

4.1. Neural networks (NN) for irrigation system

NN is an anatomical organization utilized for modelling non-linear systems using artificial intelligence methods. The NN data preparing structure is made like human neural organization and it comprises of three fundamental components, for example, input, concealed layers and yield. Shrouded layers among info and yield have number of neurons, hubs or cells. Information signal from the info layer arrives at the following connection by following all conceivable association ways and at each connection signal goes through change. NN comprises of many handling components arranged by connections and loads since its gigantic equal framework (Keskin and Terzi, 2006). The NN can gauge the cycle conduct even with halfway data. To gauge ET_0 neural organization models were utilized with various methodologies. In this section different neural organization strategies utilized for forecast of ET_0 are described.

The Artificial NN (ANN) and NN integrated with auto regressive external input (NNARX) models performance were analyzed in hot and dry environment (Piri *et al*., 2009). Multiple regression (MLR) and NN model efficiency was analyzed considering humidity and temperature data (Laaboudi *et al.*, 2012). Adaptive neuro-fuzzy inference system (ANFIS) model was analyzed for climate data of Kerman and Isfahan station (Karimaldini *et al*.,

2011). The ANN and Evolutionary NN (ENN) models were analyzed for forecast of ET_0. The feed forward back propagation NN (FFBP-NN) and second order NN (SONN) models were investigated for forecast of ET_0 (Adamala *et al*., 2013). Cuckoo search algorithm (CSA) was integrated with NN (ANN+CSA) and ANFIS was integrated with CSA (ANFIS+CSA) for forecast of ET_0 over twelve stations climate data of Serbia (Shamshirband *et al*., 2015). Back propagation neural networks (BPNN) was applied to forecast ET_0 with the help of hybrid particle based back propagation (PF-BP), Imperialist competition algorithm (ICA-BP) was used for forecast of ET_0 over Tabriz weather station data (Nazari and Shamshirband, 2018). Regression technique was applied for ET_0 prediction (Khoshravesh *et al*., 2017). The survey reveals that PF-BP and ENN model surpasses the different NN methods for forecast of ET_0.

4.2. Support vector machines (SVM) for irrigation system

SVM is a measurable learning hypothesis created by Vapnik. The informational collections of non-linearly distinct can be grouped by SVM utilizing kernels for plotting the information into high-dimensional component space. Support vector regression (SVR) is a way to deal with decide relapse through SVM. The fitting choice of bits and its boundaries portrays the performance of SVR model. Radial basis function (RBF) is the kernel function for SVM due to its favourable performance (Deo and Samui, 2017). Least square support vector machine (LSSVM) approach was applied to forecast ET_0 considering weather data from Shihez station of China and the prediction of LSSVM method was compared with ANN (Chen, 2011). The SVR approach was applied for forecast of ET_0 using regression procedures with SVM. The SVR model outperformed the other variants of SVM (Kisi and Cimen, 2009).

4.3. Genetic programming (GP) for irrigation system

The GP model discovers solution for issues utilizing traverse and change rules. Genetic calculation upholds equal inquiry dependent on Darwin development hypothesis. GP has self boundary choice potential to draw the features for improving the model without client impedance and it describes the program linearly. Genetic algorithm and back propagation (GABP) NN approach was applied to estimate ET_0 considering weather data of Tabriz station, Iran (Nazari and Shamshirband, 2018). A linear GP (LGP) was applied to forecast plant water requirement (Kisi and Guven, 2010). Gene expression programming (GEP) approach was applied to forecast plant water requirement using Egypt weather data (Mattar and Alazba, 2018). The LGP surpasses other GP variants for forecast of ET_0.

5. Intelligent Irrigation Systems

Approximately 60% of the flooded land must be smoothed out by adopting innovative irrigation methods to satisfy future global food demand and to extend WP (Alexandratos and Bruinsma, 2012; Playan *et al*., 2013). The SC strategies, agent technology, wireless Sensor Networks (WSN), Fuzzy decision support system (FDSS), Internet of Things (IoT) and have great potential to extend water savings in irrigation management. The review of innovative irrigation system exhibits the key features which help to improve the performance of irrigation system. The Fuzzy decision support system (FDSS) for irrigation was planned to address the particular issues of online water system model called IRRINET (Giusti and Marsili-Libelli, 2015). Agent based irrigation was planned considering soil properties, crop thirst affectability, development stage and net return estimation of harvest yield. The day by day water revive model was planned thinking about precipitation, ET_0, and introductory profundity of field water. The specialist model increases WP without yield reduction using

regulated deficit irrigation. The depth of water required for daily recharge to maintain soil water balance was decided using volume of soil moisture depleted. The experiment was conducted for multi-crop farm land using priority based irrigation scheduling, which exhibits increased water productivity (Anthony and Birendra, 2018). To optimize water for agricultural crops automated irrigation system was developed. An intelligent irrigation system was designed using WSN, which comprises of temperature and dampness sensors inserted in the root zone of the yields, detected and handled information moved to a web machine. Based on temperature and soil moisture data for real time monitoring and programming of irrigation graphical user interface software was implemented (Gutierrez *et al*., 2014). The drip irrigation scheduling was implemented using java application software tool called IRRIX. The water balance model was employed for forecast of plant water requirement and recharge strategy was applied to balance the soil water, based on the feedback data of soil and plant sensor. Experiments were conducted for automated full and deficit irrigation with conventional method. Automated irrigation surpasses the conventional method through increase in WP and economy (Casadesus *et al*., 2012). Multi-intelligent control system (MICS) was used with the help of IoT for irrigation management. MICS provides reliable and satisfactory solution and also increases WP and EWP over conventional irrigation system (Hadipour *et al*., 2020). A smart irrigation system was proposed using IoT and neural networks approach. Crop water requirement data set was used to train the neural networks algorithm to get the accurate results. Intelligent irrigation was compared with normal drip and conventional irrigation methods, where in intelligent irrigation model surpasses the conventional methods through increased water productivity (Nawandar and Satpute, 2019). Automated drip irrigation was proposed using smart phone and microcontroller for paddy crop. It was compared with flood and normal drip irrigation. The smart phone captures the soil image, estimates the moisture and passes the data onto the microcontroller using GSM module. Automated drip out performs the normal drip and flood irrigation system (Barkunan *et al*., 2019).

5.1. Irrigation scheduling based on crop water stress

Intelligent root zone water quality model based irrigation was used to predict crop water pressure progressively. The depth of water needed for day by day revive to deal with soil water balance was set considering the depth of soil dampness drained. The yield water pressure based water system was adjusted with field water system utilizing drip and sprinkler technique for corn and soybean crops individually. The model expands the water system proficiency in low precipitation territory and it burns-through somewhat more water in moist territories with expanded harvest yield (Gu *et al*., 2017). The software model anticipated irrigation was calibrated with field drip irrigation, which is highlighted in Table 5.

Table 5: Comparison of software model based irrigation over field drip irrigation

References	Software model based irrigation	Water savings when calibrated with field drip irrigation for 3 years			Crop	Crop yield
		2008	2009	2010		
Gu *et al*., (2017)	Simulated for full water supply	30.5%	17.3%	7.1%	Corn	Negligible decrease between 0.03-3.81%
	Simulated for 60-90% of full water supply	35%	30%	16%		

The crucial input parameters are identified in the survey considering various irrigation systems and which can be used as features for machine learning based irrigation system. Comparison of machine learning, IoT, cloud and agent based irrigation systems over water savings are outlined in Table 6. The vital input features required for efficient irrigation systems are outlined in Table 7.

Table 6: Comparison of various software based irrigation systems on water savings

References	Technology	Water savings	Irrigation method	Crop species	Additional benefits	Experiment duration
Anthony and Birendra, (2018)	Agent technology	22.11% Without affecting the crop production .	Not mentioned	Pastures Maize Tomato Potato	High profit with priority-based water allocation	Not mentioned
Gutierrez *et al*., (2014)	Wireless sensor networks	60%	Drip	Sage Thyme Origanum Basil	Energy autonomy And Low cost	18 Months
Giusti and Marsili-Libelli, (2015)	Fuzzy logic	13.55 % compare to irrinet model	Not mentioned	Corn Kiwi Potato Vegetable and Fruit crops	Robust and Consistent	2006-08
Gu *et al*., (2017)	RZWQM2	35%	Drip, Sprinkler	Corn Soybean	Crop production improvement of 291 kg/ hectare	2008-10
Niu *et al*., (2017)	Machine learning	Not mentioned	Not mentioned	Reeds Typha Orientalis Paddy	High Accuracy	Not mentioned
Severino *et al*., (2018)	Internet of things (IoT)	Not mentioned	Drip	Not Mentioned	Usage of recycled water	Not mentioned
Zhou and Li, (2017)	Cloud services	Not mentioned	Not mentioned	Not Mentioned	Great market prospect	Not mentioned

Table 7: Key features identified for efficient intelligent irrigation system

References	SM	HU	ST	IM	CS	CG	CD	ET	RF	DP	RO
Anthony and Birendra, (2018)	✔		✔		✔	✔	✔	✔	✔		
Gutierrez *et al*., (2014)	✔	✔		✔	✔						
Giusti and Marsili-Libelli, (2015)	✔	✔	✔	✔	✔	✔		✔	✔		
Gu *et al*., (2017)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Niu *et al*., (2017)	✔	✔			✔	✔		✔	✔		
Severino *et al*., (2018)	✔		✔	✔	✔	✔	✔	✔	✔		
Zhou and Li, (2017)	✔	✔			✔	✔		✔			

SM= Soil moisture, HU= Humidity, ST= Soil type, IM= Irrigation method, CS= Crop species, CG=Crop growth stage, CD= Crop drought sensitivity, RF= Rain fall, DP= Deep percolation, RO= Runoff.

6. Performance Indicators for Irrigation System

The performance indicators play a vital role in rating irrigation systems (Pereira *et al*., 2012). In this section the key terminologies used for analyzing irrigation system performance are outlined. The ET determines the plant water requirement but how efficiently the irrigation system satisfies the need is computed through application efficiency (A_e). The AE is defined as the ratio of average depth of irrigation water consumed by crops and average depth of irrigation water applied. The aim of irrigation system is every part of the field should receive same amount of water. The distribution uniformity is defined as the ratio of average points of smallest water depth accumulated and average depth of water stored in all points. The irrigation efficiency (I_e) is the ratio of beneficially used irrigation water and gross volume of irrigation water that leaves the boundary. The irrigation consumptive use coefficient (I_{cu}) is defined as the ratio of depth of irrigation water consumptively used and gross volume of irrigation water that leaves the boundary. The irrigation sagacity (I_s) is the ratio that covers water usage for societal purpose along with crops consumption and gross volume of irrigation water that leaves the boundary. The other performance indicators such as adequacy (A_q), equity of water distribution (E_q), dependability of water supply (D_p), net Returns (N_r), yield Response, deep percolation ratio (D_r), tail water ratio (T_r), yearly relative water supply (Y_{rw}), yearly relative irrigation supply (Y_{ri}), Transmission loss (T_l), Outcome per planted area (O_{pa}), outcome per unit irrigated area (O_{ui}), outcome per unit irrigation applied (O_{ia}), outcome per unit irrigation depth consumed (O_{ic}), relative water supply (R_w), relative irrigation supply (R_i), irrigation water delivery capability (I_{dc}), dependability of duration (D_d), annual income (A_i), annual profit (A_p), net irrigation requirement (N_{ir}), net regulated deficit irrigation (N_{rdi}), seasonal irrigation performance index (S_{ipi}) are outlined in the following section. The survey of irrigation performance indicator model exhibited that, the water productivity and economic water productivity models are the effective measures to understand water savings and economy (Pereira *et al*., 2012). The irrigation performance indicator model to measure application efficiency is outlined in Table 8. The Irrigation performance indicator model to measure distribution uniformity (low quarter) is outlined in Table 9. The Irrigation performance indicators considering crop transpiration, evaporation, yield and profit are outlined in Table 10 (Appendix).

Table 8: Application efficiency models used in irrigation system

References	Model	Variables considered
Burt *et al*., (1997)	$A_e = \frac{A_t}{A_a} \times 100$	A_e: Application Efficiency A_t: Average depth of irrigation water providing to target A_a: Average depth of Irrigation water applied
Ghamarnia *et al*., (2011)	$A_e = \frac{I_a + I_c}{I_s}$	A_e: Application Efficiency I_a: Irrigation depth accumulated upon root zone (m^3) I_c: Irrigation depth consumed on the root zone (m^3) I_s: Total Irrigation depth supplied (m^3)
Raghuwanshi *et al*., (2010)	$A_e = \frac{I_a}{q_0 W T_e} \times 100$	A_e: Application Efficiency I_a: Depth of irrigation water accumulated upon root zone (m^3) q_0: Flow in rate per unit border extent (m3/ m/s) W: Border extent (m) T_e: End time (s)
Reca *et al*., (2018)	$A_e = 1 + f\left(\frac{D_r}{D_g} - 1\right) - \frac{\left(C_v - \frac{v^2}{2}\right)}{\left(\sqrt{\frac{\pi}{2}}\right)}$	A_e: Application Efficiency f: Fragment of the command area unit that is adequately irrigated. D_r : Irrigation depth requirement D_g : Total irrigation depth C_v: Coefficient variation of irrigation depth applied v: Cumulative variable

Table 9: Distribution uniformity low quarter (DU_{lq}) models used in irrigation system

References	Model	Variables considered
Burt *et al*., (1997)	$DU_{lq} = \frac{AD_{lq}}{AD_{ef}}$	AD_{lq} : Average depth of irrigation water accrued in low quarter field AD_{ef} : Average depth of irrigation water accrued in entire field elements
Raghuwanshi *et al*., (2010)	$DU_{lq} = \frac{\overline{AP_{lq}}}{\overline{AP}}$	$\overline{AP_{lq}}$: Average percolated depth for low field quarter (mm) $\overline{AP}$: Average percolated depth (mm)

7. Conclusion

Irrigation practices and software techniques applied for agricultural water management was reviewed to determine the effective method considering water productivity and economy. This paper reveals that, software techniques should be integrated with traditional irrigation methods to offer economical and efficient irrigation system. The empirical irrigation

strategies were analyzed for water productivity and economy. This paper exhibits that, suppose if economy is the decision making factor, then surface irrigation is the best method over expensive micro-irrigation. Suppose if water savings is the key objective, then micro-irrigation technique is the best approach over surface irrigation system. The review of intelligent irrigation systems exhibits that, the software model based crop stress irrigation was the most effective technique with 30.5% water savings compared to field drip irrigation and this paper also reveals that software based irrigation system significantly improves water productivity. The soft computing model based forecast of reference evapotranspiration approach outperforms conventional models with minimal number of input features.

The survey opens-up future research on machine learning based surface irrigation system, which offers efficient and economical agricultural water management system. The machine learning based irrigation framework safeguards the advantage of low initial venture of conventional surface irrigation system with higher water productivity through the aid of artificial intelligence techniques. Real-time irrigation framework based on machine learning technique makes a significant improvement in water productivity.

Acknowledgments

The authors express their gratefulness to the anonymous reviewer for giving some useful suggestions that led to considerable improvement in the presentation of the results. The authors also express their thanks to the Chief Editor for his suggestions on restructuring the contents for better expression.

References

Adamala, S., Raghuwanshi, N. S., Mishra, A. and Tiwari, M. K. (2013). Evapotranspiration modeling using second-order neural networks. *Journal of Hydrologic Engineering*, **19**(**6**), 1131-1140.

Albaji, M., Boroomand-Nasab, S., Naseri, A. and Jafari, S. (2010). Comparison of different irrigation methods based on the parametric evaluation approach in Abbas plain: Iran. *Journal of Irrigation and Drainage Engineering*, **136**(**2**), 131-136.

Albaji, M., Golabi, M., Nasab, S. B. and Jahanshahi, M. (2014). Land suitability evaluation for surface, sprinkler and drip irrigation systems. *Transactions of the Royal Society of South Africa*, **69**(**2**), 63-73.

Alcamo, J., Florke, M. and Marker, M. (2007). Future long-term changes in global water resources driven by socio-economic and climatic changes. *Hydrological Sciences Journal*, **52**(**2**), 247-275.

Alexandratos, N. and Bruinsma, J. (2012). *World Agriculture Towards* 2030/2050:*The*2012*Revision,* **12(3)**. FAO, Rome: ESA working paper.

Allen, R. G., and Pruitt, W. O. (1991). FAO-24 reference evapotranspiration factors. *Journal of Irrigation and Drainage Engineering*, **117**(**5**), 758-773.

Anthony, P., and Birendra, K. C. (2018). Improving irrigation water management using agent technology. *New Zealand Journal of Agricultural Research*, **61**(**4**), 425-439.

Arnell, N. W., van Vuuren, D. P. and Isaac, M. (2011). The implications of climate policy for the impacts of climate change on global water resources. *Global Environmental Change*, **21**(**2**), 592-603.

Arunkumar, R., and Ambujam, N. K. (2010). Performance assessment of canal irrigation system. *ISH Journal of Hydraulic Engineering*, **16**(**1**), 146-155.

Barkunan, S. R., Bhanumathi, V., and Sethuram, J. (2019). Smart sensor for automatic drip irrigation system for paddy cultivation. *Computers & Electrical Engineering*, **73,** 180–193.

Behmann, J., Mahlein, A. K., Rumpf, T., Romer, C., and Plumer, L. (2015). A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, **16**(**3**), 239-260.

Broner, I. and Lambert, J. (1989). Optimal scheduling of irrigation machines. I: Model development. *Journal of irrigation and drainage engineering*, **115**(**5**), 862-879.

Burt, C. M., Clemmens, A. J., Strelkoff, T. S., Solomon, K. H., Bliesner, R. D., Hardy, L. A., Howell, T. A. and Eisenhauer, D. E. (1997). Irrigation performance measures: efficiency and uniformity. *Journal of irrigation and drainage engineering*, **123**(**6**), 423-442.

Casadesus, J., Mata, M., Marsal, J., & Girona, J. (2012). A general algorithm for automated scheduling of drip irrigation in tree crops. *Computers and Electronics in Agriculture*, **83,** 11-20.

Cetin, O., and Kara, A. (2019). Assesment of water productivity us- ing different drip irrigation systems for cotton. *Agricultural Water Management*, **223,** 105693–105693.

Chartzoulakis, K., and Bertaki, M. (2015). Sustainable water management in agriculture under climate change. *Agriculture and Agricultural Science Procedia*, **4**, 88-98.

Chen, D. (2011), October. Daily reference evapotranspiration estimation based on least squares support vector machines. In *International Conference on Computer and Computing Technologies in Agriculture*, 54-63. Springer, Berlin, Heidelberg.

Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, **151**, 61-69.

Corcoles, J. I., de Juan, J. A., Ortega, J. F., Tarjuelo, J. M. and Moreno, M. A. (2011). Evaluation of irrigation systems by using benchmarking techniques. *Journal of Irrigation and Drainage Engineering*, **138**(**3**), 225-234.

Dabach, S., Lazarovitch, N., Simunek, J., and Shani, U. (2013). Numerical investigation of irrigation scheduling based on soil water status. *Irrigation Science*, **31**(**1**), 27-36.

Deo, R. C., and Samui, P. (2017). Forecasting evaporative loss by least-square support-vector regression and evaluation with genetic programming, Gaussian process, and minimax probability machine regression: case study of Brisbane City. *Journal of Hydrologic Engineering*, **22**(**6**), 05017003.

Fang, Q., Zhang, X., Shao, L., Chen, S. and Sun, H. (2018). Assessing the performance of different irrigation systems on winter wheat under limited water supply. *Agricultural Water Management*, **196**, 133-143.

Galindo, A., Collado-Gonzalez, J., Grinan, I., Corell, M., Centeno, A., Martin-Palomo, M. J., Giron, I. F., Rodriguez, P., Cruz, Z. N., Memmi, H., and Carbonell-Barrachina, A.A. (2018). Deficit irrigation and emerging fruit crops as a strategy to save water in Mediterranean semiarid agrosystems. *Agricultural water management*, **202**, 311-324.

Ghamarnia, H., Arji, I., Sepehri, S., Norozpour, S., and Khodaei, E. (2011). Evaluation and comparison of drip and conventional irrigation methods on sugar beets in a semiarid region. *Journal of Irrigation and Drainage Engineering*, **138**(**1**), 90-97.

Giusti, E., and Marsili-Libelli, S. (2015). A Fuzzy Decision Support System for irrigation and water conservation in agriculture. *Environmental Modelling & Software*, **63**, 73-86.

Gocic, M., Motamedi, S., Shamshirband, S., Petkovic, D., Ch, S., Hashim, R., and Arif, M. (2015). Soft computing approaches for forecasting reference evapotranspiration. *Computers and Electronics in Agriculture*, **113**, 164-173.

Griffiths, T. L., and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, **12**, 1185-1224.

Gu, Z., Qi, Z., Ma, L., Gui, D., Xu, J., Fang, Q., Yuan, S., and Feng, G. (2017). Development of an irrigation scheduling software based on model predicted crop water stress. *Computers and Electronics in Agriculture*, **143**, 208-221.

Gutierrez, J., Villa-Medina, J. F., Nieto-Garibay, A., and Porta-Gandara, M. A. (2014). Automated irrigation system using a wireless sensor network and GPRS module. *IEEE transactions on instrumentation and measurement*, **63**(**1**), 166-176.

Hadipour, M., Derakhshandeh, J. F., and Shiran, M. A. (2020). An experimental setup of multi-intelligent control system (MICS) of water management using the Internet of Things (IoT). *ISA transactions*, **96,** 309-326.

Haider, M. A., Pakshirajan, K., Singh, A., and Chaudhry, S. (2008). Artificial neural network-genetic algorithm approach to optimize media constituents for enhancing lipase production by a soil microorganism. *Applied biochemistry and biotechnology*, **144**(**3**), 225-235.

Hargreaves, G. H., and Samani, Z. A. (1984). Economic considerations of deficit irrigation. *Journal of irrigation and Drainage engineering*, **110**(**4**), 343-358.

Kamilaris, A., Kartakoullis, A., and Prenafeta-Boldu, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, **143**, 23-37.

Kang, S., Hao, X., Du, T., Tong, L., Su, X., Lu, H., Li, X., Huo, Z., Li, S., and Ding, R. (2017). Improving agricultural water productivity to ensure food security in China under changing environment: From research to practice. *Agricultural Water Management*, **179**, 5-17.

Karimaldini, F., Teang Shui, L., Ahmed Mohamed, T., Abdollahi, M., and Khalili, N. (2011). Daily evapotranspiration modeling from limited weather data by using neuro-fuzzy computing technique. *Journal of Irrigation and Drainage Engineering*, **138**(**1**), 21-34.

Keskin, M. E., and Terzi, O. (2006). Artificial neural network models of daily pan evaporation. *Journal of Hydrologic Engineering*, **11**(**1**), 65-70.

KISI, O., and Cimen, M. (2009). Evapotranspiration modelling using support vector machines/Modélisation de l'évapotranspiration à l'aide de 'support vector machines. *Hydrological sciences journal*, **54**(**5**), 918-928.

Kisi, O., and Guven, A. (2010). Evapotranspiration modeling using linear genetic programming technique. *Journal of Irrigation and Drainage Engineering*, **136**(**10**), 715-723.

Khoshravesh, M., Sefidkouhi, M. A. G., and Valipour, M. (2017). Estimation of reference evapotranspiration using multivariate fractional polynomial, Bayesian regression, and robust regression models in three arid environments. *Applied Water Science*, **7**(**4**), 1911-1922.

Laaboudi, A., Mouhouche, B., and Draoui, B. (2012). Neural network approach to reference evapotranspiration modeling from limited climatic data in arid regions. *International journal of biometeorology*, **56**(**5**), 831-841.

Levidow, L., Zaccaria, D., Maia, R., Vivas, E., Todorovic, M., and Scardigno, A. (2014). Improving water-efficient irrigation: Prospects and difficulties of innovative practices. *Agricultural Water Management*, **146**, 84-94.

Martinez, J., and Reca, J. (2014). WUE of surface drip irrigation versus an alternative subsurface drip irrigation method. *Journal of Irrigation and Drainage Engineering*, **140**(**10**), 04014030.

Mattar, M. A., and Alazba, A. A. (2018). GEP and MLR approaches for the prediction of reference evapotranspiration. *Neural Computing and Applications*, 1-13.

Memon, N. A., Broughton, R. S., Madramootoo, C. A., Prasher, S. O., and Huene, B. V. H. (1986). A method of designing subsurface irrigation/drainage systems to maximize net benefits. *Canadian Water Resources Journal*, **11**(**4**), 46-57.

Nazari, M., and Shamshirband, S. (2018). The particle filter-based back propagation neural network for evapotranspiration estimation. *ISH Journal of Hydraulic Engineering*, 1-7.

Niu, C. J., Deng, W., Gu, S. X., Chen, G., and Liu, S. S. (2017). Real-time irrigation forecasting for ecological water in artificial wetlands in the Dianchi Basin. *Journal of Information and Optimization Sciences*, **38**(**7**), 1181-1196.

Orang, M. N., Scott Matyac, J., and Snyder, R. L. (2008). Survey of irrigation methods in California in 2001. *Journal of Irrigation and Drainage Engineering*, **134**(**1**), 96-100.

Pereira, L. S., Cordery, I., and Iacovides, I. (2012). Improved indicators of water use performance and productivity for sustainable water conservation and saving. *Agricultural water management*, **108**, 39-51.

Pereira, L. S., Oweis, T., and Zairi, A. (2002). Irrigation management under water scarcity. *Agricultural water management*, **57**(**3**), 175-206.

Piri, J., Amin, S., Moghaddamnia, A., Keshavarz, A., Han, D., and Remesan, R. (2009). Daily pan evaporation modeling in a hot and dry climate. *Journal of Hydrologic Engineering*, **14**(**8**), 803-811.

Playan, E., Salvador, R., Lopez, C., Lecina, S., Dechmi, F., and Zapata, N. (2013). Solid-set sprinkler irrigation controllers driven by simulation models: Opportunities and bottlenecks. *Journal of Irrigation and Drainage Engineering*, **140**(**1**), 04013001.

Provenzano, G., and Sinobas, L. R. (2014). Special Issue on Trends and Challenges of Sustainable Irrigated Agriculture.

Raghuwanshi, N. S., Saha, R., Mailapalli, D. R., and Upadhyaya, S. K. (2010). Infiltration Evaluation Strategy for Border Irrigation Management. *Journal of Irrigation and Drainage Engineering*, **137**(**9**), 602-609.

Reca, J., Trillo, C., Sanchez, J. A., Martinez, J., and Valera, D. (2018). Optimization model for on-farm irrigation management of Mediterranean greenhouse crops using desalinated and saline water from different sources. *Agricultural Systems*.

Rowshon, M. K., Mojid, M. A., Amin, M. S. M., Azwan, M., and Yazid, A. M. (2014). Improving irrigation water delivery performance of a large-scale rice irrigation scheme. *Journal of Irrigation and Drainage Engineering*, **140**(**8**), 04014027.

Safriel, U., Adeel, Z., Niemeijer, D., Puigdefabregas, J., White, R., Lal, R., Winsolow, M., Ziedler, J., Prince, S., Archer, E., and King, C. (2006). Dryland systems. In *Ecosystems and Human Well-being. Current State and Trends,* **Vol. 1,** 625-656. Island Press.

Severino, G., D'Urso, G., Scarfato, M., and Toraldo, G. (2018). The IoT as a tool to combine the scheduling of the irrigation with the geostatistics of the soils. *Future Generation Computer Systems*, **82**, 268-273.

Shamshirband, S., Amirmojahedi, M., Gocic, M., Akib, S., Petkovic, D., Piri, J., and Trajkovic, S. (2015). Estimation of reference evapotranspiration using neural networks and cuckoo search algorithm. *Journal of Irrigation and Drainage Engineering*, **142**(**2**), 04015044.

Soulis, K. X., and Elmaloglou, S. (2018). Optimum soil water content sensors placement for surface drip irrigation scheduling in layered soils. *Computers and electronics in agriculture*, **152**, 1-8.

Stambouli, T., Zapata, N., and Faci, J. M. (2011). Irrigation patterns and scheduling of a telecontrolled irrigation district in northeastern Spain. *Journal of Irrigation and Drainage Engineering*, **138**(**6**), 503-516.

Temesgen, B., Eching, S., Davidoff, B., and Frame, K. (2005). Comparison of some reference evapotranspiration equations for California. *Journal of irrigation and drainage engineering*, **131**(**1**), 73-84.

Tindula, G. N., Orang, M. N., and Snyder, R. L. (2013). Survey of irrigation methods in California in 2010. *Journal of Irrigation and Drainage Engineering*, **139**(**3**), 233-238.

Torres-Rua, A. F., Ticlavilca, A. M., Walker, W. R., and McKee, M. (2012). Machine learning approaches for error correction of hydraulic simulation models for canal flow schemes. *Journal of Irrigation and Drainage Engineering*, **138**(**11**), 999-1010.

Valipour, M. (2015). Land use policy and agricultural water management of the previous half of century in Africa. *Applied Water Science*, **5**(**4**), 367-395.

Zhou, B., and Li, L. (2017). Security monitoring for intelligent water-saving precision irrigation system using cloud services in multimedia context. *Multimedia Tools and Applications*.

APPENDIX

Table 10: List of irrigation performance indicator models

References	Performance indicators	Model	Variables considered
Arunkumar and Ambujam, (2010)	T_l: Transmission loss	$T_l = \frac{R_i - R_o}{A_w \times R_l}$	R_i: Reach flow in rate (m^3/s) R_o: Reach flow out rate (m^3/s) R_l: Reach length (m) A_w: Avg. Wet area (m^2)
	O_{pa}: Outcome per planted area (Rs/ha)	$O_{pa} = \frac{CP_v}{P_a}$	CP_v: Crop production value as per local market price (Rs) P_a: Planted area (ha)
	O_{ui}: Outcome per unit irrigated area (Rs/ha)	$O_{ui} = \frac{CP_v}{A_{ui}}$	CP_v: Crop production value as per local market price (Rs) A_{ui}: Unit irrigated area (ha)
	O_{ia}: Outcome per unit irrigation depth applied (Rs/m3)	$O_{ia} = \frac{CP_v}{D_{ia}}$	CP_v: Crop production value as per local market price (Rs) D_{ia}: Depth of irrigation applied (m^3)
	O_{ic}: Outcome per unit depth of irrigation consumed (Rs/m3)	$O_{ic} = \frac{CP_v}{D_{ic}}$	CP_v: Crop production value as per local market price (Rs) D_{ic}: Unit depth of irrigation consumed (m^3)
	R_w: Relative water supply	$R_w = \frac{G_{id}}{ET_c}$	G_{id}: Gross irrigation depth supply (m^3) ET_c: Crop ET requirement (m^3)

Arunkumar and Ambujam, (2010)	R_i: Relative irrigation supply	$R_i = \frac{I_a}{I_r}$	I_a: Irrigation applied (m^3) I_r: Irrigation need
	I_{dc}: Irrigation water delivery capability	$I_{dc} = \frac{C_o}{R_{Peak}}$	C_o: Outflow capability of irrigation water at the system head R_{Peak}: Peak consumptive requirement
	D_d: Dependability of duration	$D_d = \frac{d_a}{d_p}$	d_a: Actual span of water supply (days) d_p: Planned span of water supply (days)
Broner and Lambert, (1989)	N_r: Net Returns	$N_r = (Y * C) - (I_a * I_e)$	Y: Yield (kg/ha) C: Cost ($/kg) I_a: Irrigation depth applied (cm) I_e: Irrigation expenditure ($/cm)
Burt *et al*., (1997)	I_e: Irrigation efficiency	$I_e = \frac{D_b}{D_a - D_s} \times 100\%$	D_b: Depth of irrigation water beneficially utilized D_a: Depth of applied irrigation water D_s: Depth of irrigation water storage
	I_{cu}: Irrigation consumptive use coefficient	$I_{cu} = \frac{D_c}{D_a - D_s} \times 100\ \%$	D_c: Depth of irrigation water consumptively utilized D_a: Depth of applied irrigation water D_s: Depth of irrigation water storage
	I_s: Irrigation sagacity	$I_s = \frac{D_{b/r}}{D_a - D_s} \times 100\ \%$	$D_{b/r}$: Depth of irrigation water beneficially / reasonably utilized D_a: Depth of applied irrigation water D_s: Depth of irrigation water storage
Corcoles *et al*., (2011)	Y_{rw}: Yearly relative water supply	$Y_{rw} = \frac{Y_{id} + E_p}{ET_c}$	Y_{id}: Yearly irrigation depth release (m^3) E_p: Effective precipitation (m^3) ET_c: Crop water consumption (m^3)

Corcoles *et al*., (2011)	Y_{ri}: Yearly relative irrigation supply	$Y_{ri} = \frac{Y_{id}}{ET_c - E_p}$	Y_{id}: Yearly irrigation depth release (m^3) E_p: Effective precipitation (m^3) ET_c: Crop water consumption (m^3)
Hargreaves and Samani, (1984)	Yield response	$\left(1 - \frac{Y_a}{Y_m}\right) = K_y \left(1 - \frac{ET_a}{ET_m}\right)$	Y_a: Actual crop production Y_m: Maximum crop production K_y: Production response ET_a: Actual crop water consumption ET_m: Maximum crop water consumption
Memon *et al*., (1986)	A_i: Annual income	$A_i = R_y * P_y * P_r$	R_y: Relative yield P_y: Potential yield P_r: Price
	A_p: Annual profit	$A_p = A_i - G_{ac}$	A_i: Annual income G_{ac}: Gross annual cost
Raghuwanshi *et al*., (2010)	D_r: Deep percolation ratio	$D_r = \frac{D_{dp}}{q_o B_e T_e}$	D_{dp}: Depth of deep percolation (m^3) q_o: Flow in rate per unit border extent (m^3/ m/s) B_e: Border extent (m) T_e: End time (s)
	T_r: Tail water ratio	$T_r = 100 - D_r - A_e$	D_r: Deep percolation ratio A_e: Application efficiency
Rowshon *et al*., (2014)	A_q: The adequacy of irrigation	$A_q = \frac{1}{t}\sum_{1}^{i}\left\{\sum_{1}^{i}\left[\frac{1}{i}\left(\frac{Q_d}{Q_r}\right)\right]\right\}$	t: Time periods for water supply i: Unit area belongs to a channel released by the system over time t. Q_d: Daily actual discharge Q_r: Irrigation need
	E_q: The equity of water distribution	$E_q = 1 - \frac{1}{t}\sum_{1}^{i} C_{vr}\left(\frac{Q_d}{Q_r}\right)$	C_{vr}: Spatial coefficient of variation
Rowshon *et al*., (2014)	D_p:The dependability of the water supply	$D_p = 1 - \frac{1}{i}\sum_{1}^{i} C_{vt}\left(\frac{Q_d}{Q_r}\right)$ When $Q_d \leq Q_r$	C_{vt}: Temporal coefficient of variation

Stambouli *et al*., (2011)	N_{ir}: Net irrigation requirement	$N_{ir} = (K_c * ET_0) - E_{rf}$	ET_0: Reference plant water consumption E_{rf}: Effective rain fall K_c: Plant Coefficient
	N_{rdi}: Net regulated deficit irrigation	$N_{rdi} = (K_c * K_{rc} * ET_0) - E_{rf}$	K_{rc}: Reduction coefficient
	S_{ipi}: Seasonal irrigation performance index	$S_{ipi} = \frac{N_{ir}}{I_{ad}}$	N_{ir}: Net irrigation requirement I_{ad}: Irrigation application depth
Pereira *et al*., (2012)	WP: Water productivity (kg/ m3)	$WP = \frac{Y}{I_{ws}}$	Y: Yield (kg/ha) I_{ws}: Irrigation water supplied (m^3)
Cetin and Kara, (2019)	EWP: Economic water productivity ($/m3)	$EWP = \frac{N_r}{I_a}$	N_r: Net returns ($) I_a: Irrigation depth applied (m^3)

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 199-216

Cause Specific Shared Frailty Proportional Reversed Hazards Models

Gleeja V. L[1]. and P. G. Sankaran[2]
1. *Department of Statistics, College of Veterinary and Animal Sciences, Mannuthy-680651, Kerala, India*
2. Department of Statistics, Cochin University of Science and Technology, Cochin-682022, Kerala, India

Received: 22 May 2020; Revised: 17 December 2020; Accepted: 20 December 2020

Abstract

In family studies, usually, information on ages at onset of diseases is collected and the resulting data are often left censored. When there is a possibility of multiple types of events (disease) in a left censored family (clustered) data, the concept of cause specific reversed hazard models and the concept of frailty are needed for modeling and analysis of the data. Hence, in this paper, for the analysis of clustered multiple event data with left censored observations, frailty models in terms of cause specific reversed hazard rates are introduced. The shared gamma frailty reversed hazards model for bivariate multiple event data are developed. The first model is developed for the analysis of data without the presence of covariates. In the second model, covariates are included and regression coefficients are assumed to be different for different type of events. The estimation of the parameters of the models by maximum likelihood method, using EM algorithm, is presented. The properties of the estimates are also discussed. Finally, the models are applied to real data sets.

Key words: Reversed hazard rate; Competing risk; Frailty models.

1. Introduction

In family studies on diseases with ages at onset, assessing the familial association is often the problem of interest. When age at onset is considered, the data is often left censored. Left censored observations occur when the exact value of a response has not been observed and instead, an upper bound on that response is observed. Such observations also arise if a measuring instrument lacks the sensitivity needed to measure the observations below a known threshold. Then the measurement is taken and if the signal is below the instrument threshold, all which is known is that measurement is less than the threshold. Left censored observations also occur in studies determining the age at which a child learns to perform a specified task. Often, some children can already perform the task when they enter to the study. Such lifetimes are considered as left censored. The modeling and analysis of such left censored lifetime data is carried out using reversed hazard rate. The concept of reversed hazard rate (RHR) has been proposed as dual to hazard rate by Barlow *et al.* (1963) and is defined for a nonnegative random variable T as $(t) = \lim_{\Delta t \to 0} \frac{P(t-\Delta t < T \leq t+\Delta t)}{\Delta t}$. That is, in a small interval, the product of the RHR function and the length of the interval is the approximate probability of failure in the interval given failure before the end of the interval. RHR was used for the estimation of the survival

Corresponding Author: Gleeja V. L.
E-mail: gleeja@gmail.com

function in the presence of left censored observations by Ware and DeMets (1976) for a baboon descent data. Later RHR was used for characterization of life distributions by Shaked and Shantikumar (1994), for investigating the properties for k out of n systems by Block *et al.* (1998) and for developing nonparametric estimators for right truncated data by Lawless (2003). Different authors, Chandra and Roy (2001), Gupta and Nanda (2001), Gupta and Wu (2001), Kalbfleisch and Lawless (1989), Nair, Sankaran and Asha (2005), Sankaran and Gleeja (2006), Bartoszewicz and Skolimowska (2006) and Faith (2017), extensively studied and presented results related to RHR.

Sometimes in studies involving family or subgroups, lifetimes of individuals within the subgroup may be related. For modeling association between individual lifetimes within subgroups, the notion of frailty was introduced by Vaupel *et al.* (1979). The model assumes frailty as a common random effect that acts multiplicatively on the hazard rates of all subgroup members. The most widely used frailty model is shared frailty model with gamma distribution as frailty distribution. It has been discussed in Vaupel *et al.* (1979), Clayton and Cuzick (1985), Klein (1992) and Andersen *et al.* (2003). Some other distributions for frailty like positive stable, Weibull, lognormal *etc.* are investigated in Hougaard (2000). The estimation of the parameters of shared frailty model using maximum likelihood method via the EM algorithm is developed in Nielson *et al.* (1992) and the asymptotic normality and efficiency of the estimators are studied and proved in Murphy (1994, 1995).

The concept of frailty as a common random effect that acts multiplicatively on RHR has been introduced in Sankaran and Gleeja (2008). Let (T_1, T_2) be the lifetimes of two related individuals. Then $m_j(Z, t_j) = Z m_{0j}(t_j)$ be conditional individual RHRs given frailty Z where $m_{0j}(t_j)$, $j = 1, 2$ are the baseline reversed hazards. Assume that lifetimes (T_1, T_2) are conditionally independent given frailty Z and Z follows a gamma distribution with mean one and variance θ. Then shared gamma frailty models is introduced by Sankaran and Gleeja (2011) as the distribution function of (T_1, T_2), $F(t_1, t_2) = \left[\theta M_{01}(t_1) + \theta M_{02}(t_2) + 1\right]^{-(1/\theta)}$ where $M_{01}(t_1)$ and $M_{02}(t_2)$ are the cumulative baseline reversed hazard function and $\theta \geq 0$.

When time to failure of paired organs like kidney, lungs, eyes, ears, dental implants *etc.* are considered, it is more appropriate to model using shared frailty models. The estimation was done using maximum likelihood method via EM algorithm. Later, estimation of parameters involved in the shared frailty model by the Bayesian estimation procedure using the Markov chain Monte Carlo (MCMC) technique was discussed in Hanagal *et al.* (2014). The most commonly used frailty distribution is Gamma distribution, because of its mathematical convenience. Other distributions can be used as frailty distribution and Hanagal and Pandey (2015) developed three parametric shared frailty models with inverse Gaussian frailty using RHR. Gamma frailty models with different baseline distributions are discussed in Hanagal and Bhambure (2017) and Hanagal and Pandey (2017). Inverse Gaussian correlated frailty model with different baseline distributions are discussed in Hanagal (2020) and Hanagal and Pandey (2020). The shared frailty models are attracting recent interest of researchers and extensive research is being conducted on these models. While analyzing family data on age at onset of a particular disease, shared frailty models using RHR is very useful.

But in certain studies on age at onset of diseases, individuals may be susceptible to more than one type of diseases or in some survival studies death can occur due to any one of the two

or more causes. When there are multiple types of causes for the event, the concept of competing risks facilitates analysis. Modeling and analysis of lifetime data with multiple type of events under right censoring is discussed in Crowder (2001), Kalbfleisch and Prentice (2002) and Lawless (2003). The analysis of competing risks data under left censoring using RHR has been carried out in Sankaran and Anjana (2014). Specifically, they presented the analysis of left censored data with multiple types of events using cause-specific RHR functions. Let (T, J) be a pair of random variables where T is possibly a censored lifetime and J represents cause of event. J takes values on the set {1,2,...,*r*}. These *r* causes are mutually exclusive and exhaustive, so that the individual can have at most one realized lifetime with an identifiable cause. Then cause specific RHR of *T* is defined as $m_j(t) = \lim_{\Delta t \to 0} \frac{P(t - \Delta t < T \le t, J = j / T \le t)}{\Delta t}$, *j* = 1,2,...,*r*. Thus $m_j(t)$ specifies the instantaneous rate of failure of an individual at time *t* due to cause *j* given that it failed before time *t*. Then the marginal RHR of T was given as $m(t) = \sum_{j=1}^{r} m_j(t)$. Sankaran and Anjana (2016) introduced a proportional cause specific RHR model for modeling and analysis of left censored competing risks data in the presence of covariates. The model was given as $m_j(t \mid \mathbf{x}) = m_{0j}(t)\exp(\boldsymbol{\beta}'\mathbf{x})$, *j*=1,2,...,*r* where $m_j(t \mid \mathbf{x})$ is the cause-specific RHR due to cause *j* in the presence of covariate **x** and **x** is a vector of p covariates, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ is the vector of p regression parameters, and $m_{0j}(t)$ is the baseline cause-specific RHR due to cause j. The vector of regression parameters **β** measures the effect of the covariate vector on the cause-specific RHR. But these models are not appropriate for clustered data like family data, as it does not consider the association exist between members of the family.

Thus, in order to analyze a left censored family (clustered) data with multiple types of events (diseases), a frailty-based competing risks models using RHR is needed. Motivated by this, in this paper, a shared gamma frailty model in terms of cause specific RHR is developed.

The paper is organized as follows. In section 2, cause specific shared frailty proportional RHR model is developed with and without the presence of covariates. The estimation and the asymptotic properties of the parameters of models are studied in section 3. In section 4, the model is illustrated with data sets from Ying and Wei (1994) and McGilchrist and Aisbett (1991). Finally conclusions and discussions are given in section 5.

2. Cause Specific Shared Frailty Proportional RHR Model

The model is constructed to deal with a clustered or family data with multiple causes of event. The time to event is the variable of interest and let us consider bivariate situation.

2.1. Cause specific shared frailty proportional RHR model without covariates

Let $T = (T_1, T_2)$ be the pair of lifetimes of two related individuals defined on a common probability space (Ω, F, P) with absolutely continuous distribution function. Let $F(t_1, t_2)$ and $F_j(t_j)$ respectively denote the joint distribution function of T and the marginal distribution function of T_j, *j=1,2*. Let the support of T be $D = [0, b_1] \times [0, b_2]$ where (b_1, b_2) is such that

$b_j = \inf\left\{t \mid F_j(t) = 1\right\}$, $j = 1,2$. Assume that each of the pair (T_1, T_2) is subject to multiple causes of event. Let $C = (C_1, C_2)$ denote the cause of event for T. Suppose that there are *r* causes for the event for each individual in the process. Assume that C_j is a unique element of the set {1,2,…,*r*}, *j*=1,2. We assume that individual can have at most one realized lifetime with an identifiable cause. Observations from the same cluster or family may share common environment or some other factors. Hence it is assumed that the pair of lifetimes shares a common unobserved frailty Z. First define the cause specific RHR of T_j for given frailty *Z* as

$$m_{jk}(Z, t_j) = Z m_{0jk}(t_j),\ j=1,2 \text{ and } k=1,2,\ldots,r$$

where $m_{0jk}(t_j)$ is baseline cause specific RHR function of T_j, *j*=1,2 and Z is an unobservable random variable having a probability density function $g(z)$. The marginal RHR of T_j for given frailty *Z* is obtained as

$$m_j(Z, t_j) = \sum_{k=1}^{r} m_{jk}(Z, t_j) = Z \sum_{k=1}^{r} m_{0jk}(t_j),\ j=1,2.$$

We assume that lifetimes (T_1, T_2) are conditionally independent given frailty Z. Then the distribution function of (T_1, T_2) given frailty Z is

$$F(t_1, t_2 \mid Z) = \exp\left\{-\sum_{k=1}^{r} \int_{t_1}^{\infty} m_{1k}(Z, u)\, du - \sum_{k=1}^{r} \int_{t_2}^{\infty} m_{2k}(Z, v)\, dv\right\}.$$

Let $g(z)$ be the joint density function of Z. Then the bivariate distribution function of (T_1, T_2) is

$$\begin{aligned} F(t_1, t_2) &= \int_0^{\infty} F(t_1, t_2 \mid Z)\, g(z)\, dz \ = E\left(F(t_1, t_2 \mid Z)\right) \\ &= E\left(\exp\left\{-Z\left(\sum_{k=1}^{r} M_{01k}(t_1) + \sum_{k=1}^{r} M_{02k}(t_2)\right)\right\}\right) \end{aligned}$$

where $M_{01k}(t_1)$ and $M_{02k}(t_2)$ are the cumulative baseline cause specific reversed hazard function.

The marginal distribution function of T_j is

$$F_j(t_j) = \int_0^{\infty} F_j(t_j \mid z)\, g(z)\, dz = E\left(\exp\left\{-Z \sum_{k=1}^{r} M_{01k}(t_1)\right\}\right),\ j = 1,2.$$

Suppose that *Z* is *i.i.d.* random variable with the following gamma density function

$$g(z)=\frac{z^{(1/\theta)-1}\exp\{-z/\theta\}}{\theta^{(1/\theta)}\Gamma(1/\theta)},\qquad \theta\geq 0.$$

The mean value of Z is 1 and variance is θ. Then bivariate distribution function of (T_1,T_2) is obtained as

$$F(t_1,t_2)=\left[\theta\sum_{k=1}^{r}M_{01k}(t_1)+\theta\sum_{k=1}^{r}M_{02k}(t_2)+1\right]^{-(1/\theta)}\qquad \text{where } \theta\geq 0. \tag{1}$$

Thus, cause specific shared gamma frailty proportional RHR model can be represented by (1).

The marginal distribution function T_j is then obtained as

$$F_j(t_j)=\left[\theta\sum_{k=1}^{r}M_{0jk}(t_j)+1\right]^{(-1/\theta)},\ j=1,2.$$

Therefore, the bivariate distribution function of (T_1,T_2) can be represented in terms of marginal distribution functions as

$$F(t_1,t_2)=\left[F_1(t_1)^{-\theta}+F_2(t_2)^{-\theta}-1\right]^{-(1/\theta)}\qquad \text{where } \theta\geq 0. \tag{2}$$

Remark 1: T_1 and T_2 are independent, when $\theta = 0$.

Remark 2: The model given in (2) is identifiable. Let $F(t_1,t_2)$ be a known distribution function given as in (2), and $\theta>0$, and let $m_i(b_i)=\sum_{k=1}^{r}m_{ik}(b_i)\neq 0$, $i=1,2$. $k=1,2,...,r$ for r different causes.

We obtain the joint density function as

$$f(t_1,t_2)=\frac{(1+\theta)\sum_{k=1}^{r}m_{1k}(t_1)\sum_{k=1}^{r}m_{2k}(t_2)F_1(t_1)^{-\theta}F_2(t_2)^{-\theta}}{\left[F_1(t_1)^{-\theta}+F_2(t_2)^{-\theta}-1\right]^{\left(\frac{1}{\theta}\right)+2}}.$$

Since $m_i(b_i)=\sum_{k=1}^{r}m_{ik}(b_i)\neq 0$ and $F_i(b_i)=1$, $i=1,2$, we have

$$\theta=\frac{f(b_1,b_2)}{m_1(b_1)m_2(b_2)}-1.$$

From the above expression it is clear that the identified value is unique. Then the model is identifiable.

2.2. Cause specific shared frailty proportional RHR model with covariates

Often the lifetime of individual is influenced by age, gender, history or severity of diseases. If information is available about such factors, then the heterogeneity in a population arising from the influence of those factors can be incorporated in models by specifying them as covariates. Accordingly, cause specific RHR function of T_j in presence of covariates and frailty is defined as

$$m_{jk}\left(Z,t_j\right)=Zm_{0jk}\left(t_j\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_j)$$

for j = 1,2 and k = 1,2,…,r, where and $\mathbf{x}_j=\left(x_{j1},x_{j2},...,x_{jp}\right)'$, j = 1,2 is a $p\times1$ vector of covariates and $\boldsymbol{\beta}_k=\left(\beta_{1k},\beta_{2k},...,\beta_{pk}\right)'$ is the vector of regression coefficients and are assumed to be different for different causes of events.

Proceeding as in Section 2.1, bivariate distribution function of $\left(T_1,T_2\right)$ is obtained as

$$F(t_1,t_2)=E\left(\exp\left\{-Z\left(\sum_{k=1}^{r}M_{01k}\left(t_1\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_1)+\sum_{k=1}^{r}M_{02k}\left(t_2\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_2)\right)\right\}\right). \qquad (3)$$

Assuming that frailty variable Z follows gamma distribution with mean one and variance θ, (3) reduces to

$$F\left(t_1,t_2\right)=\left[\theta\sum_{k=1}^{r}M_{01k}\left(t_1\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_1)+\theta\sum_{k=1}^{r}M_{02k}\left(t_2\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_2)+1\right]^{-(1/\theta)} \text{ where } \theta>0. \qquad (4)$$

The marginal distribution function T_j is then obtained as

$$F_j\left(t_j\right)=\left[\theta\sum_{k=1}^{r}M_{0jk}\left(t_j\right)\exp(\boldsymbol{\beta}_k{}'\mathbf{x}_j)+1\right]^{(-1/\theta)},\ j=1,2.$$

When the bivariate distribution function of $\left(T_1,T_2\right)$ is represented in terms of marginal distribution functions, (4) reduces to (2).

The parameters of the model could be estimated from observed data only if the model is identifiable. The identifiable property of the models follows from Sankaran and Gleeja (2011).

3. Estimation

The estimation procedures are developed for cause specific shared gamma frailty proportional RHR model when the data is left censored. Let $T=\left(T_1,T_2\right)$ be the pair of lifetimes

of two related individuals and $U=(U_1,U_2)$ be a pair of corresponding censoring times defined on a common probability space (Ω,F,P) with absolutely continuous distribution function. Under bivariate censoring, one could observe $(T_1^*,T_2^*,C_1,C_2,\delta_1,\delta_2)$ where $T_j^*=\max(T_j,U_j)$ and $\delta_j=I(T_j=T_j^*)$, j=1,2 with $I(.)$ as usual indicator function. Suppose that $(T_{i1}^*,T_{i2}^*,C_{i1},C_{i2},\delta_{i1},\delta_{i2})$, i=1,2,…,n are n independent and identically distributed observations of $(T_1^*,T_2^*,C_1,C_2,\delta_1,\delta_2)$. Define $N_{ijk}(t)=I\{T_{ij}^*\le t,\delta_{ij}=1,C_{ij}=k\}$, $Y_{ijk}(t)=I\{T_{ij}^*\le t,C_{ij}=k\}$, for i=1,2,…,n, j=1,2, k=1,2,…,r. Define the predictable process $\boldsymbol{Y}(\boldsymbol{t})$ as

$$\boldsymbol{Y}(\boldsymbol{t})=(Y_{ijk}(t),\ i=1,2,...,n,\ j=1,2,\ k=1,2,...,r)$$

and

$$\boldsymbol{N}(\boldsymbol{t})=(N_{ijk}(t),\ i=1,2,...,n,\ j=1,2,\ k=1,2,...,r\ )$$

as a multivariate counting process with components N_{ijk}, where components with the same value of the first index i share the same frailty variable Z_i. Further it is assumed that conditional on Z, T and U are independent.

3.1. Estimation of cause specific shared frailty proportional RHR model without covariates

For the case without covariates, let the cause specific RHR of T_j for given frailty Z for the ith cluster is $m_{ijk}(t)=Z_iY_{ijk}(t)m_{0jk}(t)$, $i=1,2,...,n,\ j=1,2,\ k=1,2,...,r$, where Z_i is unobservable *i.i.d.* random variable from Gamma $(1/\theta,\ 1/\theta)$ distribution. The estimate of parameter θ and cumulative baseline cause specific RHR function $M_{0jk}(t)=\int_t^b m_{0jk}(s)ds$ is to be obtained. Assume that conditional on $Z=z$, censoring is non-informative. So the partial conditional likelihood based on $\boldsymbol{N}(\boldsymbol{t})$ is given by the product integral

$$dP=\prod_t\left\{\prod_i\prod_j\prod_k\left[m_{ijk}(t)dt\right]^{\Delta N_{ijk}(t)}\left[1-m_{\cdot}(t)dt\right]^{1-\Delta N_{\cdot}(t)}\right\}, \tag{5}$$

where $N_{\cdot}(t)=\sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{k=1}^{r}N_{ijk}(t)$ and $m_{\cdot}(t)=\sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{k=1}^{r}m_{ijk}(t)$. Considered as a function of Z, (5) is proportional to conditional density of $(\boldsymbol{N}(\boldsymbol{t}),\boldsymbol{Y}(\boldsymbol{t}))$ given $Z=z$. Substituting the specification of $m_{ijk}(t)$ and evaluating the product integral, $L(\theta)$ is obtained as

$$L(\theta)=\prod_i\left\{\frac{z_i^{(1/\theta)-1}\exp\{-z_i/\theta\}}{\theta^{(1/\theta)}\Gamma(1/\theta)}\prod_j\prod_k\exp\left[-z_i\int_0^b Y_{ijk}(s)dM_{0jk}(s)\right]\prod_t\left(z_iY_{ijk}(t)dM_{0jk}(t)\right)^{\Delta N_{ijk}(t)}\right\} \tag{6}$$

$$=\prod_i\frac{z_i^{(1/\theta)+N_{i\cdot\cdot}(b)-1}\exp\{-z_i(\frac{1}{\theta}+\sum_j\sum_k\int_0^b Y_{ijk}(s)dM_{0jk}(s))\}}{\theta^{(1/\theta)}\Gamma(1/\theta)}\prod_j\prod_k\prod_t\left(Y_{ijk}(t)dM_{0jk}(t)\right)^{\Delta N_{ijk}(t)}$$

where $N_{i..}(b) = \sum_{j=1}^{2}\sum_{k=1}^{r} N_{ijk}(b)$. Conditional on data, Z_i are still independent and gamma distributed with parameters $(1/\theta) + N_{i..}(b)$ and $(1/\theta) + \sum_j \sum_k \int_0^b Y_{ijk}(s)\, dM_{0jk}(s)$. Integrating out Z in (6), the marginal partial likelihood is obtained as

$$L(\theta) = \prod_i \left\{ \frac{\left(\Gamma\left((1/\theta) + N_{i..}(b)\right) \prod_j \prod_k \prod_t \left(Y_{ijk}(t)\, dM_{0jk}(t) \right)^{\Delta N_{ijk}(t)} \right)}{\left(\theta^{(1/\theta)} \Gamma(1/\theta) \left[(1/\theta) + \sum_{j=1}^{2}\sum_{k=1}^{r} \int_0^b Y_{ijk}(s)\, dM_{0jk}(s) \right]^{(1/\theta)+N_{i..}(b)} \right)} \right\}. \tag{7}$$

EM algorithm is used to maximize (7). The estimates of parameters which maximizes (7) maximizes (6) also. The E step is to estimate

$$\hat{z}_i = \frac{(1/\theta) + N_{i..}(b)}{(1/\theta) + \sum_{j=1}^{2}\sum_{k=1}^{r} \int_0^b Y_{ijk}(s)\, dM_{0jk}(s)}.$$

The M step is then to calculate $\hat{\theta}$, the maximum likelihood estimator for θ from (7), and to estimate cumulative baseline cause specific RHR function

$$\hat{M}_{0jk}(t) = \int_t^b \frac{dN_{.jk}(s)}{\sum_i \hat{z}_i Y_{ijk}(s)}, \text{ where } N_{.jk}(s) = \sum_{i=1}^{n} N_{ijk}(s).$$

The initial estimates of $\hat{z}_i$ and $\hat{M}_0(t)$ are obtained by taking $\theta = 0$. By general theory of EM algorithm, if this algorithm converges, it converges to a stationary point of $\log L(\theta)$.

3.2. Estimation of cause specific shared frailty proportional RHR model with covariates

For the model with covariates, let the vector observed be $\left(T_{i1}^{*}, T_{i2}^{*}, C_{i1}, C_{i2}, \delta_{i1}, \delta_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}\right)$. Then cause specific proportional RHR in presence of covariates and frailty is represented as

$$m_{ijk}(t) = z_i Y_{ijk}(t)\, m_{0jk}(t) \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) \tag{8}$$

for j = 1,2 and k = 1,2,…,r, where $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, ..., \beta_{pk})'$ is the vector of regression coefficients and $\mathbf{x}_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})'$, j=1,2 is a px1 vector of covariates. Then the likelihood function conditional on the covariate $\mathbf{x}_{ij}$ and frailty Z_i for the model (8) is given as

$$L(\boldsymbol{\beta}, Z_i, \mathbf{x}_{ij}) = \prod_i \left\{ \prod_j \prod_k \prod_t \left(z_i Y_{ijk}(t) \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) dM_{0jk}(t) \right)^{\Delta N_{ijk}(t)} \exp\left[-z_i \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) \int_0^b Y_{ijk}(s) dM_{0jk}(s) \right] \right\}.$$

Let Z_i be unobservable *i.i.d.* random variable from Gamma $(1/\theta, 1/\theta)$ distribution. Then $L(\theta, \boldsymbol{\beta})$ is obtained as

$$L(\theta, \boldsymbol{\beta}) = \prod_i \frac{z_i^{(1/\theta)+N_{i..}(b)-1} \exp\{-z_i (\frac{1}{\theta} + \sum_j \sum_k \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) \int_0^b Y_{ijk}(s) dM_{0jk}(s))\}}{\theta^{(1/\theta)} \Gamma(1/\theta)} \prod_j \prod_k \prod_t \left(\exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) Y_{ijk}(t) dM_{0jk}(t) \right)^{\Delta N_{ijk}(t)}.$$

Then loglikelihood function can be written as $\log L(\theta, \boldsymbol{\beta}) = l_1(\theta) + l_2(\boldsymbol{\beta})$ where

$$l_1(\theta) = \sum_i \frac{1}{\theta} \log \frac{1}{\theta} - \log \Gamma(1/\theta) + \left[(1/\theta) + N_{i..}(b) - 1 \right] \log z_i - \frac{z_i}{\theta}, \tag{9}$$

and

$$l_2(\boldsymbol{\beta}) = \sum_i (-z_i) \sum_j \sum_k \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) \int_0^b Y_{ijk}(s) dM_{0jk}(s) + \log \left[\prod_j \prod_k \prod_t \left(\exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) Y_{ijk}(t) dM_{0jk}(t) \right)^{\Delta N_{ijk}(t)} \right]. \tag{10}$$

Proceeding as in the case without covariates and using EM algorithm, estimate of Z_i is obtained as

$$\hat{z}_i = \frac{(1/\theta) + N_{i..}(b)}{(1/\theta) + \sum_{j=1}^{2} \sum_{k=1}^{r} \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) \int_0^b Y_{ijk}(s) dM_{0jk}(s)}.$$

The M step is then to calculate $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}_k$, the maximum likelihood estimator for θ and $\boldsymbol{\beta}_k$ from (9) and (10) respectively and to estimate cumulative baseline cause specific RHR function

$$\hat{M}_{0jk}(t) = \int_t^b \frac{dN_{.jk}(s)}{\sum_i \hat{z}_i \exp(\boldsymbol{\beta}_k' \mathbf{x}_{ij}) Y_{ijk}(s)}.$$

For testing of independence of variables T_1 and T_2, likelihood ratio test can be used. The case of shared frailty model with no covariates and with standard conditions on the censoring distribution is discussed in Maller and Zhou (2003). They obtained that the likelihood ratio statistic has an asymptotic null distribution which is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom. For testing $H_0 : \theta = 0$, the likelihood ratio test statistic is -2log Q = 2 (log $L(\hat{\theta})$-log $L(0)$). When -2log $Q > 0.5(\chi^2_{0,\alpha} + \chi^2_{1,\alpha})$, the null hypothesis is rejected at 5% level of significance.

The asymptotic properties of the estimators follow from Sankaran and Gleeja (2011). The consistency of the estimators is established in Theorem 1.

Let T_1 be the first jump of N, θ_0 lies in a known interval $[0, S]$ and true cumulative baseline reversed hazard M_{0jk} be strictly decreasing and continuous on $[0, b]$ for $b<\infty$.

Theorem 1: Assume that

i. Y is a non-decreasing step function and $P\left(Y\left(t\right)\geq 1\right)$ has at most finite number of discontinuities in $t\in\left(0,b\right)$,

ii. $\underset{u\in(0,b)}{Inf}\ E\left(Y\left(u\right)\right)>0$,

iii. $P\left(Y\left(T_1\right)\geq 1\right)<1$,

then $\underset{t\in(0,b)}{Sup}\left|\hat{M}_{0jk}\left(t\right)-M_{0jk}\left(t\right)\right|\to 0$ almost surely (a. s.) and $\left|\hat{\theta}-\theta_0\right|\to 0$ a. s.

Proof: The assumption (i) is used to prove that $\hat{M}_{0jk}\left(t\right)$ does not diverge to infinity, (ii) is used to ensure that counting process N has sufficient activity on the entire interval so as to estimate the parameters, and (iii) excludes the possibility of N having at most only one jump. The model becomes unidentifiable if all N_i have only one jump. The rest of the proof is similar to the one given in Murphy (1994).

The asymptotic normality of the estimators can be established in the following way. Set $M_{0jkt}\left(0\right)=\int_0^b 1+th_1\left(u\right)d\hat{M}_{0jk}\left(u\right)$ and $\theta_t=th_2+\hat{\theta}$ for h_1 a function and h_2 a scalar, and differentiate at $t=0$ to get $F_n\left(\hat{M}_{0jk},\hat{\theta}\right)\left(h_1,h_2\right)$. Then, if $\left(\hat{M}_{0jk},\hat{\theta}\right)$ maximizes $\log L\left(\theta\right)$, then $F_n\left(\hat{M}_{0jk},\hat{\theta}\right)\left(h_1,h_2\right)=0$ for all $\left(h_1,h_2\right)$. The form of F_n is given by $F_n=F_{n1}+F_{n2}$, where

$$F_{n1}\left(M_{0jk},\theta\right)\left(h_1\right)=n^{-1}\sum_{i=1}^{n}\int_0^b h_1dN_{ijk}-\frac{\theta^{-1}+N_{ijk}\left(b\right)}{\theta^{-1}+\int_0^b Y_{ijk}dM_{0jk}}\int_0^b h_1Y_{ijk}dM_{0jk}$$

and $$F_{n2}\left(M_{0jk},\theta\right)\left(h_2\right)=h_2n^{-1}\sum_{i=1}^{n}\int_0^b\frac{N_{ijk}\left(u\right)}{1+\theta N_{ijk}\left(u\right)}dN_{ijk}\left(u\right)$$

$$+\theta^{-2}\left(\log\left(1+\theta\int_0^b Y_{ijk}dM_{0jk}\right)-\frac{\theta^{-1}+N_{ijk}\left(b\right)}{\theta^{-1}+\int_0^b Y_{ijk}dM_{0jk}}\theta\int_0^b h_1Y_{ijk}dM_{0jk}\right)$$

For $\theta=0$, the last term is taken as its limit as θ approaches zero to get $\left(\left(\int_0^b Y_{ijk}dM_{0jk}\right)^2\Big/2\right)+N_{ijk}\left(b\right)\int_0^b Y_{ijk}dM_{0jk}$. The class of h is taken to be the space of bounded variation cross the reals. Define the norm to be $\left\|h\right\|_H=\left\|h_1\right\|_v+\left|h_2\right|$, where $\left\|h_1\right\|_v$ is absolute value of $h_1\left(0\right)$ plus the total variation of h_1 on the interval $\left[0,b\right]$. Define H_p to be the product space

of bounded variation functions on $[0,b]$ and real valued scalars with norm $\|h\|_H = \|h_1\|_v + |h_2| \le p$. If $p = \infty$, then the inequality is strict. In the following p is assumed to be finite unless stated otherwise. Define $\left(M_{0jk},\theta\right)(h) = \int_0^b h_1 dM_{0jk} + h_2\theta$. Then the parameter space Ψ can be considered to be a subset of $l^\infty\left(H_p\right)$, which is the space bounded by real valued functions on H_p under the supremum norm $\|U\| = \sup_{h\in H_p}|U(h)|$. The score function F_n is a random map from Ψ to $l^\infty\left(H_p\right)$ for all finite p.

Theorem 2: Assume that

i. $\underset{t\in(0,b)}{Sup}\left|\hat{M}_{0jk}(t) - M_{0jk}(t)\right| \to 0$ a.s. and $\left|\hat{\theta} - \theta_0\right| \to 0$ a. s.,

ii. There exist a constant K for which $\|Y\| \le K$ and $N(b) \le K$ a.s. ,

iii. $\underset{u\in(0,b)}{Inf}\ E\left(Y(u)\right) > 0$,

iv. $P\left(Y(T_1) \ge 1\right) < 1$,

then $\sqrt{n}\left|\hat{M}_{0jk}(t) - M_{0jk}(t)\right|.\sqrt{n}\left|\hat{\theta} - \theta_0\right| \Rightarrow \mathrm{G}$ on $l^\infty\left(H_p\right)$; G is a tight Gaussian process on $l^\infty\left(H_p\right)$ with mean zero and covariance process $Cov\left(G(h), G(h')\right) = \int_0^b h_1\tilde{\sigma}_{(1)}^{-1}(h')dM_{0jk} + h_2\tilde{\sigma}_{(2)}^{-1}(h')$, where $\tilde{\sigma} = \left(\tilde{\sigma}_1, \tilde{\sigma}_2\right)$ is a continuously invertible linear operator from H_∞ onto H_∞ with inverse $\tilde{\sigma}^{-1} = \left(\tilde{\sigma}_{(1)}{}^{-1}, \tilde{\sigma}_{(2)}{}^{-1}\right)$. The form of $\tilde{\sigma}$ is as follows:

$$\tilde{\sigma}_1(h)(u) = h_1(u)E\left(ZY(u)\right) - \sum_{j=1}^{2}\sum_{k=1}^{r} E\left(\frac{\theta_0\int_0^b Y_{jk}h_1 dM_{0jk}}{1+\theta_0\int_0^b Y_{jk}dM_{0jk}}ZY_{jk}(u)\right)$$

$$-h_2 E\left(\frac{Y(u)}{1+\theta_0\int_0^b YdM_{0jk}}\left(\int_0^b ZYdM_{0jk} - N(b)\right)\right)$$

and

$$\tilde{\sigma}_2(h) = h_2\sum_{j=1}^{2}\sum_{k=1}^{r} E\left(-\frac{\partial^2 \log L(\theta, M_{0jk})}{\partial(\theta)^2}\Big|_{(\theta_0, M_{0jk})}\right) - \sum_{j=1}^{2}\sum_{k=1}^{r} E\left(\frac{\int_0^b Yh_1 dM_{0jk}}{1+\theta_0\int_0^b YdM_{0jk}}\left(\int_0^b ZYdM_{0jk} - N(b)\right)\right)$$

where

$$-\frac{\partial^2 \log L(\theta, M_{0jk})}{\partial(\theta)^2}\Big|_{(\theta_0, M_{0jk})} = n^{-1}\sum_{i=1}^{n}\int_0^b \left(\frac{N_{ijk}(u)}{1+\theta_0 N_{ijk}(u)}\right)^2 dN_{ijk}(u) - N_{ijk}(b)\left(\frac{\int_0^b Y_{ijk}\,dM_{0jk}}{1+\theta_0\int_0^b Y_{ijk}\,dM_{0jk}}\right)^2$$

$$+2\theta_0^{-3}\left(\ln\left(1+\theta_0\int_0^b Y_{ijk}\,dM_{0jk}\right) - \frac{\theta_0\int_0^b dM_{0jk}}{1+\theta_0\int_0^b Y_{ijk}\,dM_{0jk}} - \frac{1}{2}\left(\frac{\theta_0\int_0^b dM_{0jk}}{1+\theta_0\int_0^b Y_{ijk}\,dM_{0jk}}\right)^2\right).$$

When $\theta_0 = 0$, the last term above is defined by its limit, which is $\frac{2}{3}\sum_{j=1}^{2}\sum_{k=1}^{r}\left(\int_0^b Y_{ijk}\,dM_{0jk}\right)^3$.

Proof: Proof of the theorem follows from Murphy (1995).

4. Data Analysis

The analysis of the proposed model is illustrated with data concerning the times to tumor appearance or death of mice from the same litter in a tumor genesis experiment by Mantel and Ciminera (1979), reported in Ying and Wei (1994). The observations from drug treated rat (T_1) and its litter matched control (T_2) which were either dead ($C_i = 1, i = 1,2$) or appeared with tumor ($C_i = 0, i = 1,2$) are considered for the analysis. The observations with the value 60 indicate left censored observations. The indicator function $\delta_i = 0,\ i = 1,2$ if observation is left censored, $\delta_i = 1,\ i = 1,2$ if it is not censored. The data consist of 22 pairs. The analysis of data is carried out with cause specific shared frailty proportional RHR model without covariates. Then the maximum likelihood estimate for θ is obtained as 0.7557 and is significantly greater than zero (p<0.001) using likelihood ratio test. Hence the pairs are not independent. The value of frailty variable estimated is given in Table 1.

Table 1: Estimates of frailty variable for mice in the same litter

Drug (T_1)	Control (T_2)	δ_1	δ_2	C_1	C_2	Z_i	Drug (T_1)	Control (T_2)	δ_1	δ_2	C_1	C_2	Z_i
60	60	0	0	0	0	0.1558	77	79	1	1	1	1	0.9270
60	60	0	0	0	0	0.1558	89	90	1	1	0	1	1.2745
81	69	1	1	1	1	0.3670	102	80	1	1	0	1	1.3389
60	77	0	1	1	1	0.4267	86	94	1	1	0	1	1.3499
67	68	1	1	0	0	0.4560	104	77	1	1	0	0	1.3830
80	73	1	1	0	1	0.4686	103	91	1	1	0	1	1.6442
76	74	1	1	1	1	0.4959	92	102	1	1	0	0	1.6924
73	66	1	1	0	0	0.5160	88	99	1	1	1	1	1.7022
70	77	1	1	0	1	0.5600	91	92	1	1	1	1	1.7225
80	76	1	1	0	1	0.6460	103	84	1	1	0	0	1.7461
76	78	1	1	1	0	0.7655	93	103	1	1	1	1	2.2075

If realization of Z is less than one, then members of the group tend to experience the event earlier. Hence most fragile ones have values less than one. The value of Z increases with increase in lifetime. The estimates of cumulative baseline cause specific reversed hazard function for time to event for drug treated rat and its litter matched control is shown in Figure1 and Figure 2. It can be seen that in drug treated rat, based on cumulative baseline reversed hazard function, tumor appearance is more than death without tumor. But in case of control, based on cumulative baseline reversed hazard function, tumor appearance is less than death without tumor.

Figure 1: Cumulative baseline cause specific RHR for time to event of drug treated rats

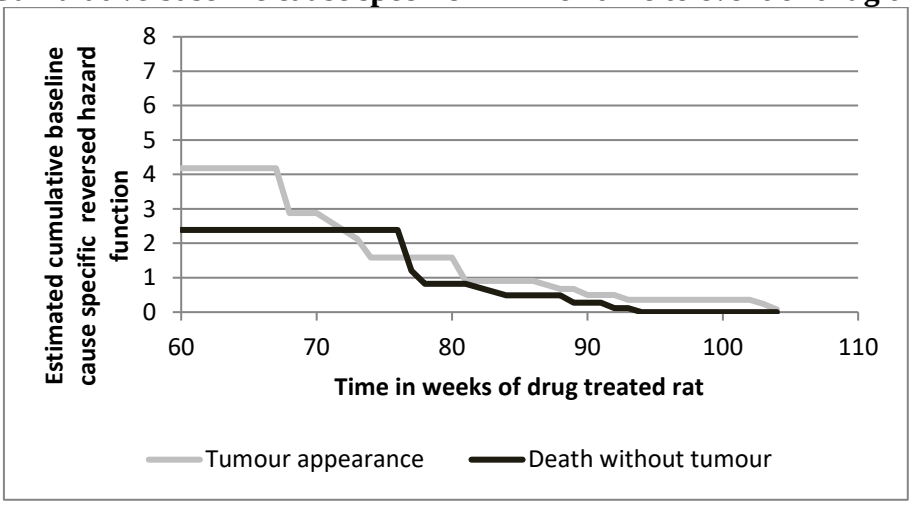
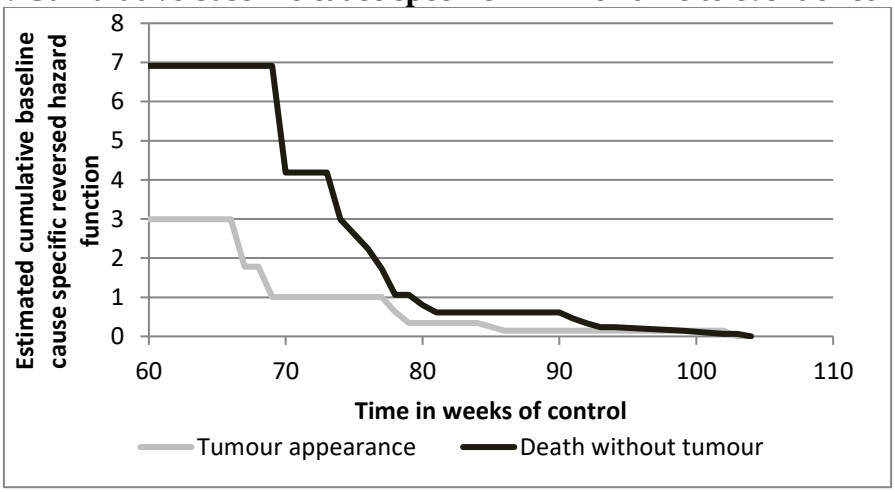


Figure 2: Cumulative baseline cause specific RHR for time to event of control rats



For illustration of the model with covariates, excerpt of the bivariate data set given in McGilchrist and Aisbett (1991), is being used. This data shows the infection times at the point of insertion of catheter for kidney patients using portable dialysis equipment. The observations with value 10 indicate left censored observations. Data for the first two occurrences of infection are given. Let T_1 and T_2 represents the first and second occurrences of infection. Disease types, glomerulo nephritis = (0), acute nephritis = (1), polycystic kidney disease = (2), others = (3), are treated as four different causes for infection. Let C_1 is the variable denoting cause for first occurrence and C_2 is the variable denoting cause for second occurrence. It takes value 0, 1, 2, or 3, depending on the disease type causing infection. Gender is considered as the

covariate for the study, 1=male and 0=female. In the model, it is assumed that regression coefficients associated with covariate gender are different for different types of diseases. Let β_1, β_2, β_3 and β_4 denote the regression coefficients associated with covariate for different disease types glomerulo nephritis, acute nephritis, polycystic kidney disease and others respectively. Estimates of parameters of the model are given in Table 2. Estimates are significant (p<0.001) using likelihood ratio test. Being a male increase the risk of getting infected at earlier time compared to that of female for all disease types. Males with disease type glomerulo nephritis and polycystic kidney disease are more prone to infection. As θ is significant, pairs are not independent. The value of frailty variable estimated is given in Table 3.

Table 2: Estimates of parameters

Parameter	θ	β_1	β_2	β_3	β_4
Estimate	0.0069	-1.1952	-0.4877	-1.1952	-0.9258
Standard error	0.0012	0.0661	0.0485	0.0661	0.0655

Table 3: Estimates of frailty variable for kidney patients

T_1	T_2	δ_1	δ_2	X_1	C_1	C_2	Z
10	16	0	1	1	3	3	0.9929
22	28	1	1	1	3	3	1.0037
447	318	1	1	0	3	3	1.0119
30	12	1	1	1	3	3	1.0024
24	245	1	1	0	3	3	1.0017
511	30	1	1	1	0	0	1.0046
53	196	1	1	0	1	1	1.0091
15	154	1	1	0	0	0	1.0071
10	333	0	1	1	1	1	0.9917
96	38	1	1	0	1	1	1.0002
185	177	1	1	0	3	3	1.0070
292	114	1	1	0	3	3	1.0066
15	108	1	1	0	3	3	0.9871
152	562	1	1	0	2	2	1.0090
13	66	1	1	1	1	1	1.0010
12	40	1	1	0	1	1	0.9996
132	156	1	1	1	0	0	1.0077
34	30	1	1	0	1	1	0.9899
10	25	0	1	0	0	0	0.9914
130	26	1	1	1	0	0	0.9928
27	58	1	1	0	1	1	1.0020
152	30	1	1	0	2	2	0.9909
119	10	1	0	0	3	3	0.9811

If realization of Z is less than one, then members of the group tend to experience the event earlier. Hence most fragile ones have values less than one. The value of Z increases with increase in lifetime. The estimates of cumulative baseline cause specific reversed hazard function for time to first and second occurrence of infection are presented in Figure 3 and Figure 4.

Figure 3: Cumulative baseline cause specific RHR for time to first occurrence of infection

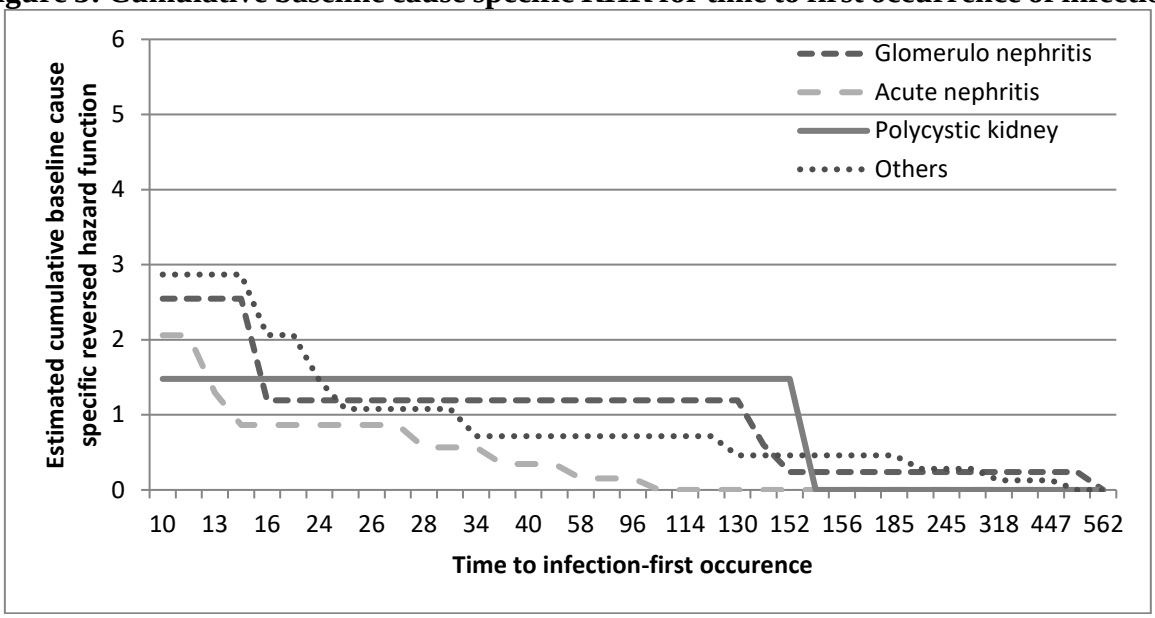
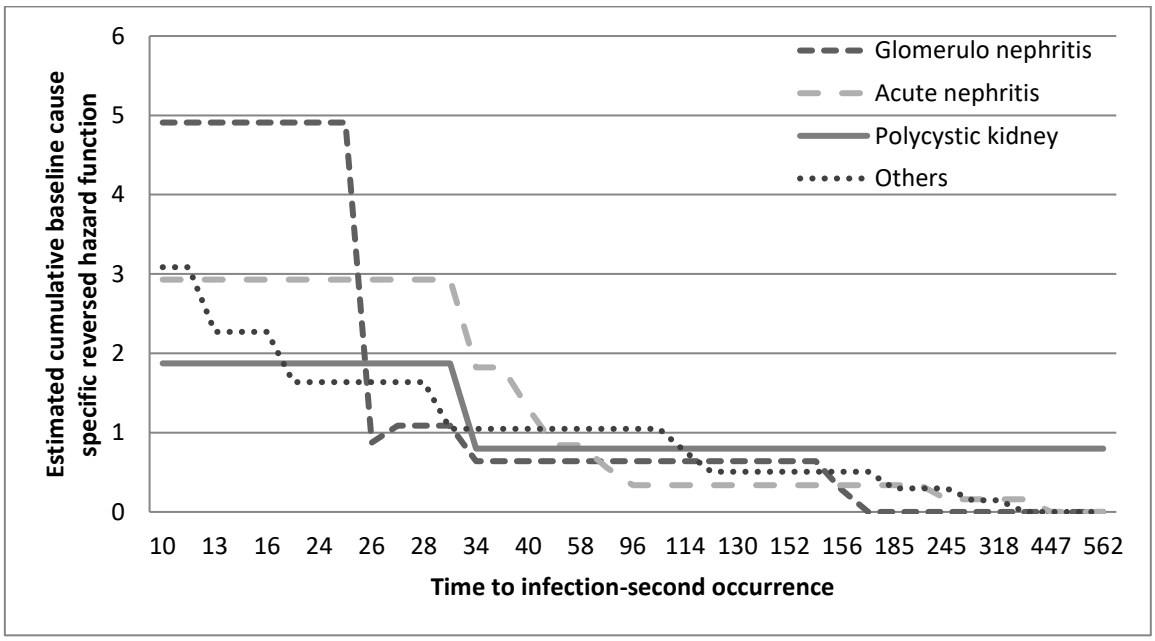


Figure 4: Cumulative baseline cause specific RHR for time to second occurrence of infection



5. Conclusion and Discussion

In this paper, a shared gamma frailty model in terms of cause specific RHR has been introduced for the analysis of competing risks data under left censoring. The gamma distribution with mean one and variance θ is chosen as distribution of the frailty random

variable. The model is discussed with and without the presence of covariates. The parameters of the models were estimated by maximum likelihood method, using EM algorithm, and discussed the properties of the estimators. The proposed models were applied to real life data sets. The data in Mantel and Ciminera (1979) was analyzed for checking the adequacy of gamma frailty distribution with marginal proportional hazard model by Cui and Sun (2004). They obtained the estimate of parameter of gamma distribution as 0.888 which is very close to the value obtained by the present model. Existence of strong association and dependency in litter matched pairs is reported in Anisha (2012). The present model helps to quantify the strength of association in litter matched pairs for left censored data. The data in McGilchrist and Aisbett (1991) were analysed by several authors. The report on analysis by Hanagal and Dabade (2013) and Hanagal (2020) depicts gender as the significant covariate and observes that females are at lower risk. The same result holds for the present model and different regression coefficients are estimated for different causes in the present model. It was observed that more fragile individuals are having realization of frailty variable as less than one and those who experience the event of interest at a later stage are having the value greater than one. The models discussed in Anisha (2012), Cui and Sun (2004) and Hanagal and Dabade (2013) were able to consider only right censored or complete observations. Those models were not dealing with left censored data. So, in order to analyse a left censored family data with multiple type of diseases, shared gamma frailty model in terms of cause specific reversed hazard rates is more appropriate and is recommended.

References

Andersen, P. K., Borgan O., Gill, R. D., and Keilding, N. (1993). *Statistical Models Based on Counting Processes.* Berlin, Springer- Verlag, New York.

Anisha, P. (2012) *Modelling and Analysis of Recurrent Event Data With Multiple Causes*. Ph.D. Thesis. Cochin University of Science and Technology, Kerala, India.

Barlow, R. E., Marshall, A. W., and Proshan, F. (1963). Properties of probability distributions with monotone hazard rate. *The Annals of Mathematical Statistics*, **34**, 375-389.

Bartoszewicz , J., and Skolimowska, M. (2006). Preservation of classes of life distributions and stochastic orders under weighting. *Statistics and Probability Letters*, **76**, 587-596.

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

Chandra, N. K., and Roy, D. (2001). Some results on reversed hazard rate. *Probability in Engineering and Informational Sciences,* **15**, 95-102.

Clayton, D., and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, **148** (**2**), 82-117.

Crowder, M. J. (2001). *Classical Competing Risks*. CRC Press, London.

Cui S., and Sun Y. (2004). Checking for the gamma frailty distribution under the marginal proportional hazards frailty model. *Statistica Sinica,* **14**, 249-267.

Faith K. (2017). The E-Bayesian and hierarchical Bayesian estimations for the proportional reversed hazard rate model based on record values. *Journal of Statistical Computation and Simulation,* **87**(**11**), 2253-2273.

Gupta, R. D and Nanda, A. K. (2001). Some results on reversed hazard rate ordering. *Communications in Statistics - Theory and Methods*, **30**, 2447-2457.

Gupta, R. C., and Wu, H. (2001). Analyzing survival data by proportional reversed hazard model. *International Journal of Reliability and Applications,* **2**(**1**), 1-26.

Hanagal, D. D. (2020) Analysis of kidney infection data using correlated Inverse Gaussian frailty model. *Statistics and Applications*, **18**(**1**), 1-19.

Hanagal, D. D., Arvind, P., and Sankaran, P. G. (2014). Shared frailty model based on reversed hazard rate for left censored data. *Communications in Statistics - Simulation and Computation,* **46**(**1**), 230-243.

Hanagal, D., and Bhambure, S. M. (2017). Shared gamma frailty models based on reversed hazard rate for modeling Australian twin data. *Communications in Statistics - Theory and Methods,* **46**(**12**), 5812-5826.

Hanagal, D. D., and Dabade, A. D. (2013). Compound negative binomial shared frailty models for bivariate survival data. *Statistics and Probability Letters*, **83**, 2507–2515.

Hanagal, D. D., and Pandey, A. (2015). Inverse Gaussian Shared Frailty Models with Generalized Exponential and Generalized Inverted Exponential as baseline distributions. *Journal of Data Science,* **13**, 569-602.

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. (2020). Correlated inverse Gaussian frailty models for bivariate survival data. *Communications in Statistics - Theory and Methods*, **49**(**4**), 845–863.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

Kalbfleisch, J. D., and Lawless, J. F. (1989). Inference based on retrospective ascertainment: an analysis of the data based on transfusion-related AIDS. *Journal of the American Statistical Association,* **84**, 360-372.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure time data*. John Wiley and Sons, New York,

Klein, J. P. (1992). Semi parametric estimation of random effects using Cox model based on EM algorithm. *Biometrics*, **48**, 795-806.

Lawless J. F. (2003). *Statistical Models and Methods for Lifetime data*. Wiley, New York,

Maller, R. A., and Zhou X. (2003). Testing for individual heterogeneity in parametric models for event history data. *Mathematical Methods of Statistics*, **12**, 276–304.

Mantel, N., and Ciminera, J. L. (1979). Use of logrank scores in the analysis of litter-matched data on time to tumor appearance. *Cancer Research*, **39**, 4308-4315.

McGilchrist, C. A., and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461-466.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, **22** (**2**), 712-731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics*, **23**(**1**), 182-198.

Nair, N. U., Sankaran, P. G., and Asha, G. (2005). Characterizations of distributions using reliability concepts. *Journal of Applied Statistical Science*, **14,** 237-241.

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.

Sankaran, P. G., and Anjana, S. (2014). A class of tests for the equality of cause- specific reversed hazard rates in competing risks models. *Journal of the Indian Statistical Association*, **52**, 161-176.

Sankaran, P. G., and Anjana, S. (2016). Nonparametric estimation of cumulative cause specific reversed hazard rates under masked causes of failure. *Journal of Biostatistics and Biometric Applications,* **1**(**2**), 201-211.

Sankaran, P. G., and Gleeja, V. L. (2006). On bivariate reversed hazard rates. *Journal of Japan Statistical Society,* **36**(**2**), 213-224.

Sankaran, P. G., and Gleeja, V. L. (2008). Proportional reversed hazard and frailty models. *Metrika*, **68**, 333–342.

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

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

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16** (**3**), 439-454.

Ware, J. H., and DeMets, D. L. (1976). Reanalysis of some baboon descent data. *Biometrics*, **32**, 459-463.

Ying, Z., and Wei, L. J. (1994). The Kaplan-Meier estimator for dependent failure time observations. *Journal of Multivariate Analysis*, **50**,17-29.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No.2, 2021(New Series), pp 217 – 230

Uniform spacings — a Bird's-Eye View

Sthitadhi Das[1] **and Saran Ishika Maiti**[1]
[1]*Department of Statistics, Visva-Bharati, Santiniketan, India*

Received: 17 July 2020; Revised: 20 December 2020; Accepted: 23 December 2020

Abstract

For a distribution, spacing is defined as the gap between order statistics. In characterization of any distribution, spacings play a pivotal role. Spacing originating from uniform distribution is called uniform spacing. Identical distribution of the first and any k-th spacings for some $k = 2, \cdots, n$ of a sample of size n guarantees a uniform distribution structure of parent population, subject to some underlying conditions. The uniqueness and tractability of uniform spacings propelled them as the focal point of many statistical investigations. However, for the regular statistics practitioners, the theory of spacings remain outside frontiers. In an effort to fill the lacuna, this article presents a succinct and lucid review of related results and applications of uniform spacings.

Key words: Order statistics; Spacings; Uniform distribution; Exponential distribution; Characterization of distribution.

AMS Subject Classifications: 60E05; 62G30; 62E10

1. Introduction

Spacing literally means *gaps* or *distance between two successive points*. In statistics, spacings gauge the distance between two successive order statistics. Let $X_1, X_2, \cdots, X_n$ be a set of independent and identical random variables from a continuous distribution function F with support $[a, b]$. Let the corresponding order statistics be $a < X_{1:n} < X_{2:n} < \cdots < X_{n:n} < b$. By j-th spacing $Y_{j:n}$, we mean

$$Y_{j:n} = X_{j+1:n} - X_{j:n}\,;\; j = 0, 1, 2, \cdots, n. \tag{1}$$

In particular, we assume two marginal order statistics as $X_{0:n} = a$ and $X_{n+1:n} = b$. Thus, in particular, $Y_{0:n} = X_{1:n} - a$, $Y_{n:n} = b - X_{n:n}$. So clearly, for n random variables there would be $(n+1)$ gaps or spacings. Theory of spacings gained steam in many fields of statistics —- goodness of fit tests, statistical estimation theory, reliability analysis, survival analysis and applications to name a few. For inciting readers' interest, a quick flavour on applicability of spacings may be presented from its latest advancement in estimation theory.

In estimating an unknown parameter $\theta \in \Theta$, under a distribution $F_\theta(.)$, Maximum Likelihood Estimation (MLE) is a widely used technique. Moreover, MLE is asymptotically unbiased and efficient under some regularity conditions. As an alternative to MLE some

Corresponding Author: Saran Ishika Maiti
Email: saranishika.maiti@visva-bharati.ac.in

authors proposed estimation process based on spacings. Cheng and Amin (1983) suggested the Maximum Product Spacing Estimator(MPSE) by maximizing

$$G = \left(\prod_{i=0}^{n} Y_{i:n} \right)^{\frac{1}{n+1}}$$

where $Y_{i:n}$ being the i-th spacing. The estimator of θ by maximizing G is known as the MPSE of θ and denoted by $\hat{\theta}_n$. The MPS estimator of the underlying distribution F_θ would be $F_{\hat{\theta_n}}$. MPSE is specially suited to the cases where one of the parameters is an unknown shift origin. This occurs, for example, in the three parameter lognormal, gamma and Weibull models. For such J-shaped distributions, under the condition of shape parameter less than unity, no stationary point can yield a consistent MLE due to unboundedness of likelihood equations. Thus not only global but also local maximum likelihood estimator breaks down. In fact, Johnson and Kotz (1976) recommended in the three parametric gamma/Weibull distribution that MLE should not be used if shape parameter < 2. On the contrary to MLE, MPS estimation too ensures consistent estimators but under much more general conditions than ML estimation. Also Cheng and Amin (1983) showed that MPSE is asymptotically normal and asymptotically efficient as MLE when both exist.

Later Renneby (1984) proposed independently another attractive alternative method as an approximation for the Kullback-Leibler measure of information. This spacing estimator is called Maximum Spacing Estimator (MSE), that can be obtained by maximizing

$$S = \frac{1}{n-1} \sum_{j=0}^{n} ln[(n+1)(F_\theta(X_{j+1:n}) - F_\theta(X_{j:n}))]$$

where $F(X_{0:n}) = 0$ and $F(X_{n+1:n}) = 1$. Maximizing S would furnish MSE of θ. MSE is conducive in tracking the true distribution from the angle of empirical distribution function. In order to address on consistency of MPSE/MSE under much weaker regularity conditions as compared to those required in MLE, Shao and Hahn (1999) can be endorsed. The regularity conditions proposed by them are very general in the sense that they cover most of the known counter examples against the universal appeal of the ML method.

Regarding the robustness of spacing estimator, both MPSE and MSE are minimax robust in Hellinger metric neighbourhoods of the given parametric model. A threadbare discussion about the robustness on general m-th order spacing with respect to certain information measure was accommodated in Ekstrorm (2001). He introduced a class of estimation methods, ensuring asymptotically efficient and robust estimator. MSE/MPSE method includes as a special case.

Following the footsteps of MPSE and MSE, Ghosh *et al.* (2001) introduced a general class of maximum spacing estimator.

$$T(\theta) = \sum_{i=1}^{n} h(n(F_\theta(X_{(i)}) - F_\theta(X_{(i-1)}))$$

where $h : (0, \infty) \to \mathbb{R}$ is a strict convex function. For better understanding a handy example might be referred below.

Example 1: Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be ordered observations from $U(0,\theta)$, $\theta \in (0,\infty)$. The MLE of θ is $\hat{\theta} = X_{(n)}$. Note that here MLE fails to be asymptotic efficient as the regularity conditions for Cramer Rao lower bound of an estimate are not met. On the contrary, generalized spacing estimate of θ is $\hat{\theta_{GSE}} = \frac{n}{n-1}X_{(n)}$ which is obtained by minimising $T(\theta) = \sum_{i=1}^{n} h(n(F_\theta(X_{(i)}) - F_\theta(X_{(i-1)}))$ with $h(x) = -logx$. Both MLE and GSE are consistent but their asymptotic distributions are different.

$$n(\theta - \theta_{MLE}) \xrightarrow{d} Exponential(0,\theta)$$
$$n(\theta - \theta_{GSE}) \xrightarrow{d} Exponential(-\theta,\theta)$$

The two asymptotic distributions have the same variance, but the first has expectation θ while the second has expectation zero and thus is centered better. Moreover,
$E(n(\theta - \theta_{MLE}))^2 \to 2\theta^2$ while $E(n(\theta - \theta_{MSE}))^2 \to \theta^2$. This yields MLE less admissible. Eventually, generalized spacing estimate is also the UMVUE of θ.

Generalized spacing estimator, under the assumption of existence of continuous derivative of p.d.f. with respect to θ, is consistent, asymptotically normal and robust (Ghosh *et al.* (2001)). These estimators are not always explicitly obtainable but can always be computed through numerical methods. Theory of spacings evolved its journey primarily in goodness of fit test and characterization of distributions. Characterization of distribution, through the view point of spacing, stems from the concept of elementary uniform spacings. Due to its comprehensiveness and computational tractability, uniform spacings are considered as benchmarks in spacing theory.

Under the set up, mentioned in equation (1), if F_θ is considered as $uniform(0,1)$ then $\{Y_{j:n}, j \geq 1\}$ is the sequence of uniform spacing variables. Clearly the sum $Y_{0:n} + Y_{1:n} + Y_{2:n} + \cdots + Y_{n:n} = 1$. Due to this linear constraint the random vector $\mathbf{Y} = (Y_{0,n}, Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n})$ has a singular distribution. Moreover, $\mathbf{Y}$ has the joint probability density function as $f_{\mathbf{Y}}(y_{0:n}, y_{1:n}, y_{2:n}, \cdots, y_{n:n}) = n!$ if $y_{i:n} \geq 0\ \forall i$. Also, the distribution of $\mathbf{Y}$ affirms that distribution function is unaltered under any permutation of the co-ordinates. Using this fact, p.d.f of $Y_{i:n}$ can be easily computed. The p.d.f. is $f_{Y_{i:n}}(x) = n(1-x)^{n-1}$, $\forall i$ where $0 < Y_{i:n} < 1$. Clearly, this form is a $beta(1,n)$ distribution.

As the order statistics from any absolutely continuous distribution with distribution function $F(x)$ can be transformed by order preserving probability integral transformation $u = F(x)$ to the order statistics from a uniform distribution, spacing from any continuous distribution can be explained through uniform spacings. This signifies the prime importance of uniform spacings in distribution free interval estimation and many nonparametric applications.

Theory of spacings received its pioneering thrust from Greenwood's (1946) foundational work, although some ground works by Bortkiewicz(1915) and Morant (1920) left a stamp in literature. In those primitive studies, hints of uniform spacings emerged as the distribution of intervals between successive events of Poisson process given the number of events in a specified interval. The basic methodology on the characterization through uniform spacings was documented in the literature by Darling(1953). The most general method on limit theorems of spacings was disseminated by Lecam(1958). A little later, R.Pyke's classic paper 'Spacing' (1965) grabbed the readers' attention wholly on the wide applicability of spacing

theory in the context of distribution free goodness of fit tests as well as characterization problem. Pyke's article, much inclined to theory of uniform spacings, unfurled the idea of construction and limiting theory of spacing at length. Further, Wichura (1968) and Bickel (1969) generalized Le Cam's result and thus provided a concise collection of limit theorems in the context of uniform spacings.

More recently, Ali and Mead (1969), Ahsanullah (1989), Gather *et al.* (1968) and Hamedani and Volkmer (2005) did a good deal of work on spacings. Specifically, Huang *et al.* (1979) established that under the assumptions of continuity and super-additivity the identical distribution of the first and the k-th ($k = 2, \cdots, n$) spacing characterizes uniform distribution.

The current article surveys the theoretical developments in the context of uniform spacings that exist in literature so far. This review, mostly, presents a bunch of useful results without delving into the intricate mathematical exposition. Of most interest of this review is results on uniform spacing as they arise in characterization of distributions. Authors' objective is to popularize the results of uniform spacing in characterization of distribution, crafted under a milder tone of discussion. The rest of the article is outlined as follows. Section 2 contains main results related to characterization, based on uniform spacing. Some preliminary ideas on uniform characterization are also included. Additionally, results on ordered uniform spacings are mentioned. In third section, results on asymptotic properties on uniform spacing are documented. Finally, a short conclusion ends the article.

2. Main Results

2.1. Genesis of uniform spacings

This subsection discusses some preliminaries that explore a few basic construction techniques of uniform spacings. Recalling the setting, already mentioned in Introduction, $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$ being the spacings formed from uniform$(0, 1)$, the following results are presented.

Result 1: As $f_{Y_{i:n}}(x) = n(1-x)^{n-1}$ for $0 < Y_{i:n} < 1$ and $f_{Y_{i:n},Y_{j:n}}(x, y) = n(n-1)(1-x-y)^{n-1}$ standard technique nails down to deduce

$$E(Y_{i:n}) = \frac{1}{n+1},\ V(Y_{i:n}) = \frac{n}{(n+1)^2(n+2)},\ Cov(Y_{i:n}, Y_{j:n}) = -\frac{1}{(n+1)^2(n+2)}$$

Higher order moments of uniform spacings were derived by Greenwood (1946); Renyi(1953).

Result 2: For positive constants v_i ; $i = 1, 2, \ldots, r$ with $r \leq n$ and $\sum_{i=1}^{r} v_i \leq 1$ the joint survival function of $Y_{1:n}, Y_{2:n}, \cdots, Y_{n:n}$ is given by

$$Prob(Y_{1:n} > v_1 \cdots Y_{r:n} > v_r) = (1 - \sum_{i=1}^{r} v_i)^{n-1}.$$

This leads that under $n \to \infty$,

$$Prob(nY_{1:n} > v_1 \cdots, nY_{r:n} > v_r) = \prod_{i=1}^{r} \{exp(-v_i)\}, v_1, \cdots, v_r > 0.$$

So the limiting distribution of $nY_{i:n}$ is Exponential(1).

The following theorem exudes a connective relation for the characterization of uniform spacings variables through exponential distribution.

Theorem 1: Let $E_1, E_2, \cdots, E_n, E_{n+1}$ be a sequence of iid exponential variables. $Y_{1:n}, \cdots, Y_{n+1:n}$ is distributed as $\frac{E_1}{\sum_{i=1}^{n+1} E_i}, \cdots, \frac{E_{n=1}}{\sum_{i=1}^{n+1} E_i}$. Furthermore, let G_{n+1} be a gamma variable with parameter $(n+1)$. Then $Y_{1:n}G_{n+1}, \cdots, Y_{n+1:n}G_{n+1}$ is distributed as $E_1, E_2, \cdots, E_{n+1}$.

As we know that sum of i.i.d. exponentials follow Gamma distribution. Ratio of an exponential random variable and Gamma variable (of which numerator is a member) lies between 0 to 1, Tacitly, from basic sampling distribution theory each ratio $\frac{E_i}{\sum_{i=1}^{n+1} E_i}$ follows a $beta(1, n)$ distribution. Theorem 1 is an important theorem as a lot of results follow from it. Any standard, degree course text book would be a sufficient resource of those basic results.

Further, Theorem 1 could be applied in order to generate spacing variables from Uniform(0,1) directly. By this we mean that it is not necessary to generate $U_1, U_2, \cdots, U_n$ first and then apply some sorting method, rather generating a bunch of uniform spacing variables at first hand. First we generate iid exponential random variables $E_1, E_2, \cdots, E_n$. Next we would compute the sum of all these random variables G. Then using Theorem 1 we could run a loop of continuation as $U_{(j-1)} + \frac{E_j}{G} = U_{(j)}$.

2.2. Results on characterization of uniform spacings

Characterization of any distribution is a certain distributional property of statistic/statistics that uniquely ascertains the probability structure of underlying distribution. Characterization of uniform distribution can be studied via spacings. An insightful investigation was done by Huang *et al.* (1979) where they asserted the identical distributions of two or more spacings characterize an uniform parent distribution. Keeping in mind that all uniform spacings are identically distributed as $beta(1, n)$, one can characterize the parent cdf uniquely with the aid of some distributional properties of the spacings. Naturally the question arises if identical distributions of two or more spacings are sufficient to characterize a uniform distribution. To address this question Huang *et al.* (1979) assumed that under the condition of (i) continuity, (ii) super-additivity (or, sub-additivity), and (iii) boundedness of support of F identical distributions of $Y_{1:n}$ and $Y_{k:n}$ for $k = 1, \ldots, n$ characterize a uniform distribution. Before unveiling the crux of the result, let us have a sneak peek on what super-additivity (or, sub-additivity) of F is.

Definition 1: A distribution function F is super-additive for all $x, y \in \mathbb{R}$ and $x, y, x + y \in support(F)$ if

$$F(x + y) \geq (\leq)\ F(x) + F(y).$$

The final statement of the result comes as follows.

Result 3: Continuity and super-additivity of F, under which $Y_{0;n}$ and $Y_{k;n}$ for some $k = 1, 2, \ldots, n$, have identical distribution, characterize a uniform distribution.

Instead of super-additivity, the bounded support of F too leads to the similar charac-

terization.

Result 4: If F has bounded support and continuous density, then the identical distribution of $Y_{0;n}$ and $Y_{k;n}$ characterizes a uniform distribution.

Next lemma is conducive in establishing some properties of F which emerge as immediate consequences of the identical distribution of $Y_{0;n}$ and $Y_{k;n}$ for some $k = 1, \ldots, n-1$.

Lemma 1: If F is continuous and the spacings $Y_{0;n}$ and $Y_{k;n}$ for some $k = 1, \ldots, n$ have identical distribution, then (a) F(0)=0, (b) if $F(x_1) = F(x_2)$ for some $0 < x_1 < x_2$, then $F(x_1) = 1$, and (c) $F(x) > 0$ for all positive x.

Remark 1: Using this lemma, we can conclude that for some $k = 1, \ldots, n-1$ when $Y_{0;n}$ and $Y_{k;n}$ have identical distribution, the support of F is either a finite interval $[0, a]$ or an infinite one $[0, \infty)$.

The forthcoming results are going to be discussed upon the condition that support of F is either finite or some other stronger conditions implying the support to be finite.

Remark 2: (1) If F is *sub-additive* and *support*(F) is finite, then F is uniform on $[0, a]$, for some $a > 0$.

(2) Neither the identical distribution of $Y_{0;n}$ and $Y_{k;n}$ nor that of $Y_{k;n}$ and $Y_{j;n}$ for some $1 \leq k < j \leq n-1$ solely guarantees that the parent distribution is uniform distribution on $[0, a]$, $a > 0$.

Clearly, all symmetric distribution functions F will satisfy the identical distribution of $Y_{0;n}$ and $Y_{n;n}$ as well as the identical distribution of $Y_{k;n}$ and $Y_{n-k;n}$, $k = 1, \ldots, n-1$. This type of identical distribution of spacings has been utilized in characterization theory by other authors as well. Some of the profound works include characterizations of exponential and geometric distribution by Puri *et al.* (1970), only exponential by Ahsanullah (1976) and the fairly recent work on general class of continuous distributions by Mirakhmedov *et al.* (2013).

Another meticulous finding of Huang *et al.* would surely grip the researchers' attention.

Theorem 2: Let F be a continuous distribution function of a bounded variable X. $Y_{0;n}$ and $Y_{1;n}$ have identical distribution. Moreover, if F has a density f which is continuous on $(0, a)$ with finite limits $f(0+)$ and $f(a-)$, then F is uniform on $[0, a]$.

Remark 3: As the byproduct of Theorem 8, one can list few remarks as mentioned below.

(1) Theorem does not state that for any arbitrary k, $Y_{0;n}$ and $Y_{k;n}$ have identical distribution.

(2) If F has bounded support $[0, a]$ and the regularity conditions mentioned in the theorems hold then the identical distribution of $Y_{n;n}$ and $Y_{n-1;n}$ leads to the fact that F has uniform distribution on $[0, a]$.

(3) However in Theorem 2 some smoothness conditions on F is required, otherwise it might be misleading. For example, if a random sample $X_1, \ldots, X_n$ is drawn from a Bernoulli

distribution with the probability of success $n/(n+1)$, then $Y_{0;n}$ and $Y_{1;n}$ have identical distribution, even though the parent distribution is not uniform.

Again taking cue from Huan *et al.* (1979) two more further explorations are mention-worthy in characterization theory by uniform spacing.

Result 5: (Ahsanullah (1989))
Let F be absolutely continuous with density function f, $F(0) = 0$, $F(1) = 1$, and f is monotonic on $(0, 1)$. Then $F \sim U[0, 1]$ iff there exists a pair (r, n), $2 \leq r \leq n$, such that

$$X_{r;n} - X_{r-1;n} \sim X_{r-1;n} - X_{r-2;n}.$$

Let F be absolutely continuous, symmetric, either super-additive or sub-additive, $F(0+) = 0$, $F(1) = 1$. Then $F \sim U[0, 1]$ iff $X_{n;n} - X_{1;n} \sim X_{n-1;n}$ for some $n \geq 2$.

Result 6: (Madreimov, Petumin (1983))
Let F be continuous and let $X_1, \ldots, X_n, X \sim F$ be independent random variables. Then

(1) $F \sim U[0, 1]$ iff $E(X_{n,n} - X_{i,n}) = Probability(X \in (X_{i,n}\ ,\ X_{n,n}))$ for all $i \in \mathbb{N}$ and $n \geq i$.

(2) $F \sim U[0, 1]$ iff there exists a pair $(i, j),\ 1 \leq i < j \leq n$, such that

$$E(X_{i,n}) = P(X \in (X_{j-1,n}\ ,\ X_{j,n}))\ for\ all\ n \geq 2.$$

Characterization of the uniform distribution is further extended by Hamedani and Volkmer (2005) in the probability structure of more than one step spacing variables.

Result 7: Let us imagine $(s - r)$ th step spacing, (unlike the distance between consecutive order statistics) $X_{s:n} - X_{r:n},\ s > r$. If F is uniform$(0, a)$, $X_{s:n} - X_{r:n} \sim X_{s+k:n} - X_{r+k:n}\ \forall 1 \leq r < s < s + k \leq n$.

Special case of the above result appears when $s = r + 1$ and $k = 1$.

Theorem 3: Let X be a positive-valued random variable having an absolutely continuous cdf F. If the corresponding pdf f is monotone on support(F) which is an interval, and the above result holds on $s = r + 1$, $k = 1$ for some r, then X has a uniform distribution on $support(F)$.

Remark 4: (1) Sometimes Result 7 and the consequent theorem might hold for some $s = r + 1$ and $k = 1$ for some r but it still does not guarantee that F is uniformly distributed. For example, whenever f is symmetric on $[a, b]$ (*i.e.* $f(x) = f(b - a - x)$) for all x, then Theorem 3 holds for the choices $r = k = 1$ and $n = 3$. That is, the monotonicity of f on $[a, b]$ is important to satisfy the condition of the theorem.

(2) If the assumption that support(F) fails to be an interval the above theorem fails as well. Here is a thought provoking example from Hamidani *et al.* Consider the following

pdf

$$f(x) = \begin{cases} \frac{3}{2} & \text{if } x \in \left[0, \frac{1}{3}\right] \bigcup \left[\frac{2}{3}, 1\right] \\ 0 & \text{otherwise} \end{cases}$$

Then f is monotone on support(F)=$\left[0, \frac{1}{3}\right] \bigcup \left[\frac{2}{3}, 1\right]$. Since f is symmetric as $f(1-x) = f(x)$, Result 7 holds for a particular choice, say, $r = k = 1$ and $n = 3$, but clearly F is not $U(0, 1)$.

2.3. Results on ordered uniform spacings

Arranging the spacings $Y_{1:n}, \cdots, Y_{n+1:n}$ in increasing order we obtain ordered uniform spacings. In Levy(1939); Renyi(1953); Barton and David (1956); Pyke (1965) and Devroye (1981) distributional and asymptotic results on smallest and largest uniform spacings, are discussed at length. Relatively little was done in the context of general ordered uniform spacings. A concrete idea on ordered uniform spacings is developed in Bairamov (2010).

Let us denote the ordered spacings as $\Delta_{0:n} < \Delta_{1:n} < \cdots < \Delta_{n+1:n}$ where $\Delta_{i:n}$ being the i-th largest spacing. Tacitly, for $(n+1)$ number of spacings we would have same number of ordered spacings.

Result 8: $(n-k-1)(\Delta_{k+1:n} - \Delta_{k:n}) \stackrel{d}{=} \Delta_{1:n} \quad (k = 0, \cdots, n)$ where $\stackrel{d}{=}$ means that the statistics are distributionally same (Pyke (1965)).

In particular, distribution of the k-th $(1 \leq k \leq n+1)$ ordered uniform spacing $\Delta_{k:n}$ is deduced by Bairamov *et al.* (2010). Let the domain of uniform distribution $(0, 1)$ be presented as the sum of non overlapping intervals

$$(0, 1) = I_{1,n} \cup I_{2,n} \cup \cdots \cup I_{n+1,n}$$

where $I_{1,n} = (0, \frac{1}{n+1})$ and $I_{m,n} = (\frac{1}{n+3-m}, \frac{1}{n+2-m})$.

Theorem 4: The distribution of the k-th $(1 \leq k \leq n+1)$ ordered uniform spacing is presented by

$$P\{\Delta_{k:n} \leq x\} = 0 \ \ (x < 0)$$
$$P\{\Delta_{k:n} \leq x\} = 1 \ \ (x(n+2-k) \geq 1)$$

and for $x \in I_{m,n}, m = 1, 2 \cdots, k$

$$P\{\Delta_{k:n} > x\} = (-1)^{k-1}(n+1)\binom{n}{k-1} \sum_{i=m}^{k} \frac{(-1)^{i-1}}{(n+2-i)} \binom{k-1}{i-1} (1 - x(n+2-i))^n.$$

As a corollary of the last theorem the distribution of minimal ordered spacing and maximal spacing can be presented.

Corollary 1: The distribution of the minimal spacing is given by

$$P\{\Delta_{1,n} > x\} = (1 - x(n+1))^n, \ x \in I_{1,n}$$

Simultaneously, the distribution of maximal spacing is given by

$$P\{\Delta_{n+1,n} > x\} = (-1)^n(n+1)\sum_{i=m}^{n+1} \frac{(-1)^{i-1}}{n+2-i}\binom{n}{i-1}(1-x(n+2-i))^n$$

.

Theorem 4 allows the readers to compute the expectation of k-th ordered uniform spacing which is $\frac{1}{n+1}\sum_{i=n+2-k}^{n+1}\frac{1}{i}$ $(k = 1, \cdots, n+1)$.

3. Limiting Results on Uniform Spacings

Limiting theories for spacings depicts some of the more interesting results. Here, we present several results from the limiting theory of uniform spacings chronologically. Mostly every case affirms on the asymptotic distribution as normal distribution.

Result 9: Levy (1939) obtained the limiting distributions of the maximal spacing $\Delta_{n+1,n} =$ maximum ordered spacing from the uniform distribution on $[-1, 1]$ as

$$P\{\frac{n\Delta_{n+1,n}}{\log n} \leq x\} \to exp(-exp(-x)); \;\; x \in \mathbb{R}$$

Result 10: Devroye (1981) established that

$$\lim_{n\to\infty} \sup \left[\frac{n\Delta_{n+1,n}}{2\log\log n}\right] = 1 \; a.s.$$

Result 11: Let $F(x)$ be a continuous distribution function. If $X_{1:n}, \ldots, X_{n:n}$ is an ordered sample of n values from the population whose distribution function is $F(x)$ then the random variable

$$\omega_n = \frac{1}{2}\sum_{i=1}^{n}\left|F(x_{i:n}) - F(x_{i-1:n}) - \frac{1}{n+1}\right|$$

is asymptotically normally distributed with mean $E(\omega_n)$ and variance $var(\omega_n)$ (Sherman (1947)), *i.e.*, the standardized random variable

$$\frac{\omega_n - E(\omega_n)}{\sqrt{var(\omega_n)}}$$

approaches towards a Standard Normal variate as $n \to \infty$.

Result 12: Kimball (1947) proved the asymptotic normal distribution of

$$\alpha_n = \frac{1}{2}\sum_{i=1}^{n}\left(F(X_{i:n}) - F(X_{i-1:n}) - \frac{1}{n+1}\right)^2$$

which is also a measure of deviation from uniform spacing.

Result 13: Moran (1972) considered a similar statistic

$$\beta_n = \sum_{i=1}^{n+1} \left(F(X_{i,n}) - F(X_{i-1,n})\right)^2$$

and proved that β_n is asymptotically normal.

All of these results on asymptotic theory of uniform spacings could serve as handy tools for constructing the goodness of fit tests as the distributions are asymptotically normal. For the exact expressions of expectation and variance, readers are recommended to go through the respective articles. Recently, Eryilmaz and Stepanov (2008) studied runs based uniform order statistics and developed even more complex limit results related to uniform spacings from the results obtained for runs. One may be interested in the asymptotic behavior of ordered uniform spacing $\Delta_{k,n}$. A solid and meticulous discussion is found in Bairamov *et al.* (2010). In this article we restrain from mentioning all those critical results as that might spoil the flow of simplicity of the article. Instead we can concentrate our attention on the asymptotic behavior of the expectation of $\Delta_{k,n}$ (k th ordered uniform spacing) which is relatively simpler.

Result 14: (1) For finite $k \geq 1$ and $n \to \infty$,

$$E(\Delta_{k:n}) = O\left(\frac{1}{n^2}\right) \to 0.$$

(2) For finite $k \geq 1$ and $n \to \infty$ we have

$$E(\Delta_{n+2-k:n}) \sim \frac{logn}{n} \to 0.$$

(3) If $k = k_n \to \infty$ such that $k_n = o(n)$, then

$$E(\Delta_{k_n:n}) \sim \frac{k_n}{n^2} \to 0.$$

(4) If $k = k_n \to \infty$ such that $k_n = o(n)$, then

$$E(\Delta_{n+2-k_n:n}) \sim \frac{log(n/k_n)}{n} \to 0.$$

The recent trend in characterization through spacings is escalated by investigations in probability distribution of adjacent spacings. By adjacent spacing we mean the neighborhood around an order statistic, *i.e.*, the points encapsulated in $(X_{k:n} - d, X_{k:n})$ or $(X_{k:n}, X_{k:n} + d)$ where the d may or may not be dependent on n. Pakes and Steutel (1997); Balakrishnan and Stepanov(2005); Dembinska *et al.* (2007); Dembinska and Balakrishnan (2010) are few worth references. Further Nagaraja *et al.* (2014) accelerated this route by discussing joint limiting distribution of adjacent spacings $(Y_{k:n}, \cdots, Y_{k+r:n})$ and $(Y_{k:n}, \cdots, Y_{k-s:n})$ around three types of order statistics in particular – central, intermediate or an extreme order statistic. When $n \to \infty$, these three different scenarios arise and (i) Central case where $\frac{k}{n} \to p,\ 0 < p < 1$,

(ii) Intermediate case where $k, n-k \to \infty$ and $\frac{k}{n} \to 0$, (iii) Extreme case where k or $n-k$ is kept fixed. Borrowing the knowledge of extreme value theory they showed that in the first two cases $(r+s)$ spacings converge weakly to a batch of i.i.d. standard exponential random variables but in extreme case, this weak convergence would hold only in the domain of attraction of Gumbel (heavy tail distribution) or Weibull type (short ended, finite end point distribution).

4. Conclusion

The main purpose of this review article is to elucidate some instructive results related to uniform spacings arising in the context of characterization theory, without much dipping down to mathematical complexity. Other than the usual uniform and exponential spacings, some relevant investigations were done by several authors on the characterization of the Gamma, Normal and Weibull distributions as well as on some discrete distributions like Geometric, Poisson, Negative Binomial distributions etc. Still characterization of uniform distribution through spacings draws fundamental importance in a wide variety of fields. Apart from characterization theory; spacing, specially uniform spacing, is applicable in few other fields as well e.g. distribution-free goodness of fit test, information theory, time series analysis etc. A couple of topics of interest might be worth mentioning here.

Goodness of fit test is used to verify if the distribution function is equal to a specified one. By probability integral transformation, any specified continuous distribution function can be converted to uniform(0,1). Thus the null hypothesis of interest boils down to $H_0 : F(x) = x$, $0 \le x \le 1$. To test this H_o, Greenwood (1946) introduced a statistic on the basis of sum of squares of disjoint uniform spacings. The statistic, called Greenwood statistic, is $G = \frac{1}{n}\sum_0^{n-1}(nY_{(i:n)})^2$. By virtue of Pitman asymptotic efficiency (AE) Greenwood test is proved as optimal among the goodness of fit tests within the class of symmetric tests.

All the more, Greenwood statistic could be generalized by $G_n = n^{-1}\sum_{i=0}^{n-1} h(nY_{(i:n)})$ where $h(.)$ is a function that satisfies some mild regularity conditions. Clearly, the choice of $h(x) = x^2$ would turn it to Greenwood's original statistic. $h(x) = x^r$ for $r > 0$ was proposed by Kimball (1950) who obtained the limiting distribution of statistic under H_0. Another popular choice of $h(.)$ is $h(x) = logx$, suggested by Darling (1953).

Periodogram analysis is an effective device in frequency analysis of time series. For a stationary, random time series periodogram ordinates are exponentially distributed and independent. Actually ordinates have the same distribution as the uniform distributions of the spacings. Thus to test for the peak , *i.e.*, the largest ordinate of periodogram test statistic might be considered as largest uniform spacing (For details see Durbin (1960)). Also, in the context of distribution of serial/auto correlation coefficients; any order of serial correlation coefficient can be put in the form of linear functions of spacings. The joint distribution of several linear functions of the spacings was derived by Watson (1956) which could explain out the distribution of serial correlations.

In applied economics, auction theory holds current trend of attraction. Motivated by the upsurge of auctions in online advertisements, like auction through eBay and Amazon, the query on expected revenues in auctions is quite of interest in recent years. Uniform

spacings (or spacings as a whole) might come out as an effective tool in such stochastic auction theory where the following spacings–$Y_{2:n}$ and $Y_{n:n}$ would represent auction rents in buyers' auction and reverse auction in the second-price business auction under identical bids. One might be inquisitive on stochastic modeling of second price (explained by second order uniform spacing) or if the expected revenues depending on the number of bidders.

So far what we presented here is all about univariate spacings. An intrinsic question might trigger regarding the exact distribution of bivariate/multivariate uniform spacings. There are myriad examples in which samples are drawn from bivariate/multivariate set-up for which it is pertinent to study the spacings of the observations. As a stepping stone, one can start with multivariate uniform distribution and investigate on multivariate uniform spacings accordingly. Barton and David (1962) studied the distribution on spacings computed on the random points drawn on the two-dimensional plane but still now number of organized works in multivariate spacings is almost nil, probably due to its degree of computational difficulty.

Acknowledgement

This research work has been made under the financial support from the major research project (EMR/2017/005421) sanctioned by Science and Engineering Board(SERB), DST, Govt. of India. Both of the authors are indebted to the Editor, unknown referee and Prof. Jyotirmoy Sarkar, IUPUI,USA for their constructive suggestions which led to much improvement of this article.

References

Ahsanullah, M. (1976). On a characterization of the exponential distribution by order statistics. *Journal of Applied Probability*, **13**, 818-822.

Ahsanullah, M. (1989). On characterizations of the uniform distribution based on functions of order statistics. *Aligarh Journal of Statistics*, **9**, 1-6.

Ali, Mir M., and Mead E. R. (1969). On the distribution of several linear combinations of order statistics from the uniform distribution. *Bulletin of the Institute of Statistical Research and Training*, **3**, 22-41.

Bairamov, I., Alexandre B., and Stepanov A. (2010). Limit results for ordered uniform spacings. *Statistical Papers*, **51(1)**, 227-240.

Balakrishnan, N., Pakes, A. Gn., and Stepanov, A. (2005). On the number and sum of near-record observations. *Advances in Applied Probability*, **37(3)**, 765-780.

Barton, D. E., and F. N. David (1956). Some notes on ordered random intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*, **18(1)**, 79-94.

Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, **40(1)**, 1-23.

Bortkiewicz, L. Von (1915). *Ü*ber die Zeitfällinger Ereignisse. *Bulletine of International Statistical Institute*, **20(2)**, 30-111.

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

Darling, Donald A. (1953). On a class of problems related to the random division of an interval. *The Annals of Mathematical Statistics*, **24**, 239-253.

David, H. A. and H. N. Nagaraja (2003). *Order Statistics*, John Wiley & Sons, New York.

Dembinska, A., Stepanov, A. and Wesolowski, J. (2007). How many observations fall in a neighborhood of an order statistic? *Communications in Statistics-Theory and Methods*, **36(5)**, 851-867.

Dembinska, A. and Balakrishnan, N. (2010). On the asymptotic independence of numbers of observations near order statistics. *Statistics*, **44(5)**, 517-528.

Devroye, Luc (1981). Laws of the iterated logarithm for order statistics of uniform spacings. *The Annals of Probability*, **9**, 860-867.

Durbin, J. (1960). Some methods of constructing exact tests. *Biometrika*, **48(1-2)**, 41-65.

Ekstrom, M. (2001). Consistency of generalized maximum spacing estimates, *Scandinavian Journal of Statistics*, **28(2)**, 343-354.

Eryilmaz, S., and Stepanov, A. (2008). Runs in an ordered sequence of random variables. *Metrika*, **67(3)**, 299-313.

Gather, U., Udo Kamps., and Schweitzer, N. (1998). Characterizations of distributions via identically distributed functions of order statistics. *Handbook of Statistics*, **16**, 257-290.

Ghosh, K., and Rao Jammalamadaka, S. (2001). A general estimation method using spacings. *Journal of Statistical Planning and Inference*, **93(1-2)**, 71-82.

Greenwood, M. (1946). The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, **109(2)**, 85-110.

Hamedani, G. G., and Volkmer, H. W. (2005). Certain characterizations of the uniform dstribution. *Metrika*, **61(2)**, 117-125.

Huang, J. S., Arnold B. C., and Ghosh, M. (1979). On characterizations of the uniform distribution based on identically distributed spacings. *Sankhya: The Indian Journal of Statistics, Series B*, 109-115.

Johnson, N. J., and Kotz, S. (1970). *Continuous Univariate Distributions-2.* Houghton Mifflin, Boston.

Kimball, Bradford F. (1947). Some basic theorems for developing tests of fit for the case of the non-parametric probability distribution function. *The Annals of Mathematical Statistics*, **18(4)**, 540-548.

LeCam, L. (1958). Un théoreme sur la division d'un intervalle par les points pris au hasard. *Publications de Institut de Statistique de l'Université de Paris*, **7**, 7- 16.

Levy, P. (1939). Sur la division d'un segment par des points choisis au hasard. *Comptes rendus de l'Académie des Sciences* , **208**, 147-149.

Madreimov, I., and Petunin, Y. I. (1983). A characterization of the uniform distribution with the aid of order statistics. *Theory of Probability and Mathematical Statistics*, **27**, 105-110.

Mirakhmedov, Sherzod M., and Rao Jammalamadaka, S. (2013). Higher-order expansions and efficiencies of tests based on spacings. *Journal of Nonparametric Statistics*, **25(2)**, 339-359.

Moran, P. A. P. (1947). The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, **9(1)**, 92-98.

Morant, G. M. (1920). On random occurrences in space and time, when followed by a closed interval. *Biometrika*, **13**, 309-337.

Pakes, A. G., and Steutel, F. W. (1997). On the number of records near the maximum. *Australian Journal of Statistics*, **39(2)**, 179-192.

Puri, P. S., and Herman, R. (1970). A characterization based on the absolute difference of two iid random variables. *The Annals of Mathematical Statistics*, **41(6)**, 2113-2122.

Pyke, R. (1965). Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, **27(3)**, 395-436.

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

Renyi, A. (1953). On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, **4(3-4)**, 191-231.

Shao, Y., and Hahn, M. G. (1999). Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Annals of Institute of Statistical Mathematics*, **51**, 31-49.

Sherman, B. (1950). A random variable related to the spacing of sample values. *The Annals of Mathematical Statistics*, **21(3)**, 339-361.

Watson, G. S. (1956). On the joint distribution of the circular serial correlation coefficients. *Biometrika*, **43(1/2)**, 161-168.

Wichura, M. J. (1968). On the weak convergence of non-Borel probabilities on a metric space. Unpublished Dissertation, Columbia University, USA.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 231-240

Estimating the Area under the ROC Curve in the Framework of Lindley Centered Distributions

Balaswamy, S.[1] and Vishnu Vardhan, R.[2]
[1]*Department of Statistics, Indira Gandhi National Tribal University, Madhya Pradesh*
[2]*Department of Statistics, Pondicherry University, Puducherry*

Received: 23 May 2020; Revised: 22 December 2020; Accepted: 27 December 2020

Abstract

In the context of ROC curve analysis, the most widely used ROC form is the Binormal ROC curve. But due to the theoretical structures and distributional assumptions, many more bi-distributional ROC curve models have been proposed over the years. In this paper, an attempt has been made to overcome few limitations of ROC curve that emanated from exponential distribution. To address this, we have considered different forms of Lindley distributions and taking its mathematical advantages and mathematical flexibility, three new ROC curves are proposed. The proposed methodology is supported by simulation studies.

Key words: ROC curves; Lindley distribution; Power Lindley; Extended Lindley distributions and AUC.

1. Introduction

In statistical theory and practice, classification problems have gained lot of attention by many researchers in solving problems that are trivial as well as complex. Identification of class label is one of the major objectives in classification, for which several statistical techniques have been developed and proposed. Basing on the prominence and demand to handle such problems, those varieties of statistical tools have emerged and were brought under the hub of Statistical Decision Theory (SDT). The common problem of interest in classification is in allocating an individual or object to one of the predefined groups (or populations) by using a threshold. These problems were addressed by using a performance tool namely, Receiver Operating Characteristic (ROC) Curve, which evolved during World War II.

ROC Curve analysis was first presented to Psychologists by Tanner and Swets (1954), who brought out the concept from the Theory of Signal Detectability (TSD), which was introduced by Peterson *et al.* (1954) during World War II for analyzing the radar signals to detect enemy objects in battlefield *i.e*., identifying the signal as signal and noise as noise. Its expansion to other fields was prompt, for instance, in Psychology it was used to study the perceptual detection of stimuli (Swets, 1996). In medicine, one of the earliest applications was proposed by Lusted (1971), in which he postulated that to measure the worth of a diagnostic test, one must measure the performance of observers with the test and argued that ROC Curve provides an ideal means of studying observer performance.

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

Swets and Pickett (1982) noted two other key features of ROC Curves that make them ideal for studying diagnostic tests. First, the curves display all possible cut points and thus supply estimates of the frequency of various outcomes (*i.e.*, true positives, true negatives, false positives, and false negatives) at each cut point. Second, the curve allows the use of previous probabilities of condition, as well as calculations for the benefits of correct and incorrect decisions, to determine the best cut point for a given test in a given set up. Suppose the outcome S of a medical test is a measurement on a continuous scale (score), then there exists a threshold t of the test score, which can be used to classify subjects. For instance, a person with $S \leq t$ may be classified as healthy (normal or benign), otherwise as diseased (abnormal or malignant). Basing on the above classification, a 2 × 2 contingency table, namely the "*confusion matrix*" can be generated with four possible states, *viz*., True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Relatively few diagnostic tests correctly classify all subjects tested as diseased "D" (abnormal) or healthy "H" (normal). Sometimes, the threshold considered for classification, classifies few healthy ones as diseased and vice versa. This wrong classification leads to the terminology of False Positive Rate (FPR) and False Negative Rate (FNR). The probabilistic definitions pertaining to four possible states are given below:

i. The probability that an individual from D is correctly classified.
   True Positive Rate, TPR = $P(S > t \mid D)$ (*Sensitivity*)
ii. The probability that an individual from H is misclassified.
   False Positive Rate, FPR = $P(S > t \mid H)$ (1-*Specificity*)
iii. The probability that an individual from H is correctly classified.
   True Negative Rate, TNR = $P(S \leq t \mid H)$ (*Specificity*)
iv. The probability that an individual from D is misclassified.
   False Negative Rate, FNR = $P(S \leq t \mid D)$. (1-*Sensitivity*)

These four probabilities describe the performance of the test at this cutoff. It is important to note that all the intrinsic measures correspond to a given value t. As t changes these measures change. One of the problems of interest is to determine such t which optimizes one or more intrinsic measures, usually referred as *"optimal cutoff"*. With these probabilistic definitions of intrinsic measures, the ROC Curve can be defined as $ROC(t) = f(FPR(t), TPR(t))$. This means that the ROC Curve is generated by a set of pairs of FPR and TPR, which are obtained at every threshold point, that are actually observed test scores. So, each test score will act as a cutoff, which in turn generates the co-ordinates (FPR, TPR). ROC Curve is a tradeoff between FPR and TPR at every t. A test is said to be a *better* one, if it has maximum TPR and a *reasonably low* FPR.

An assumption in ROC curve is that the test scores of diseased populations will be greater than that of the healthy populations. For instance, if the Creatinine (one of the indicators of severe kidney impairment) levels are 5.0 or more in adults, then those adults are classified as risk group and the rest are non-risk group. Similar examples are HbA1c, LDL, Cholesterol, *etc.*, The bi-distributional ROC curves namely Bi-Normal, Bi-Exponential, Bi-Gamma *etc*., will fit to the above situation. But, these ROC forms do not fit to deal with situation where lower values of variables indicate risk and higher indicate non-risk group. For example, if the HDL is less than 40mg/dL (for men aged more than 20), then such individuals are considered as at risk and may be prone to cardiac issues and higher values relates to non-risk group. Few more variables that take similar phenomena as that of HDL are copper, iron, *etc*. Thus, it is important to address the problem of defining an ROC curve and also the corresponding intrinsic measures that can fit for the situation of lower test scores indicating risk group and higher scores indicate non-risk group.

With this background, we have considered Lindley (L) (Lindley, 1958; Ghitany *et al.*, 2008), Power Lindley (PL) (Ghitany *et al.*, 2013) and Extended Power Lindley (EPL) distributions (Said, 2015) to propose new ROC forms. The parameter combinations of these distributions have some interesting points that help us in constructing the ROC forms of the required nature. Another main reason to consider variant forms of Lindley is that it has a better fit than the exponential distribution (Ghitany *et al.*, 2008). It is also known that the power transformation and having additional parameter for basic Lindley form provides a lot of mathematical flexibility in explaining the shape and dispersion of heavy tail.

2. Family of Some Lindley Distributions

In this section, we start with the probability density functions and cumulative distribution function of the three distributions *i.e.*, Lindley, Power Lindley and Extended Power Lindley respectively.

2.1. Lindley distribution

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

$$F(x) = 1 - \left(1 + \frac{\theta x}{\theta+1}\right)e^{-\theta x} \quad ;\ \theta, x > 0 \tag{2}$$

where $\theta \in (0,1)$, is a scale parameter.

2.2. Power Lindley distribution

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

$$F(x;\theta,\alpha) = 1 - \left(1 + \frac{\theta}{\theta+1}x^\alpha\right)e^{-\theta x^\alpha} \quad ;\theta,\alpha,x > 0 \tag{4}$$

where θ is a scale parameter and α is a shape parameter. The purpose and reason to work on Power Lindley distribution is to overcome the theoretical and practical limitations of Lindley distribution. PL distribution is more flexible and this can be viewed as mixture of Weibull distribution due to the power transformation (shape α and scale θ), and a generalized gamma distribution (with shape parameters 2, α and scale θ), with mixing proportion $p = \theta/(\theta + 1)$ (Ghitany *et al*. 2013). For the values of α between 0 and 1, and with $\theta > 0$, we can have the increasing and decreasing nature of the density function.

2.3. Extended Power Lindley distribution

$$f(x;\theta,\beta,\alpha) = \frac{\alpha\theta^2}{\theta+\beta}(1+\beta x^\alpha)x^{\alpha-1}e^{-\theta x^\alpha} \quad ;\theta,\beta,\alpha,x > 0 \tag{5}$$

$$F(x;\theta,\beta,\alpha) = 1 - \left(1 + \frac{\theta\beta}{\theta+\beta}x^\alpha\right)e^{-\theta x^\alpha} \quad ;\theta,\beta,\alpha,x > 0 \tag{6}$$

EPL distribution can be shown as the mixture of Weibull distribution (with shape α and scale θ), and a generalized gamma distribution (with shape parameters 2, α and scale θ), with mixing proportion $p = \theta/(\theta + \beta)$.

3. Family of Three Lindley ROC Curves

In this section, we have developed a family of Lindley ROC Curves based on the considered Lindley distributions.

3.1. Bi-Lindley (L) ROC curve

It is assumed that the test scores (*S*, which is attributed as random variable) of normal/population I (denoted with "0") and abnormal/population II (denoted with "1") follow Lindley Distribution and the expression for the *FPR (1- specificity)* is defined as

$$FPR = x(t) = \left(1 + \frac{\theta_0}{\theta_0+1} t\right) e^{-\theta_0 t} \tag{7}$$

The threshold values can be obtained using the following formula

$$t = \left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right] \tag{8}$$

Here, an approximation of the type $\log(1+x) \approx x$ is used in driving the expression of "*t*", since our interest is only involved in the first order term in ROC form and the *TPR* (*sensitivity*) is obtained as

$$TPR = y(t) = \left(1 + \frac{\theta_1}{\theta_1+1} t\right) e^{-\theta_1 t} \tag{9}$$

on substituting the "*t*" value in above expression, the Lindley ROC Curve can be estimated as

$$y(t) = \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]} \tag{10}$$

Further, the Area under the Lindley ROC Curve can be estimated as

$$AUC = \int_0^1 \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]} dx(t) \tag{11}$$

on further simplification, the expression for AUC is

$$AUC = \frac{\theta_0^2}{\theta_0^2+\theta_1(\theta_0+1)}\left[\frac{\theta_1(\theta_0+1)(\theta_1+1)+\theta_0^2(\theta_1+1)+\theta_1(\theta_0+1)}{(\theta_1+1)\left(\theta_0^2+\theta_1(\theta_0+1)\right)}\right] \tag{12}$$

3.2. Bi-Power Lindley (PL) ROC curve

The *FPR* for the Power Lindley distribution can be derived as follows

$$FPR = x(t) = \left(1 + \frac{\theta_0}{\theta_0+1} t^{\alpha_0}\right) e^{-\theta_0 t^{\alpha_0}} \tag{13}$$

From the above expression, the threshold value can be found at each and every test score as

$$t = \left[\left(\frac{\theta_0+1}{\theta_0^2}\right) \log\left(\frac{1}{x(t)}\right)\right]^{\frac{1}{\alpha_0}} \tag{14}$$

Further, the expression for the *TPR* under Power Lindley distribution is derived as

$$TPR = y(t) = \left(1 + \frac{\theta_1}{\theta_1+1} t^{\alpha_1}\right) e^{-\theta_1 t^{\alpha_1}} \tag{15}$$

on substituting the expression for "*t*" in the above equation, the Power Lindley ROC Curve can be obtained as

$$y(t) = \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} \tag{16}$$

Further, the Area under the Power Lindley ROC Curve can be estimated as follows

$$AUC = \int_0^1 \left(1 + \frac{\theta_1}{\theta_1+1}\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+1}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} dx(t) \tag{17}$$

The above expression does not have a closed form solution and has to be evaluated by numerical integration method.

3.3. Bi- Extended Power Lindley (EPL) ROC curve

ROC Curve based on the EPL distribution is constructed as follows. The False Positive Rate is given by

$$FPR = x(t) = \left(1 + \frac{\theta_0\beta_0}{\theta_0+\beta_0}t^{\alpha_0}\right)e^{-\theta_0 t^{\alpha_0}} \tag{18}$$

on further simplification, the expression for the threshold 't' is given by

$$t = \left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{1}{\alpha_0}} \tag{19}$$

The True Positive Rate is given by

$$TPR = y(t) = \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}t^{\alpha_1}\right)e^{-\theta_1 t^{\alpha_1}} \tag{20}$$

The Extended Power Lindley ROC (EPLROC) Curve can be defined on substituting the expression for "t" in the above equation as follows.

$$y(t) = \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} \tag{21}$$

Further, the Area under the EPLROC Curve can be derived as follows,

$$AUC = \int_0^1 \left(1 + \frac{\theta_1\beta_1}{\theta_1+\beta_1}\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right) e^{-\theta_1\left[\left(\frac{\theta_0+\beta_0}{\theta_0^2}\right)\log\left(\frac{1}{x(t)}\right)\right]^{\frac{\alpha_1}{\alpha_0}}} dx(t) \tag{22}$$

The above expression does not have a closed form solution and has to be evaluated by numerical integration method.

4. Simulation Studies

The application of the proposed three new ROC forms is demonstrated using simulated data. For each ROC type, that is L, PL and EPL, the random numbers (RNs) are generated according to their distribution functionalities. With respect to Lindley distribution, the RNs are generated using quantile function. For Power Lindley and Extended Power Lindley forms, the RNs are generated using mixture of Weibull and Generalized Gamma distributions (Ghitany *et al.* (2013) and Said (2015)).

To demonstrate different forms of ROC curves (worst, moderate and best), simulations are carried out with different parameter combinations and the optimal threshold is deduced using Youden's index (J) for each of the combinations. In table 1, the results pertaining to all the three ROC curves are reported and the figures (ROC Curves) are depicted in Figure 1, which shows the comparison between the proposed ROC curves. The ROC curves are in the order of the parameter combination which is displayed as in Table 1.

Table 1: AUC and J values of L, PL and EPL ROC curves

θ_0	θ_1	AUC_L	J_L
0.6	0.5	0.4530	0.095
0.9	0.5	0.6265	0.2958
1	0.5	0.6675	0.3442
1.3	0.5	0.7567	0.4564
1.8	0.5	0.8396	0.5765

θ_0	θ_1	α_0	α_1	AUC_{PL}	J_{PL}
0.6	0.5	2	0.5	0.7486	0.6438
0.9	0.5	2	0.5	0.786	0.6784
1	0.5	2	0.5	0.795	0.6870
1.3	0.5	2	0.5	0.8148	0.7077
1.8	0.5	2	0.5	0.8357	0.7318

θ_0	θ_1	β_0	β_1	α_0	α_1	AUC_{EPL}	J_{EPL}
0.6	0.5	3	2.5	2	0.5	0.7487	0.7006
0.9	0.5	3.2	2.5	2	0.5	0.7945	0.7335
1	0.5	3.5	2.5	2	0.5	0.8016	0.7413
1.3	0.5	3.8	2.5	2	0.5	0.8243	0.7604
1.8	0.5	4	2.5	2	0.5	0.8504	0.7826

In Figure 1, the advantage of power transformation to Lindley and its extension by having additional parameter can be seen clearly. That is, the ROC curve of Lindley is very close to the chance line, which is not a preferable form for a better classification, whereas with additional shape parameter, the ROC forms of PL and EPL have shifted towards the top left corner of the unit square plot. In the context of ROC methodology, any test's or procedure's ROC curve should be far away from the chance line indicating that the test/procedure can classify the subjects with greater accuracy. Here, with these simulations, the advantage of having additional shape parameter has boosted the performance of a classifier witnessing a better ROC curve. However, with increase in the scale parameter values of population I, the ROC curve of Lindley distribution gradually shifted towards the top left corner of unit square plot. The gradual improvement in ROC curve of each L, PL and EPL ROC curves can be seen in Figure 2.

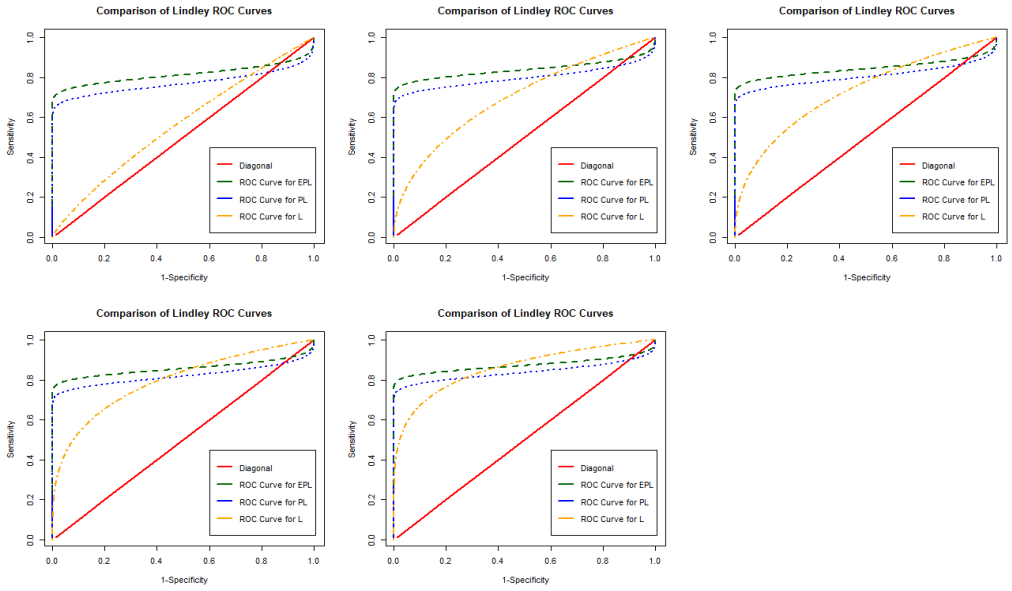


Figure 1: Graphical Comparison between L, PL and EPL ROC curves

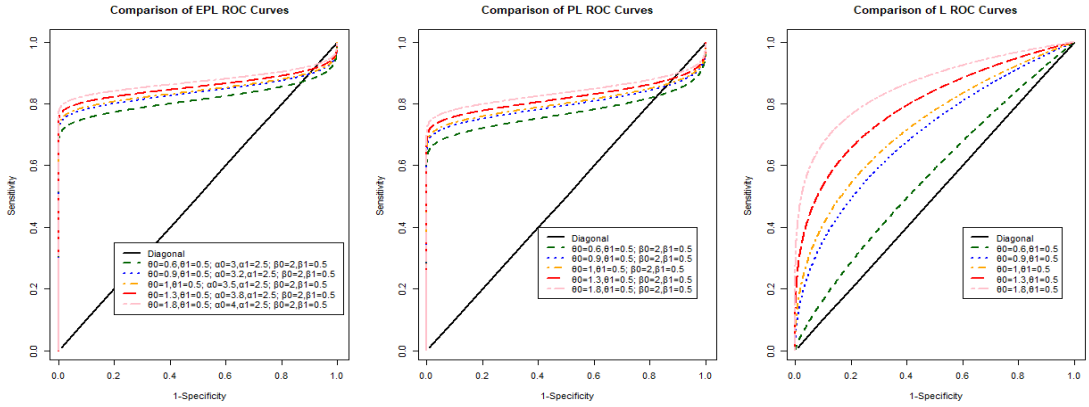


Figure 2: L, PL and EPL ROC curves with different parameter combinations

One more illustration is also carried out to address the question; "what happens to the nature of the ROC curve if the parameter combinations of population I are kept constant and varying parameter values in population II?". In the previous illustration, the simulations and ROC curves are exhibited for the case where the parameter values of population II are fixed. With the second illustration, it is observed that it affects the performance of the classifier and will not have impact in having a better accuracy (Table 2). With a large difference of scale and shape values between the populations I and II, some sort of improvement in ROC curves can be witnessed (Figures 3 and 4). This is due to the basic nature of the distribution forms that the values of population I should be at the higher side than that of population II, which is a very rare phenomenon in the general context of ROC methodology. Hence, the distributional forms of L, PL and EPL distributions has a very rare functionality of having higher values on population I (Normal or Healthy) than that of population II (Abnormal or Diseased), and these proposed ROC forms can be applied to such situations to explain the accuracy and other measures.

Table 2: AUC and J values of L, PL and EPL ROC curves

θ_0	θ_1	AUC_L	J_L
1.8	1.8	0.4766	0.0000
1.8	1.5	0.5342	0.0871
1.8	1.2	0.6052	0.1942
1.8	0.9	0.6929	0.3295
1.8	0.6	0.8001	0.5057

θ_0	θ_1	α_0	α_1	AUC_{PL}	J_{PL}
1.8	1.8	2	2	0.4766	0.0000
1.8	1.5	2	1.7	0.5201	0.0916
1.8	1.2	2	1.4	0.5858	0.2183
1.8	0.9	2	1.1	0.6788	0.3927
1.8	0.6	2	0.8	0.7974	0.6201

θ_0	θ_1	β_0	β_1	α_0	α_1	AUC_{EPL}	J_{EPL}
1.8	1.8	4	4	2	2	0.3615	0.0000
1.8	1.5	4	3.7	2	1.7	0.4167	0.1094
1.8	1.2	4	3.4	2	1.4	0.5052	0.2558
1.8	0.9	4	3.1	2	1.1	0.6350	0.4470
1.8	0.6	4	2.8	2	0.8	0.7924	0.6769

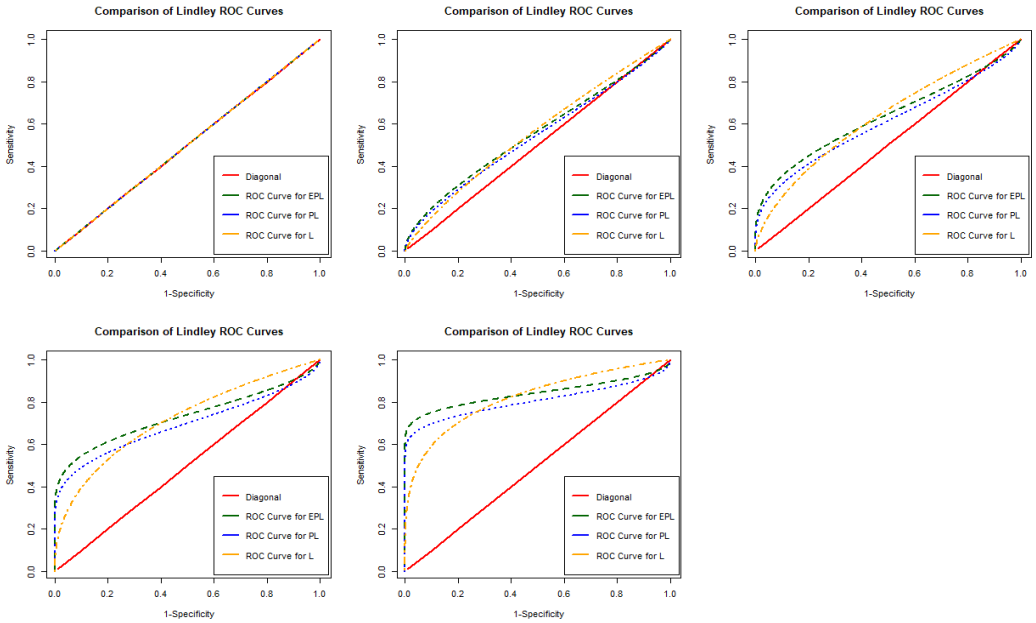


Figure 3: Graphical Comparison between L, PL and EPL ROC curves

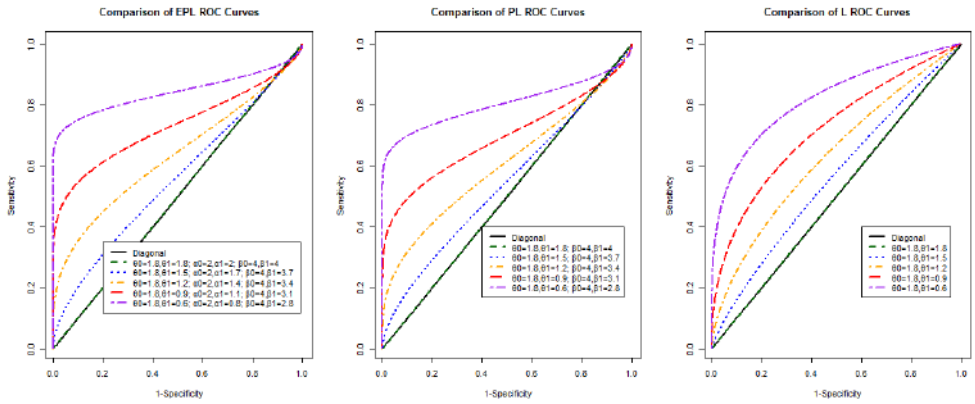


Figure 4: L, PL and EPL ROC curves with different parameter combinations

5. Conclusion

Different ROC curves have been studied in this paper by taking into consideration the three variant forms of Lindley distribution. The reason being certain mathematical and functional advantages such as superiority of Lindley over Exponential, the ease of Power transformation to the basic Lindley form and having additional shape parameter to the Power Lindley distribution. These considerations have been the support and motivation to propose three ROC curves namely, L, PL and EPL ROC forms. The advantages and flexibility of having power transformation and additional parameter is well demonstrated through simulation studies and also using graphical comparisons. Further, from the simulations and parameter combinations, an interesting fact that the ROC pattern and assumption of scores in population I and population II are in reverse pattern than that of the usual assumption made in several bi-distributional ROC forms such as Bi-Normal, Bi-Gamma and Bi-Exponential ROC curves *etc.*, was revealed. At most attention to the type of data is needed before fitting the proposed ROC Curves. The three L, PL and EPL ROC curves are quite applicable and apt to the practical contexts where the above said situation is witnessed.

References

Balaswamy. S., and Vishnu Vardhan. R. (2016). An anthology of parametric ROC models. *Research and Reviews: Journal of Statistics*, **5(2)**, 32-46.

Ghitany, M., Atieh, B., and Nadadrajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**, 493-506.

Ghitany, M., Al-Mutairi, D., Balakrishnan, N., and Al-Enezi, I. (2013). Power Lindley distribution and associated inference. *Computational Statistics and Data Analysis*, **64**, 20-33.

Krzanowski, W. J., and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Monographs on Statistics and Applied Probability. New York, NY: CRC Press (ISBN: 978-1-4398-0021-8).

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society*, **B20**,102-107.

Lusted, L. B. (1971). Signal detectability and medical decision making, *Science*, **171**(**3977**), 1217-1219.

Said Hofan Alkarni (2015). Extended power Lindley distribution: A new statistical model for non-monotone survival data. *European Journal of Statistics and Probability*, **3**(**3**), 19-34.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Lawrence Erlbaum Associates, New Jersey (ISBN 978-1-1389-8191-1)

Swets, J. A., and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press (ISBN 978-0-12-679080-1).

Tanner Jr, W. P., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**(**6**), 401-409.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 241 - 253

Measuring interchangeability in school lunch intervention data using concordance correlation coefficient

Gokul, T.[1], Srinivasan, M. R.[1] and Swaminathan, V. D.[2]
[1]*Department of Statistics, University of Madras, Chennai*
[2]*Department of Psychology, University of Madras, Chennai*

Received: 02 March 2020; Revised: 26 October 2020; Accepted: 30 December 2020

Abstract

In this paper, the measure of agreement is evaluated in terms of interchangeability among the set of treatment groups through the concordance correlation coefficient (CCC) for the longitudinal data using generalized linear mixed model (GLMM) and zero inflated model to account for the presence of zero observations in the data. CCC allow us to retain/interchange the treatment groups and provide scope for many experimental researches. Apart from CCC, this study also considers intra class correlation coefficient (CC), precision and accuracy for the evaluation. A simulation study is carried out to evaluate the performance of CCC followed by an application to a psychometric data that reveals there is no interchangeability in the nutritional supplements based on a school lunch intervention study.

Key words: Concordance correlation coefficient; Generalized linear mixed model; Poisson model; Zero inflated model.

1. Introduction

In many psychological studies, the measurement of variables could be ordinal, count or continuous using different methods. Further, in the case of longitudinal data, each subject is measured at different time points to obtain repeated measurements. Generally, measurements of agreement such as kappa measure in statistics are commonly used to identify the degree of concordance between the two or more observers. However, for the longitudinal data, there exists a necessity to assess the agreement between repeated measurements produced by a single observer or among multiple measurement methods.

The concordance correlation coefficient (CCC) proposed by Lin (1989, 1992) is one of the most widely applied procedures to assess agreement between observers on a quantitative scale by measuring the variation of linear relationship between each pair of data from a 45-degree line through the origin and it degenerates into kappa and weighted kappa for binary and ordinal data. The CCC was formerly defined as the Euclidean distance between paired data from two observers and the concordance line, and it was conveniently scaled to [-1, 1] interval, where -1 indicates perfect inverse agreement, 1 indicates perfect agreement, and value 0 is interpreted as complete disagreement. Further, each pair of measurements should fall on the 45-degree line, otherwise some disagreement is present in the data and hence that particular method becomes interchangeable. Traditionally, CCC is said to be a measure of total agreement and it can also be expressed as intra class correlation coefficient (CC) discussed in Carrasco and Jover (2003) and demonstrated its equivalence using mixed effect model by two methods namely variance components and moment method. Carrasco and

Corresponding Author: Srinivasan M.R.
E-mail: mrsrin8@gmail.com

Jover (2005) extended the CCC for measuring agreement with count data by means of intra class CC derived from a GLMM.

Literature is abundant in studying the measure of agreement in psychological studies (Barchard, 2012; Ma *et al.,* 2010) and for more than two observers in discrete data (King and Chinchilli, 2001; Carrasco and Jover, 2005; Carrasco *et al.,* 2009). Lin *et al.,* (2007) studied CCC based on the variance components under linear mixed model for quantitative/ qualitative data. Here, we have extended CCC to generalized version so that this can be applied to various kinds of data based on the generalized linear mixed model (GLMM) for the longitudinal and repeated measure data (Ge *et al.,* 2016). Carrasco (2010) proposed an index to measure the degree of agreement as the extended version of concordance correlation coefficient (CCC) through variance component (VC) approach for count data using GLMM framework. Moreover, CCC has been estimated with weights of a diagonal matrix between various repeated measurements over time for the longitudinal data. In the case of count data, Carrasco (2010) adopted a two-way GLMM with subject and observer specific random effects and random subject observer interaction effects. We extend this to a three-way GLMM and estimate the CCC for longitudinal count data. Thus, the aim of this work is to evaluate the impact of the CCC for count data with the presence of zeroes by means of a generalized expression in the CCC (GCCC) based on the intra class CC through GLMM approach count data. The idea behind the GCCC is first to fit the data using the most appropriate GLMM, and subsequently to develop the expression of the CCC based on the model parameters. Usually, the correlations have been studied for the identification of the relationship as strong or weak, but the methods/treatments are not interchangeable. In the context of CCC, it is expressed to improve the measure of agreement by interchanging the variables. CCC and intra class CC are the statistics that quantify the proportion of variance explained by a random factor in multilevel/hierarchical data.

The focus of this article is to explore the applicability of the measure of agreement CCC and intra class CC for each observer through GLMM for longitudinal data. Further, an extended three-way GLMM for longitudinal count data in the presence of zeros is considered to measure the agreement between the variables together with inter class CC, intra class CC, and total agreement. The paper is organized as follows. Section 2 introduces the dataset considered to examine the application of CCC. The existing methodology of generalized linear mixed model is reviewed together with the measurement of agreements CCC in Section 3. The results based on application of CCC for longitudinal data is discussed in Section 4. Section 5 provides the conclusion.

2. School Lunch Intervention Data

The cognitive data is a secondary data consisting of school lunch intervention given to children in rural Kenya (Neumann *et al.,* 2003). The intervention study is designed with three feeding groups of school children and also a control group who received no nutritional supplements. Each treatment group is comprised of 12 centres with children aged 6–14 years. The school lunch intervention was carried out in 9 out of 12 schools and students at the other three schools formed a control group. Data collected in Round 1 served as baseline before the intervention and called as pre-intervention scores. Round 2 was taken during the term after the intervention started and data in rounds 3, 4, and 5 were recorded during the second, fourth, and sixth terms after intervention started as post intervention scores. A total of 554 participants have been recorded including missing entries in the data. Data associated with 374 participants, excluding the missing observations is considered for this study: out of which

188 were boys and 186 were girls. Among the 374, 97 children were given calorie supplement, 127 children were given meat supplement, 78 were given milk and 72 were considered as control group in this study.

For the intervention study, recorded data was on general intelligence factor called G-factors (Raven's coloured progressive matrix) and other three S-factors (Verbal meaning, Arithmetic score and Digit span total), with nutritional supplements in order to study the measurement of agreement of interchangeability. Five repeated rounds of lunch intervention have been recorded from the schools and a summary of the data is presented in Table 1. From Table 1, it is evident that the there is an increasing trend in the overall mean for all the response variable considered. In addition, verbal meaning shows higher mean value since higher order cognitive function involving reasoning abilities on linguistic domain. The recorded data has been subjected to GLMM for count data and inflated GLMM for data with zeroes.

Table 1: Human intelligence - Overall mean for all response group

Time	RCPM	AS	VM	DS
Round 1	17.11	7.02	26.68	4.90
Round 2	17.47	7.16	27.27	5.37
Round 3	18.02	7.52	29.03	6.20
Round 4	18.41	8.04	32.03	6.97
Round 5	19.40	8.75	33.96	7.80

RCPM: Raven's Coloured Progressive Matrices; AS: Arithmetic Score; VM: Verbal Meaning; DS: Digit Span

3. Models and methods

Consider a study where n subjects are measured m times by J observers. Let Y_{ijkl} be the l^{th} reading $(l=1,...,L)$ made by the j^{th} observer $(j=1,...,J)$ at the k^{th} time $(k=1,...,K)$ on the i^{th} subject $(i=1,...,n)$. Since a sample of subjects is included in the data, the subject and the subject-by-observer interaction effects are assumed to be random effects. If the interest lies only in measuring the agreement among those observers included in the dataset as in many situations, then the same could be considered as observer-specific and are assumed as fixed effect. However, as mentioned by Carrasco and Jover (2003), when defining the agreement index, it is convenient to consider the observer effect as random to account for the systematic differences between observers as a source of disagreement. The index would otherwise, measure consistency rather than agreement. Thus, the present study, has considered variance components model as appropriate to fit the data with, subject-specific, subject-observers, subject–time interaction effects, $u_{ijk}=(\alpha_i,\alpha\beta_{ij},\alpha\gamma_{ik})$ as sources of variability.

The random-effects vector u_{ijk} are independently distributed from an exponential family with mean $E(Y_{ijkl}\,|\,u_{ijk})=\mu_{ijk}$ and variance $\text{var}(Y_{ijkl}\,|\,u_{ijk})=v_{ijk}=\phi v(\mu_{ijk})$ where $v(\cdot)$ is a user specified variance function and ϕ is a unknown dispersion parameter. Through the link function $g(\mu_{ijk})=\eta_{ijk}$, the conditional mean associated with a linear predictor is given by

$$\eta_{ijk} = x_{ijk}\beta^* + z_{ijk}u_{ijk}, \tag{1}$$

where x_{ijk} $(p\times 1)$ and z_{ijk} $(q\times 1)$ are independent variables with the fixed effects β^* and the random effects u_{ijk}.

We extend the three-way LMM proposed by Carrasco *et al.*, (2009) and Tsai (2017) to the GLMM incorporating repeated measurements rated by an observer at a certain time for each subject. The extended three-way GLMM can be written as

$$\eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}, \tag{2}$$

where μ is the overall mean, α_i is the subject-specific random effect assumed to be distributed as $\alpha_i \sim N(0,\sigma_\alpha^2)$, β_j is the observer-specific fixed effect, γ_k is the time-specific fixed effect, $\alpha\beta_{ij}$ is the random subject–observer interaction effect assumed to be distributed as $\alpha\beta_{ij} \sim N(0,\sigma_{\alpha\beta}^2)$, $\alpha\gamma_{ik}$ is the random subject–time interaction effect assumed to be distributed as $\alpha\gamma_{ik} \sim N(0,\sigma_{\alpha\gamma}^2)$, and $\beta\gamma_{jk}$ is the fixed observer–time interaction effect and all the three parameters are mutually independent. From the equation (1) and (2), fixed effect is expressed as $\beta^* = (\mu,\beta_1,...,\beta_J,\gamma_1,...,\gamma_K,\beta\gamma_{11},.....,\beta\gamma_{JK})$ and random effect as $u_{ijk} = (\alpha_i,\alpha\beta_{ij},\alpha\gamma_{ik})$, $u_{ijk} \sim MVN(0,G)$ where G is a diagonal matrix with elements $\sigma_\alpha^2, \sigma_{\alpha\beta}^2$ and $\sigma_{\alpha\gamma}^2$ on the diagonal and zero otherwise.

Following, Barnhart *et al.*, (2005) and Lin *et al.*, (2007), the total, intra class CC and inter class CC based on the GLMM approach can be expressed as follows

$$\rho_{CCC} = \frac{\text{cov}(Y_{ijkl}, Y_{ijk'l'})}{\text{var}(Y_{ijkl})} \tag{3}$$

where $\text{cov}(Y_{ijkl}, Y_{ijk'l'})$ and $\text{var}(Y_{ijkl})$ stand for the marginal covariance of the l^{th} reading $(l = 1,...,L)$ made by the j^{th} observer $(j = 1,...,J)$ at the k^{th} time $(k = 1,...,K)$ on the i^{th} subject $(i = 1,...,n)$. The marginal variance and covariance are developed as

$$\begin{aligned}
\text{var}(Y_{ijkl}) &= \text{var}_u\{E(Y_{ijkl} \mid u_{ijk})\} + E_u\{\text{var}(Y_{ijkl} \mid u_{ijk})\} \\
&= \text{var}_u(\mu_{ijk}) + E_u\{\phi h(\mu_{ijk})\} \\
\text{cov}(Y_{ijkl}, Y_{ijk'l'}) &= \text{cov}_u\{E(Y_{ijkl} \mid u_{ijk}), E(Y_{ijk'l'} \mid u_{ijk'})\} + E_u\{\text{cov}(Y_{ijkl}, Y_{ijk'l'} \mid u_{ijk}, u_{ijk'})\}.
\end{aligned}$$

Since Y_{ijkl} and $Y_{ijk'l'}$ are conditioned effects considered independent, the marginal covariance reduces to (McCulloch and Searle, 2001)

$$\begin{aligned}
\text{cov}(Y_{ijkl}, Y_{ijk'l'}) &= \text{cov}_u\{E(Y_{ijkl} \mid u_{ijk}), E(Y_{ijk'l'} \mid u_{ijk'})\} \\
&= \text{cov}_u(\mu_{ijk}, \mu_{ijk'}).
\end{aligned}$$

In addition, the CCC may also be divided into two components namely the precision and accuracy (Lin, 1989; Lin *et al.*, 2007) so that

$$\rho_{CCC} = \rho_p \cdot \chi_a \tag{4}$$

The precision component ρ_p, is the intra class CC considering the observers as fixed in equation (6), *i.e.*, not considering the between-observers variability as a source of disagreement. Additionally, the accuracy index χ_a, measures the distance between the observers' means, *i.e.*, the systematic differences among the observers. Thus, using the expressions of the marginal variance and covariance in the CCC equation (3), the following is obtained:

$$\rho_{GCCC} = \frac{\text{cov}_u(\mu_{ijk}, \mu_{ijk'})}{\text{var}_u(\mu_{ijk}) + E_u\{\phi h(\mu_{ijk})\}} \tag{5}$$

where ϕ is the dispersion parameter and $h(\cdot)$ the corresponding variance function associated random effects, which can be considered as a generalization of CCC (GCCC) to fit GLMM approach. Further, Lin *et al.*, (2007) defined the intra class CC as a measure of proportion of total variance attributable to the subjects and can be expressed as

$$\begin{aligned}\rho_{GCCC}^{\text{intra}} &= \frac{\text{cov}_u(Y_{ijkl}, Y_{ijk'l'})}{\text{var}_u(Y_{ijkl})}\Bigg|_{j(l,l')} \\ &= \frac{\text{var}_u(\mu_{ijk})|_{j(l,l')}}{\text{var}_u(\mu_{ijk})|_{j(l,l')} + E_u\{\phi h(\mu_{ijk})\}}\end{aligned} \tag{6}$$

where $j(l,l')$ be the reading measured m times based on j^{th} observer. Additionally, the conditional variance and covariance of $Y_{ijkl}, Y_{ijk'l'}$ given $j(l,l')$ is defined in equation (6). Furthermore, an intra class CC can also be defined as a measure of intra-observer agreement (Barnhart *et al.*, 2005; Lin *et al.*, 2007) where the observer effect is considered as fixed. This index should be interpreted, for each rater, as a measure of the proportion of the total variance (subjects plus error) attributable to subjects. It is also possible to define an inter class CC if the data have replicated readings $(m > 1)$ by considering the data as the average of those m readings

$$\begin{aligned}\rho_{GCCC}^{\text{inter}} &= \frac{\text{cov}(\bar{Y}_{ijk.}, \bar{Y}_{ijk'.})}{\text{var}\ (\bar{Y}_{ijk.})} \\ &= \frac{\text{cov}_u(\mu_{ijk}, \mu_{ijk'})|_{j(l,l')}}{\text{var}_u(\mu_{ijk}) + \dfrac{E_u\{\phi h(\mu_{ijk})\}}{m}}\end{aligned} \tag{7}$$

where $\bar{Y}_{ijk.}$ stands for the average of m readings of the j^{th} observer on the i^{th} subject and along with k^{th} time point. Specifically, two models are considered for the count data based on the presence or absence of zero observations.

3.1. Poisson model

Let $Y \mid X, u$ follows a Poisson distribution and the conditional mean of Y_{ijkl} given u_{ijk} is $\mu_{ijk} = e^{\lambda_i + \alpha_i + \beta_j + \gamma_k}$ and the conditional variance is given by $\text{var}(Y_{ijkl} \mid u_{ijk}) = \mu_{ijk}$. The marginal expectation over the random effects is expressed in terms of its generating function and is given by

$$\begin{aligned} E_u(\mu_{ijk}) &= E_u(e^{\lambda_i + \alpha_i + \beta_j + \gamma_k}) = e^{\lambda} E_u(e^{\lambda_i + \alpha_i + \beta_j + \gamma_k}) \\ &= e^{\lambda} M_u(\alpha_i + \beta_j + \gamma_k), \end{aligned}$$

where $\lambda = \lambda_0 + \lambda_1 X_1 + ... + \lambda_p X_p$. Since, $u \sim MVN(0, G)$ under the assumption of random effects, the expected value can be obtained from $E_u(\mu_{ijk}) = e^{\lambda + \frac{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}{2}}$. Thus, the specific CCC for the Poisson GLMM as explained by Carrasco (2010) becomes

$$\begin{aligned} \rho_{CCC} &= \frac{E_u(\mu_{ijk})^2 (e^{\sigma_\alpha^2} - 1)}{E_u(\mu_{ijk})^2 (e^{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} - 1) + E_u(\mu_{ijk})} \\ &= \frac{E_u(\mu_{ijk})(e^{\sigma_\alpha^2} - 1)}{E_u(\mu_{ijk})(e^{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} - 1) + 1} \end{aligned}$$

Further, CCC would reduce to

$$\rho_{CCC} = \frac{\mu(e^{\sigma_\alpha^2} - 1)}{\mu(e^{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2} - 1) + 1} \tag{8}$$

in the case of no subject-observer interaction. When $\sigma_\alpha^2 \to 0$ or $\sigma_\beta^2, \sigma_\gamma^2 \to \infty$, the CCC will tend to 0 (independence) and conversely, when $\sigma_\alpha^2 \to \infty$ or $\sigma_\beta^2, \sigma_\gamma^2 \to 0$, it will reach a value of 1 (perfect agreement) if $\sigma_\alpha^2 \rangle\rangle \sigma_\beta^2, \sigma_\gamma^2$. It is to be noted that the CCC is defined using the variance components, thus it cannot result in negative values.

3.2. Zero Inflated Poisson (ZIP) model

Let Y_{ijk} denote the longitudinal response for j^{th} observer $(j = 1,...,J)$ at the k^{th} time $(k = 1,...,K)$ on the i^{th} subject $(i = 1,...,n)$. Then, the distribution of Y_{ijk} is expressed as

$$Y_{ijk} = \begin{cases} 0 & \text{with prob. } \phi_{ijk} \\ Poisson\ (\lambda_{ijk}) & \text{with prob. } 1 - \phi_{ijk} \end{cases} \tag{9}$$

where ϕ_{ijk} denotes the probability of the observation arising from the degenerated distribution at zero and λ_{ijk} represents the mean of the Poisson distribution. This formulation incorporates more zeros than permitted under the Poisson assumption (*i.e.*, where $\phi_{ijk}=0$). The probability distribution function of the longitudinal ZIP model can be written as

$$\begin{aligned} p(Y_{ijk}=0\,|\,Z_{ijk}) &= \phi_{ijk}+(1-\phi_{ijk})e^{-\lambda_{ijk}} \\ p(Y_{ijk}=0\,|\,X_{ijk}) &= (1-\phi_{ijk})\frac{\lambda_{ijk}^{y_{ijk}}e^{-\lambda_{ijk}}}{y_{ijk}!}, \qquad y_{ijk}=1,2,... \end{aligned} \tag{10}$$

where $0\le\phi_{ijk}\le 1$ and $0<\lambda_{ijk}<\infty$. Here X_{ijk} and Z_{ijk} can be mutually exclusive, partially, or completely overlapping to achieve modeling flexibility. This model has the same specification as that of the Poisson model but the variance function is expressed by $E_u\{\phi_{ijk}E_u(\mu_{ijk})\}=\phi_{ijk}E_u(\mu_{ijk})$ and the total CCC for ZIP model becomes

$$\rho_{CCC}=\begin{cases} \dfrac{E_u(\mu_{ijk})(e^{\sigma_\alpha^2}-1)}{E_u(\mu_{ijk})(e^{\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2}-1)+\phi_{ijk}} & \text{with prob. } \phi_{ijk} \\ \dfrac{E_u(\mu_{ijk})(e^{\sigma_\alpha^2}-1)}{E_u(\mu_{ijk})(e^{\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2}-1)+(1-\phi_{ijk})} & \text{with prob. } 1-\phi_{ijk} \end{cases} \tag{11}$$

To estimate the CCC and related quantities for the ZIP case, the data are fitted by penalized quasi-likelihood (PQL) through the SAS GLIMMIX and NLMIXED (Zhu *et al.*, 2015). The goal of the analysis is to determine the degree of agreement between each treatment group in order to decide if a new type of treatment could replace any other treatment. The interchangeability of treatment in this study will be considered only if the CCC value is at least 0.9 (Carrasco 2010). The following section describes the application of CCC for school lunch intervention study.

4. Simulation Study

The aim of the simulation study is to evaluate the impact of zero counts present in the dataset and assess the performance of CCC obtained from ZIP model. The parameter values of each combination were used as initial values to start the estimation process, with regard to the ML procedure. Carrasco (2010) studied to evaluate the impact of the over dispersion when estimating the CCC assuming a Poisson distribution. Similar, CCC estimate from a normal linear mixed model (Normal) whose behavior is related to the classical Lin's sample moment approach (Lin, 1989; Carrasco and Jover, 2003).

Following Carrasco (2010) simulation has been further extended to ZIP model and built three levels of agreement (as discussed in Section 1) through CCC values namely, below 20 as low, between 20 and 200 as medium and above 200 as high. Similarly, simulating the data into multiple zero values and classified as low (below 30), moderate (between 30 and 120), high (between 120 and 300) and extremely high (above 300) and repeated 1000 times. For each case, CCC was estimated using ZIP model using ML method of estimation and the

simulation is performed in PROC GLIMMIX and PROC NLMIXED using SAS software. Table 2 explains the CCC values for the above mentioned scenarios.

Table 2: Results of the simulation scenarios

Multiple zeros	Agreement level	Comb.	Mean	CCC
Low	Low	1	11.65	0.2922
	Medium	2	21.06	0.3397
	High	3	29.78	0.4812
Moderate	Low	4	65.18	0.5461
	Medium	5	97.87	0.6424
	High	6	115.75	0.6997
High	Low	7	150.02	0.7495
	Medium	8	208.12	0.7948
	High	9	282.52	0.8542
Extremely	Low	10	565.14	0.8999
	Medium	11	729.45	0.9206
	High	12	1314.19	0.9958

The following are the findings from Table 2:

(i) The CCC values show increasing trend irrespective of the zero counts with the increasing levels.

(ii) In the case of extremely high zero counts, there is an improvement in the measure of agreement with an increasing rate of mean. Similarly, CCC increases as the mean increases for the case of medium and high zero counts in the data.

(iii) It is clear that the CCC value reaches 0.9 when the mean value of the data is extremely high irrespective of the zero counts and this provides an insight in interchangeability of the methods used for the study.

The simulation results paved the way for a better understanding of CCC with different categories of mean and zero counts in the data and motivates us to incorporate the same in the real time data set as explained in the following section.

5. Data Analysis

The school lunch intervention data introduced in Section 2 were considered for data analysis. To capture the measurement of agreement, various assumptions about the distribution of the random effects has been made. There is also baseline covariate information on each subject including age, gender, socio economic status, intake of nutritional supplements such as milk and meat, duration of the follow-up study. Furtther, measurement of G-factor in the analytical ability as assessed by Raven's coloured progressive matrices and S-factor involving reasoning ability, linguistic ability and immediate memory and are assessed by Verbal meaning, Arithmetic scores and Digit span total respectively. Obviously, it is expected that improvement of children cognitive skills is correlated with the nutrition supplements and this association is studied using GLMM. We considered Poisson regression model and zero inflated Poisson model for analyzing the concept of interchangeability of nutritional supplements such as milk, meat, calories and control groups. From $u_{ijk} = (\alpha_i, \alpha\beta_{ij}, \alpha\gamma_{ik})$, a variance component model is used to fit the data including subjects-specific, subject-observers, subject–time interaction effects as sources of variability. Following Carrasco and Jover (2003), the observer effect is treated to be random since it accounts for the systematic difference between the observers by means of disagreement.

However, if the measures of agreement in CCC value is at least 0.9, then the particular treatment can be interchanged (Carrasco 2010).

The CCC is estimated using the variance components of a Poisson GLMM and the model is fitted by maximum likelihood (ML) using Gauss-Hermite quadrature through the NLMIXED SAS procedure. For each case, the CCC and its standard error are estimated. Table 3 gives the results of CCC on (8), precision, accuracy, intra class CC based on (3) to (7) for the Poisson model.

Table 3: Human intelligence - Results for Poisson model

	CCC	Precision	Accuracy	Intra class CC Observer 1	Intra class CC Observer 2	Intra class CC Observer 3
Analytical ability by Raven's coloured progressive matrices test						
Calorie	0.3926 (0.0298)	0.897 (0.0798)	0.4329 (0.0333)	0.897 (0.0798)	0.9029 (0.0805)	0.9085 (0.0810)
Meat	0.4089 (0.0299)	0.8951 (0.0786)	0.4587 (0.0352)	0.8951 (0.0786)	0.9011 (0.0802)	0.9068 (0.0808)
Milk	0.4356 (0.0327)	0.8979 (0.0798)	0.4931 (0.0398)	0.8979 (0.0798)	0.9037 (0.0806)	0.9093 (0.0812)
Control	0.3797 (0.0279)	0.8964 (0.0797)	0.4063 (0.0316)	0.8964 (0.0798)	0.9024 (0.0803)	0.908 (0.0809)
Numerical ability by Arithmetic score						
Calorie	0.3127 (0.0239)	0.6513 (0.0552)	0.3609 (0.0273)	0.6513 (0.0551)	0.6634 (0.0565)	0.6755 (0.0577)
Meat	0.3263 (0.0242)	0.6465 (0.0545)	0.3934 (0.0298)	0.6465 (0.0545)	0.6589 (0.0559)	0.6712 (0.0570)
Milk	0.3589 (0.0256)	0.6526 (0.0553)	0.4082 (0.0302)	0.6526 (0.0554)	0.6647 (0.0566)	0.6768 (0.0581)
Control	0.3004 (0.0213)	0.6477 (0.0546)	0.3589 (0.0255)	0.6477 (0.0546)	0.6602 (0.0561)	0.6725 (0.0572)
Linguistic ability by Verbal meaning						
Calorie	0.4871 (0.0383)	0.8298 (0.0729)	0.5329 (0.0432)	0.8298 (0.07287)	0.8484 (0.0746)	0.8653 (0.0768)
Meat	0.4936 (0.0399)	0.8234 (0.0723)	0.5412 (0.0447)	0.8234 (0.0723)	0.8426 (0.0736)	0.8601 (0.0764)
Milk	0.5031 (0.0403)	0.8329 (0.0731)	0.5532 (0.0452)	0.8329 (0.0731)	0.8511 (0.0752)	0.8678 (0.0775)
Control	0.4724 (0.0372)	0.829 (0.0728)	0.5216 (0.043)	0.8290 (0.0728)	0.8477 (0.0743)	0.8647 (0.0767)
Immediate memory by Digit span total						
Calorie	0.3264 (0.0226)	0.8422 (0.0743)	0.3721 (0.0277)	0.8422 (0.0743)	0.8567 (0.076)	0.8702 (0.0778)
Meat	0.3315 (0.0233)	0.8383 (0.0730)	0.3824 (0.0288)	0.8383 (0.073)	0.8532 (0.0753)	0.867 (0.077)
Milk	0.3561 (0.0256)	0.8452 (0.0744)	0.3987 (0.0293)	0.8452 (0.0242)	0.8594 (0.0762)	0.8727 (0.078)
Control	0.3129 (0.0213)	0.8387 (0.07308)	0.3621 (0.0262)	0.8387 (0.0248)	0.8537 (0.0757)	0.8674 (0.0773)

The following are the observations from Table 3:

(i) Concerning Raven's coloured progressive matrices with milk supplement produces a high degree of agreement (with CCC 0.4356) than other treatments but insufficient to declare that the treatment is interchangeable.

(ii) Based on the Koo and Li (2016) guidelines for interpreting the intra class CC be classified as poor (below 0.5), moderate (0.5 to 0.75) and excellent (above 0.9). In this study, the intra class CC was closer to one in all cases, so one could conclude that the methods adopted are reliable. However, we can see that there are few cases where the intra class CC falls between 0.6 and 0.8, but still are not closer to zero. Thus, we can say that these methods are reliable to fit under GLMM.

(iii) For the S factor through responses namely Arithmetic, Verbal meaning and Digit span, the treatment group milk produces a higher degree of agreement than other treatments.

Further, to handle the zero counts in the data, we fit the Zero inflated Poisson model on (9), the CCC behaves the same as in Poisson model yielding that there can be no interchangeability in treatments.

Table 4: Human intelligence - Results for Zero Inflated Poisson model

	CCC	Precision	Accuracy	Intra class CC Observer 1	Intra class CC Observer 2	Intra class CC Observer 3
			Analytical ability by Raven's coloured progressive matrices test			
Calorie	0.3327 (0.0237)	0.9008 (0.0801)	0.3818 (0.0279)	0.9008 (0.0801)	0.8777 (0.0783)	0.8846 (0.0786)
Meat	0.3428 (0.0241)	0.8953 (0.0789)	0.3836 (0.028)	0.8953 (0.0791)	0.9062 (0.0807)	0.9117 (0.0812)
Milk	0.3538 (0.0256)	0.8914 (0.079)	0.3863 (0.0281)	0.8914 (0.0789)	0.9068 (0.0808)	0.9123 (0.0814)
Control	0.3288 (0.0232)	0.9 (0.08)	0.3806 (0.0278)	0.9 (0.0800)	0.8907 (0.07987)	0.8969 (0.0792)
			Numerical ability by Arithmetic score			
Calorie	0.2684 (0.0152)	0.6554 (0.0556)	0.3017 (0.0205)	0.6554 (0.055)	0.6381 (0.0531)	0.6507 (0.0552)
Meat	0.2736 (0.0171)	0.6448 (0.0541)	0.3069 (0.021)	0.6448 (0.0541)	0.6711 (0.0568)	0.6831 (0.0582)
Milk	0.2883 (0.0189)	0.6452 (0.0546)	0.3082 (0.0214)	0.6452 (0.0545)	0.6615 (0.0562)	0.6737 (0.0573)
Control	0.2491 (0.0144)	0.6552 (0.0554)	0.3002 (0.0201)	0.6552 (0.0554)	0.6273 (0.0527)	0.64 (0.0536)
			Linguistic ability by Verbal meaning			
Calorie	0.4186 (0.0313)	0.8346 (0.0726)	0.4949 (0.0397)	0.8346 (0.0725)	0.8091 (0.0705)	0.8294 (0.0715)
Meat	0.4318 (0.0325)	0.831 (0.0720)	0.4919 (0.0395)	0.831 (0.0719)	0.8368 (0.0729)	0.8548 (0.0758)
Milk	0.4536 (0.0352)	0.8218 (0.0712)	0.4962 (0.0399)	0.8218 (0.0712)	0.8441 (0.0739)	0.8613 (0.0765)
Control	0.4003 (0.0301)	0.876 (0.0781)	0.4938 (0.0396)	0.676 (0.0579)	0.8532 (0.07539)	0.8678 (0.0774)

	Immediate memory by Digit span total					
Calorie	0.2201 (0.0118)	0.8484 (0.0746)	0.3092 (0.0218)	0.8484 (0.0746)	0.804 (0.0702)	0.8214 (0.0709)
Meat	0.2239 (0.0126)	0.8382 (0.0735)	0.3131 (0.0223)	0.8382 (0.0732)	0.8534 (0.0755)	0.8671 (0.0771)
Milk	0.2282 (0.014)	0.8335 (0.0726)	0.3134 (0.0227)	0.8335 (0.0722)	0.8687 (0.0776)	0.8812 (0.0784)
Control	0.2164 (0.011)	0.7287 (0.0631)	0.2014 (0.0101)	0.7853 (0.0753)	0.8760 (0.0781)	0.8914 (0.0789)

The result in Table 4 based on (3) to (7) revealed that CCC on (11) is higher for the treatment milk in the entire response group namely Raven's coloured progressive matrices, Arithmetic score, Verbal meaning and Digit span total. This is significant because identifying the correct nutritional supplements in the development of cognitive function improves the intelligence of school children.

6. Concluding Remarks

The field of psychometric studies focus on developing a proper measure that can accurately summarize or give an idea of an individual's intellectual abilities and mental state. Longitudinal studies are common in many psychometric studies particularly on cognitive ability of school children and psychometric factors involving mental illness. Further, the policy makers are interested in identifying suitable interventions by providing nutritional supplements like milk, meat, pulses *etc.*, to enhance the intellectual abilities of the students. There is also a need to examine the interchangeability of various nutritional supplements provided to children.

Gokul *et al.*, (2021) proposed a joint model, based on GLMM approach, for Kenya school lunch intervention study and suggested that the nutritional supplements show gradual improvement in cognitive behavior among the students. However, the choice of nutritional supplements, also play an important and unique role in promoting children's growth and development. There have been arguments and counter arguments through various studies that nutritional supplements like meat, milk and calories provide suitable interventions in the intellectual abilities of children. It is in this direction the present study considers CCC approach as an appropriate measure to study the agreement or otherwise of various nutritional supplements in enhancing the mental abilities (Carrasco 2010).

In this study, the concept of interchangeability in the treatment groups through CCC, intra class CC, proposed by Tsai and Lin (2018) is adopted for analyzing the longitudinal school lunch intervention data. Further, in psychological studies, data are often of count or ordinal in nature involving more number of zeroes. Thus, we considered Poisson GLMM for count data, and inflated models in the presence of zero observations to capture the measure of agreement through the concordance correlation co-++efficient. The performance and applicability of the CCC has been first demonstrated with a simulation study followed by the Kenyan real time dataset from a psychological study. The results of CCC based on the real data suggests lesser degree of agreement for the interchangeability among the four considered treatment groups. The study has established in a limited way that the nutritional supplement of milk as an appropriate intervention for the growth of intellectual abilities among children. On the whole, the methodology provides an insight to researchers working on longitudinal data with zeros to derive the benefit of using CCC method based on GLMM as a suitable measure of agreement.

Acknowledgements

The authors are grateful to an anonymous referee for the constructive and insightful comments made on the manuscript which led to a highly improved version.

References

Barchard, K. A. (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. *Psychological Methods*, **17**(**2**), 294-308.

Barnhart, H. X., Song, J., and Haber, M. (2005). Assessing intra, inter, and total agreement with replicated measurements. *Statistics in Medicine*, **24**, 1371-1384.

Carrasco, J. L. (2010). A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for over dispersed count data. *Biometrics*, **66**, 897-904.

Carrasco, J. L., and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, **59**, 849-858.

Carrasco, J. L., and Jover, L. (2005). Concordance correlation coefficient applied to discrete data. *Statistics in Medicine*, **24**, 4021-4034.

Carrasco, J. L., King, T. S., and Chinchilli, V. M. (2009). The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of Biopharmaceutical Statistics*, **19**, 90-105.

Ge, L., Tu, J. X., Zhang, H., Wang, H., He, H., and Gunzler, D. (2016). Modern methods for longitudinal data analysis, capabilities, caveats and cautions. *Shanghai Archives of Psychiatry*, **28(5)**, 293-300.

Gokul, T., Srinivasan, M. R., and Gallo, M. (2021). Modelling human intelligence using mixed model approach. *In: Mariani P., Zenga M. (eds) Data Science and Social Research II. Studies in Classification, Data Analysis, and Knowledge Organization*, 199-216.

King, T. S., and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, **20**, 2131-2147.

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic*, **15**, 155-163.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255-268.

Lin, L. I.-K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, **48**, 599-604.

Lin, L. I.-K., Hedayat, A. S., and Wu, W. A. (2007). Unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics*. **17**, 629-652.

Ma, Y., Tang, W., and Yu, Q. (2010). Modeling concordance correlation coefficient for longitudinal study data. *Psychometrika*, **75**(**1**), 99-119.

McCulloch, C. E., and Searle S. R. (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.

Neumann, C. G., Bwibo, N. O., Murphy, S. P., Sigman, M., Whaley, S., Allen, L. H., Guthrie, D., Weisszz, R. E., and Demment, M. W. (2003). Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in Kenyan

school children: background, study design and baseline findings. *The Journal of Nutrition*, **133**, 3941S-3949S.

Tsai, M. Y., and Lin, C. C. (2018). Concordance correlation coefficients estimated by variance components for longitudinal normal and Poisson data. *Computational Statistics and Data Analysis*, **121**, 57-90.

Tsai, M. Y. (2017). Concordance correlation coefficients estimated by generalized estimating equations and variance components for longitudinal repeated measurements. *Statistics in Medicine*, **36**, 1319-1333.

Zhu, H., Luo, S., and DeSantis, S. M. (2015). Zero-inflated count models for longitudinal measurements with heterogeneous random effects. *Statistical Methods in Medical Research*, **26**(**4**), 1-16.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 255 - 266

Automation of USDA Triangle Soil Texture Classification Using Finite State Machine: A Novel Conceptual Modeling Approach

Pradeep H K[1], Jasma Balasangameshwara[2], K Rajan[3], Madhu M[4], Archana B K[5]

[1]Department of Computer Science & Engineering, JSS Academy of Technical Education, Bengaluru & Research Scholar, Visvesvaraya Technological University – RRC, Belgaum, Karnataka, India.
[2]Department of Computer Science & Technology, Dayananda Sagar University, Bengaluru, Karnataka, India.
[3]ICAR - Indian Institute of Soil and Water Conservation, Ooty, India.
[4]Indian Institute of Soil & Water Conservation (IISWC), Research Centre, Sunabeda, Koraput Dist,Odisha, India.
[5]Department of Electronics & Communication Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India.

Received: 23 June 2020; Revised: 18 December 2020; Accepted: 31 December 2020

Abstract

The USDA triangle is the most widely used model for soil texture classification. The problem with USDA triangle model was, it needs critical analysis for identification of soil textural class. To simplify the soil textural class prediction process the USDA triangle model was automated using finite state machine technique. The experimental results exhibited the equivalence between USDA triangle and automated soil textural classification model. The proposed automated model is efficient, reliable and user-friendly for prediction of soil textural class.

Key words: Clay fraction, Sand fraction, Silt fraction, Software model.

1. Introduction

Soil includes supplements, water, minerals and micro-organism, which gives living environment to all plants. Jha and Ahmad (2018). The dirt quality varies overtime due to changes in properties. Karlen *et al.* (2003), Ghosh *et al.* (2017), Doran *et al.* (1999), Rajan *et al.* (2016). The organic and physical property of soil has immense impact on fertility. Schoenholtz *et al.* (2000), Crittenden and de Goede (2016). Soil fertility is the ability to give supplements to the yield development. Peigne *et al.* (2017). The poor soil surface influences hydro coherent and biochemical procedures. Moncada *et al.* (2017). Soil properties variation has high effect on irrigation management. The dirt properties and land suitability are integral factor for structuring water system frameworks. Cho *et al.* (2016). Artificial Intelligence approaches are efficiently used for soil classification. Wu *et al.* (2018), Sirsat *et al.* (2017). The dirt texture has high impact on tillage practices, plant nutrients and liming application. Jovic *et al.* (2019). Modeling soil classes play crucial role in irrigation system water

Corresponding Author: Pradeep H.K.
Email: phk.contact@gmail.com

productivity. Zeng *et al.* (2016). The soil classification has long history, wherein the USDA triangle model is the widely used model worldwide. Hartermink (2015). The objective of the proposed study is automation of USDA triangle model. The finite state machine (FSM) approach is most widely used technique for automation of multidiscipline theoretical concepts. In the proposed model the USDA triangle model is automated and also retained the logical equivalence of manual approach over soil texture classification. In USDA triangle model, for many cases there are multiple transitions for a same sand, silt or clay fraction value, hence we have chosen non-deterministic finite state machine to design automated framework of USDA triangle soil texture classification.

2. Materials and Methods

The USDA triangle soil texture model and FSM concepts are integrated to design soil texture automation framework. An input string is passed to the model one character at a time, in which the model considers the current state and the new character and chooses the next state. In FSM model one of the states is designated as start state and consists of one or more final states. Final or accepting states are represented using double circle. In FSM model, if it runs out of the input and halts at final state then it accepts the input string otherwise, it rejects. The number of steps FSM executes is exactly equal to number of characters present in the string. The FSM has two variants, Non-Deterministic Finite State Machine (NDFSM) and Deterministic Finite State Machine (DFSM). In NDFSM, there will be multiple moves for one input symbol, the behavior is non-deterministic. In this section the USDA triangle model represented in Figure 1 is automated using NDFSM model. Groenendyk*et al.* (2015).

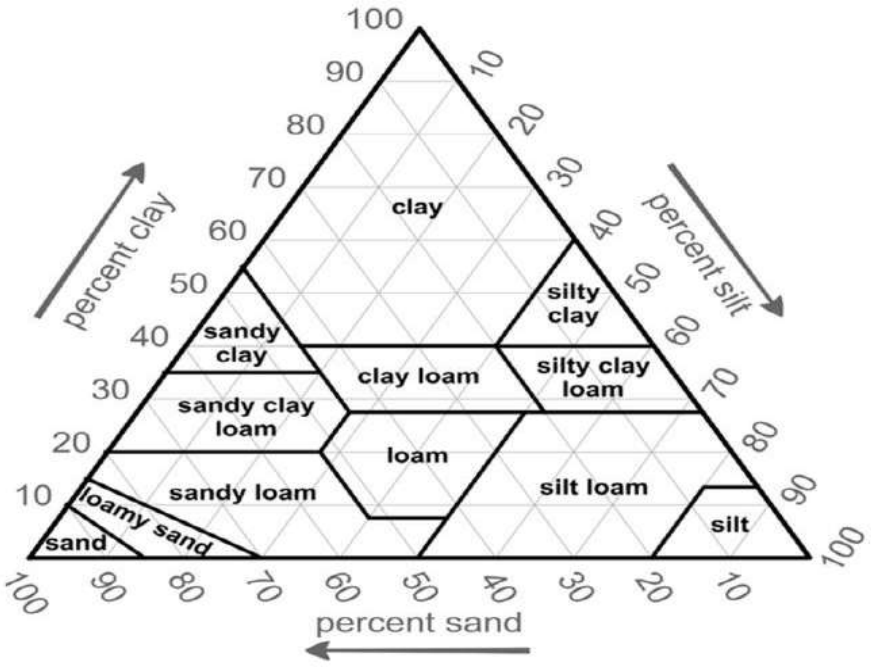


Figure 1: USDA triangle soil textural classification model

The sand, silt and clay fraction threshold values of twelve USDA triangle model classes are considered to identify the input parameters for NDFSM framework. The NDFSM model variables are defined in Table 1.

Table 1: Preprocessing of USDA triangle soil texture data to fit into NDFSM

Sand Fraction (%)	Sand Input Variables	Silt fraction (%)	Silt Input Variables	Clay fraction (%)	Clay Input Variables
0-20	*a*1	0-15	*b*1	0-7	*c*1
20-23	*a*2	15-20	*b*2	7-10	*c*2
23-42	*a*3	20-28	*b*3	10-12	*c*3
42-45	*a*4	28-30	*b*4	12-15	*c*4
45-50	*a*5	30-40	*b*5	15-20	*c*5
50-52	*a*6	40-50	*b*6	20-27	*c*6
52-65	*a*7	50-52	*b*7	27-35	*c*7
65-70	*a*8	52-60	*b*8	35-40	*c*8
70-80	*a*9	60-73	*b*9	40-55	*c*9
80-85	*a*10	73-80	*b*10	55-60	*c*10
85-90	*a*11	80-87	*b*11	60-100	*c*11
90-100	*a*12	87-100	*b*12	-	-

2.1.1. Design of automated model for soil texture classification using NDFSM

The NDFSM approach is one of easiest method of finite automata used for designing abstract machines. In the proposed model automated soil texture classification model is designed using NDFSM. NDFSM is formally defined as set of five attributes which are described in the following section for USDA triangle model.

NDFSM = {$S, \Sigma, F, s0, \delta$}

States (S): {$s0$, $s1$, $s2$, $s3$, $s4$, $s5$, $s6$, $s7$, $s8$, $s9$, $s10$, $s11$, $s12$, $s13$, $s14$, $s15$, $s16$, $s17$, $s18$, $s19$, $s20$, $s21$, $s22$, $s23$, $s24$, $s25$, $s26$, $s27$, $s28$, $s29$, $s30$, $s31$, $s32$, $s33$, $s34$}.

Start State is $s0$ and $\in S$.

A state is a circumstance of a framework relying upon past sources of info and causes a response on current information sources. States indicate the step by step procedure for soil textural class identification based on the sand, silt and clay fraction input. Suppose if sand fraction is 85-100%, silt fraction is 0-15% and clay fraction is 0-10% then in the FSM model state transitions takes place in the path $s0$->$s1$->$s2$->$s3$. The state $s0$ is the initial state, $s1$ and $s2$ are intermediate states and $s3$ is the final state, which represents sand soil textural class. Suppose if sand fraction is 70-90%, silt fraction is 0-30% and clay fraction is 0-15% then in the FSM model state transitions takes place in the path $s0$->$s4$->$s5$->$s6$. The state $s0$ is the initial state, $s4$ and $s5$ are intermediate states and $s6$ is the final state, which represents loamy sand textural class. Similarly for all the 12 soil texture classes there are different state transition paths based on the sand, silt and clay fraction values which are represented in Figure 2.

Input Alphabets (Σ): {$a1$, $a2$, $a3$, $a4$, $a5$, $a6$, $a7$, $a8$, $a9$, $a10$, $a11$, $a12$, $b1$, $b2$, $b3$, $b4$, $b5$, $b6$, $b7$, $b8$, $b9$, $b10$, $b11$, $b12$, $c1$, $c2$, $c3$, $c4$, $c5$, $c6$, $c7$, $c8$, $c9$, $c10$, $c11$}. The sand fraction values are represented using the template "ai", in which the symbol 'a' represents the sand fraction and 'i' represents the parameter number. The parameter number is assigned

based on the sand fraction threshold values of USDA triangle model soil textural classes. Suppose if sand fraction value is 0-20% then the corresponding input parameter is mapped as "*a*1". Suppose if sand fraction value is 20-23% then the corresponding input parameter is mapped as "*a*2". Likewise, for all the unique sand fraction range the input parameters are assigned, which are reported in Table 1.

The silt fraction values are represented using the template "*bi*", in which the symbol '*b*' represents the silt and '*i*' represents the parameter number. The parameter number is assigned based on the silt fraction threshold values of USDA triangle model soil textural classes. Suppose if silt fraction value is 0-15% then the corresponding input parameter is mapped as "*b*1". Suppose if silt fraction value is 15-20% then the corresponding input parameter is mapped as "*b*2". Likewise, for all the unique silt fraction range the input parameters are assigned, which are reported in Table 1.

The clay fraction values are represented using the template "*ci*", in which the symbol '*c*' represents the clay and '*i*' represents the parameter number. The parameter number is assigned based on the clay fraction threshold values of USDA triangle model soil textural classes. Suppose if clay fraction value is 0-7% then the corresponding input parameter is mapped as "*c*1". Suppose if clay fraction value is 7-10% then the corresponding input parameter is mapped as "*c*2". Likewise, for all the unique clay fraction range the input parameters are assigned, which are reported in Table 1.

Final States (F): {*s*3, *s*6, *s*9, *s*12, *s*15, *s*18, *s*21, *s*24, *s*26, *s*29, *s*31, *s*34}

In USDA triangle model there are 12 soil texture classes accordingly in FSM model 12 final states are defined. Each final state represents a soil texture class. The state "*s*3" represents sand class, "*s*6" represents loamy sand class, "*s*9" represents sandy loam, "*s*12" represents loam, "*s*15" represents silty loam, "*s*18" represents silt, "*s*21" represents sandy clay loam, "*s*24" represents clay loam, "*s*26" represents silty clay loam, "*s*29" represents sandy clay, "*s*31" represents silty clay and "*s*34" represents clay soil texture. For all valid input patterns the FSM model halts at one of the final state based on sand, silt and clay fraction values.

Transition functions (δ): It maps from S (state) $\times$ Σ (Input symbol) = S (States), the outcome of transition function can have set of states in NDFSM}.In the following section the NDFSM model is designed for soil texture classification considering the transition functions represented in Table 2.

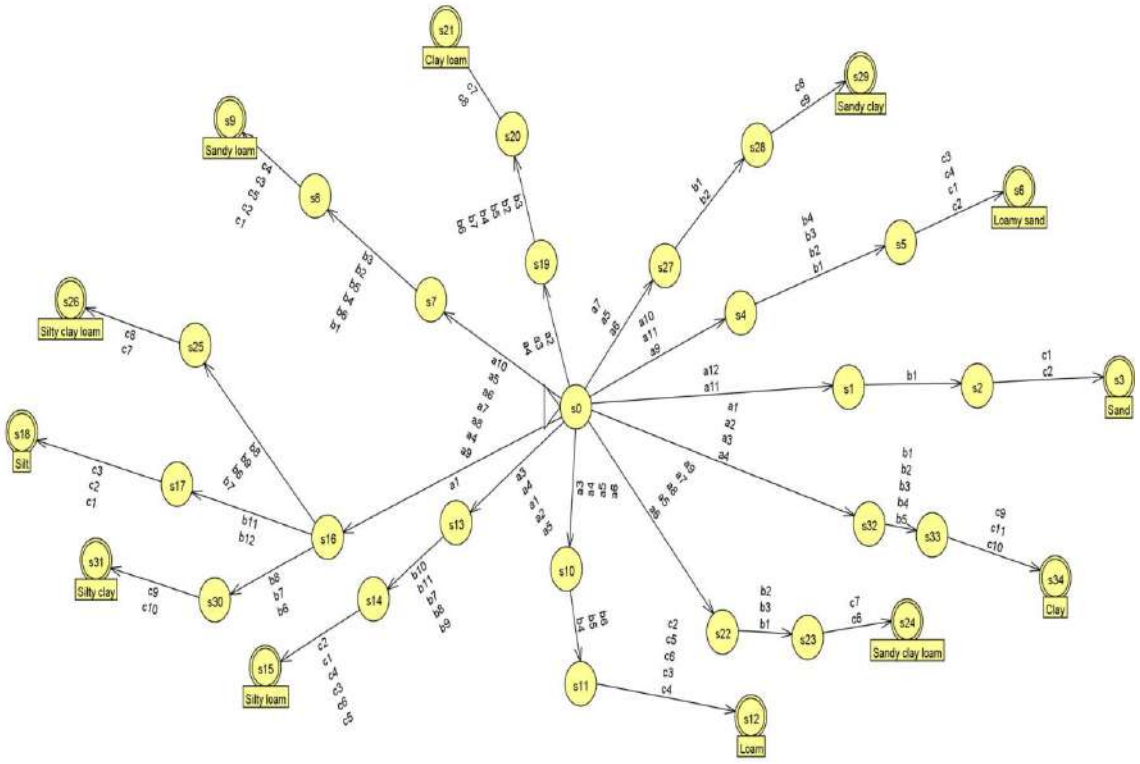


Figure 2: USDA triangle automated model for soil textural classification

Transition functions of proposed NDFSM model are highlighted in the following section over all the input symbols. For each state the possible movements on all the input parameters are represented using transition function. Suppose if the input comprises of sand fraction 91% then form start state *s*0 on input "*a*12" FSM moves to state "*s*1", followed by suppose if silt fraction is 5% then from state *s*1 on input "*b*1" FSM moves to state "*s*2" and followed by suppose if clay fraction is 6% then from state "*s*2" on input "*c*1" FSM moves to state "*s*3" and the corresponding input pattern is accepted as sand soil texture class.

Table 2: Transition functions defined for automated soil texture classification

Transition functions for states *s*0,*s*1,*s*2,*s*4,*s*5,*s*6,*s*7	Transition functions for states *s*8,*s*10,*s*11,*s*13,*s*14	Transition functions for states *s*16,*s*17,*s*19,*s*20,*s*22	Transition functions for states *s*23,*s*25,*s*27,*s*28, *s*30,*s*31,*s*32,*s*33
Transitions from state *s*0: *(s*0, *a*1*) = (s*13, *s*16, *s*32*)* *(s*0, *a*2*) = (s*13, *s*19, *s*32*)* *(s*0, *a*3*) = (s*10, *s*13, *s*19, *s*32*)* *(s*0, *a*4*) = (s*7, *s*10, *s*13, *s*19, *s*32*)* *(s*0, *a*5*) = (s*7, *s*10, *s*13, *s*22, *s*27*)* *(s*0, *a*6*) = (s*7, *s*10, *s*22, *s*27*)* *(s*0, *a*7*) = (s*7, *s*22, *s*27*)* *(s*0, *a*8*) = (s*7, *s*22*)* *(s*0, *a*9*) = (s*4, *s*7, *s*22*)* *(s*0, *a*10*) = (s*4, *s*7*)*	Transitions from state *s*8: *(s*8, *c*1*) =* ***(s*9*)*** *(s*8, *c*2*) =* ***(s*9*)*** *(s*8, *c*3*) =* ***(s*9*)*** *(s*8, *c*4*) =* ***(s*9*)*** *(s*8, *c*5*) =* ***(s*9*)*** Transitions from state *s*10: *(s*10, *b*4*) = (s*11*)* *(s*10, *b*5*) = (s*11*)* *(s*10, *b*6*) = (s*11*)* Transitions from	Transitions from state *s*16: *(s*16, *b*11*) = (s*17*)* *(s*16, *b*12*) = (s*17*)* *(s*16, *b*6*) = (s*25*)* *(s*16, *b*7*) = (s*25*)* *(s*16, *b*8*) = (s*25*)* *(s*16, *b*9*) = (s*25*)* *(s*16, *b*6*) = (s*30*)* *(s*16, *b*7*) = (s*30*)* *(s*16, *b*8*) = (s*30*)* Transitions from state *s*17:	Transitions from state *s*23: *(s*23, *c*6*) =* ***(s*24*)*** *(s*23, *c*7*) =* ***(s*24*)*** Transitions from state *s*25: *(s*25, *c*7*) =* ***(s*26*)*** *(s*25, *c*8*) =* ***(s*26*)*** Transitions from state *s*27: *(s*27, *b*1*) = (s*28*)* *(s*27, *b*2*) = (s*28*)* Transitions from

(s0, a11) = (s1,s4) *(s0, a12) = (s1)* Transitions from state *s1*: *(s1, b1) = (s2)* Transitions from state *s2*: *(s2, c1) =* ***(s3)*** *(s2, c2) =* ***(s3)*** Transitions from state *s4*: *(s4, b1) = (s5)* *(s4, b2) = (s5)* *(s4, b3) = (s5)* *(s4, b4) = (s5)* Transitions from state *s5*: *(s5, c1) =* ***(s6)*** *(s5, c2) =* ***(s6)*** *(s5, c3) =* ***(s6)*** *(s5, c4) =* ***(s6)*** Transitions from state *s7*: *(s7, b1) = (s8)* *(s7, b2) = (s8)* *(s7, b3) = (s8)* *(s7, b4) = (s8)* *(s7, b5) = (s8)* *(s7, b6) = (s8)*	state *s11*: *(s11, c2) =* ***(s12)*** *(s11, c3) =* ***(s12)*** *(s11, c4) =* ***(s12)*** *(s11, c5) =* ***(s12)*** *(s11, c6) =* ***(s12)*** Transitions from state *s13*: *(s13, b7) = (s14)* *(s13, b8) = (s14)* *(s13, b9) = (s14)* *(s13, b10) = (s14)* *(s13, b11) = (s14)* Transitions from state *s14*: *(s14, c1) =* ***(s15)*** *(s14, c2) =* ***(s15)*** *(s14, c3) =* ***(s15)*** *(s14, c4) =* ***(s15)*** *(s14, c5) =* ***(s15)*** *(s14, c6) =* ***(s15)***	*(s17, c1) =* ***(s18)*** *(s17, c2) =* ***(s18)*** *(s17, c3) =* ***(s18)*** Transitions from state *s19*: *(s19, b2) = (s20)* *(s19, b3) = (s20)* *(s19, b4) = (s20)* *(s19, b5) = (s20)* *(s19, b6) = (s20)* *(s19, b7) = (s20)* Transitions from state *s20*: *(s20, c7) =* ***(s21)*** *(s20, c8) =* ***(s21)*** Transitions from state *s22*: *(s22, b1) = (s23)* *(s22, b2) = (s23)* *(s22, b3) = (s23)*	state *s28*: *(s28, c8) =* ***(s29)*** *(s28, c9) =* ***(s29)*** Transitions from state *s30*: *(s30, c9) =* ***(s31)*** *(s30, c10) =* ***(s31)*** Transitions from state *s32*: *(s32, b1) = (s33)* *(s32, b2) = (s33)* *(s32, b3) = (s33)* *(s32, b4) = (s33)* *(s32, b5) = (s33)* Transitions from state *s33*: *(s33, c9) =* ***(s34)*** *(s33, c10) =* ***(s34)*** *(s33, c11) =* ***(s34)***

3. Results and Discussions

An analysis has been planned to scrutinize 12 classes in USDA soil textural triangle and its soil fraction ranges and developed a soft computing model to arrive at textural class. The objective of the proposed work is automation of USDA triangle soil texture classification concept using NDFSM. The data set comprises of 5000 records, in which each sample has sand, silt and clay particle size distribution. The summation of all three parameters particle size must be exactly 100 for all input samples. The 70% data was used for training, 20% data was used for testing and 10% data was used for validation. The testing and validation phase of experiment results exhibited the equivalence between USDA triangle model and FSM based automated software model. The model has been traced for many observed input patterns using JFALP. Rodger and Gramond (1998).The validation phase of the NDFSM soil texture classification model also obtained equivalence with USDA triangle over soil texture classification.

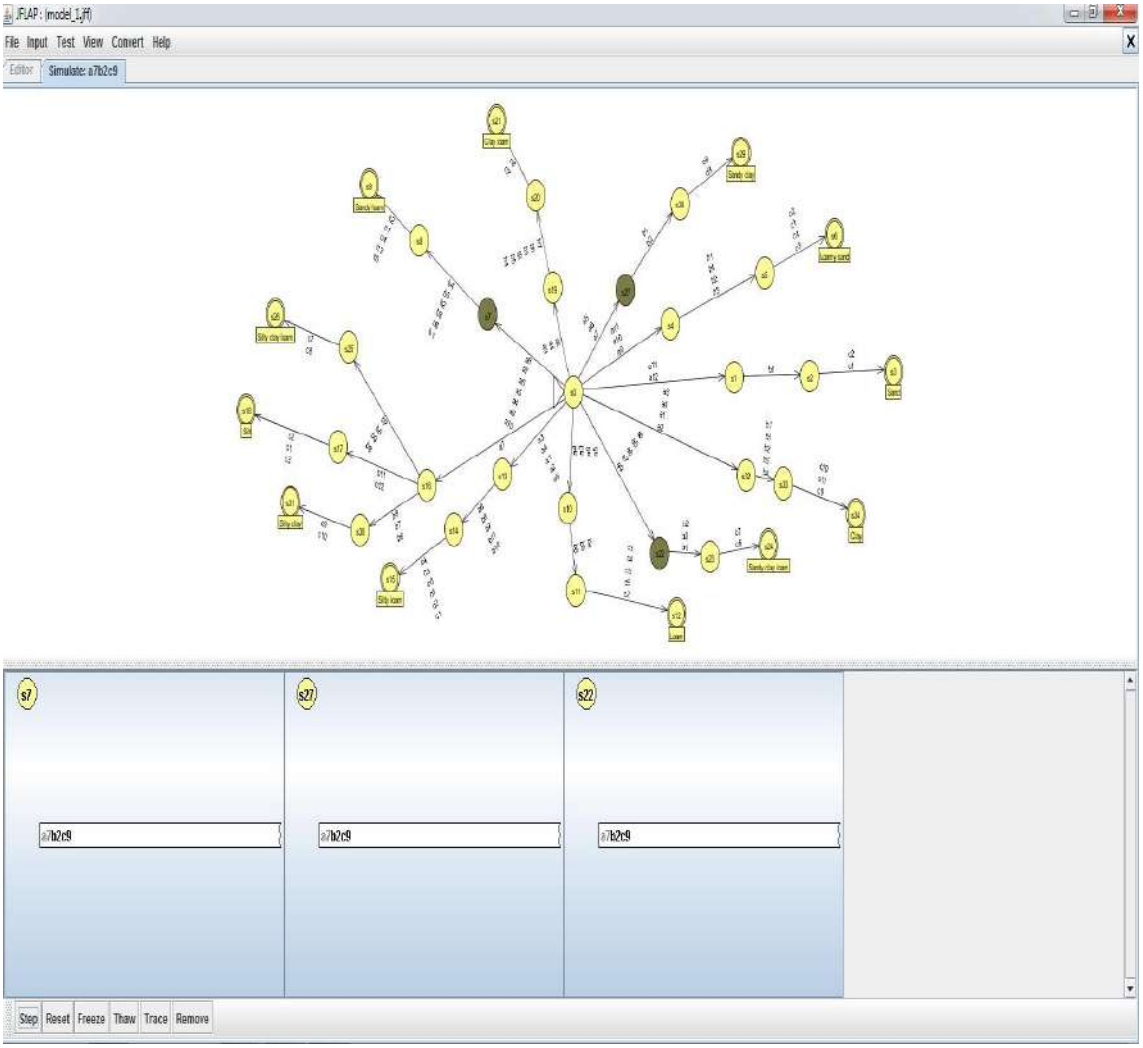


Figure 3: NDFSM model step by state tracing over the sand fraction value "*a7*"

The input pattern *"a7b2c9"* was traced using Java Formal Languages and Automata Package (JFLAP), in which the state transitions are observed over the sand fraction input *"a7"*. The transitions indicate the possible movements from state *s*0 over the input *"a7"* are *s*7, *s*22 and *s*27 which are highlighted in Figure 3.

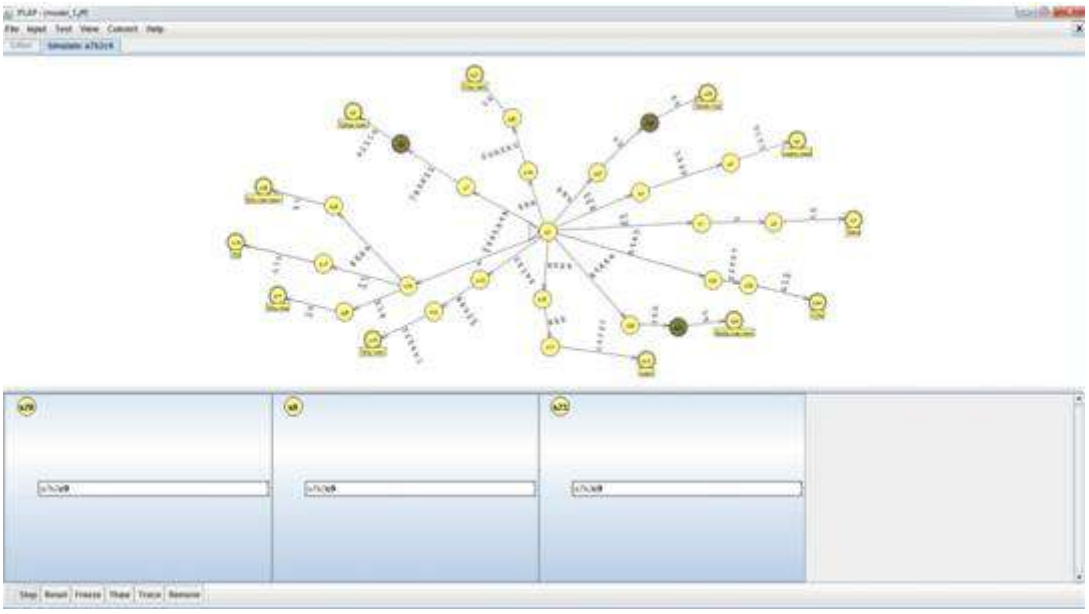


Figure 4: NDFSM model step by state tracing over the silt fraction input "*b*2"

The input pattern *"a*7*b*2*c*9*"* was traced using JFLAP, in which the state transitions are observed over sand fraction input *"a*7*"* followed by the silt fraction *"b*2*"*. The transitions indicate the possible movements over the input *"a*7*b*2*"* are *s*8, *s*23 and *s*28 which are highlighted in Figure 4.

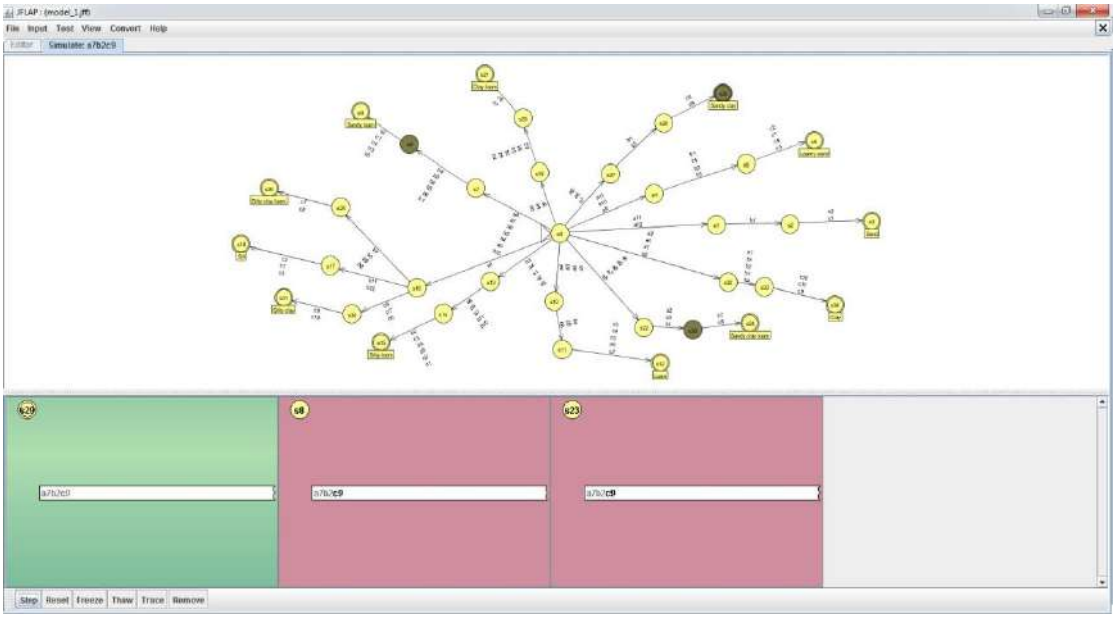


Figure 5: NDFSM model step by state tracing over the input symbol '*c*9'

The input pattern *"a*7*b*2*c*9*"* was traced using JFLAP, in which the state transitions are observed over sand fraction input *"a*7*"* followed by the silt fraction *"b*2*"*and followed by clay fraction *"c*9*"*. The transitions indicate the possible movements over the input *"a*7*b*2*c*9*"* are *s*29 which is final state highlighted in Figure 5 and represents sandy clay texture. Initially the execution starts from start state *s*0 over the input symbol "*a*7", from *s*0 the control moves

to *s*7, *s*27 and *s*22 because from *s*0 there are transitions to all the above mentioned states on the input symbol "*a*7*"*. Further, from state *s*7 on input symbol "*b*2" the control moves to state *s*8, from state *s*22 on input symbol '*b*2' control moves to state *s*23 and from state *s*27 it moves to state *s*28 over the input "*b*2". Finally, the transitions are checked from the states *s*8, *s*23 and *s*28 over the input symbol "*c*9", wherein only the state *s*28 has transition to the state *s*29. The state *s*29 is the accepting state because it's represented using double circle and it accepts the input pattern and predicts the soil texture as Sandy clay for the input "*a*7*b*2*c*9". The same pattern is also traced using state by state execution method, in which the path obtained is *s*0->*s*27->*s*28->*s*29 and the corresponding process is represented in Figure 6. Automated model has been validated considering soil textural data set of Jangamakotte and Bhaktarahallipedonds of Kolar district, Karnataka, India. Rajan *et al.* (2014).

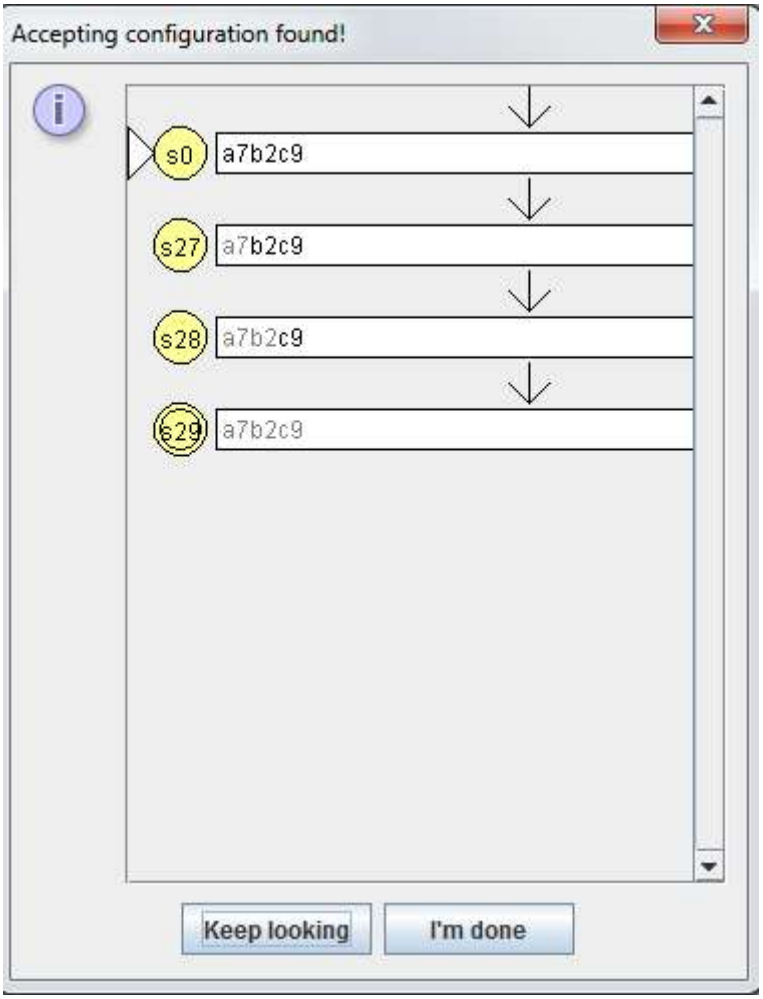


Figure 6: NDFSM model state by state tracing over the input pattern "*a*7*b*2*c*9"

The input pattern *"a*9*b*9*c*9*"* was tested using automated model which is represented in Figure 7.

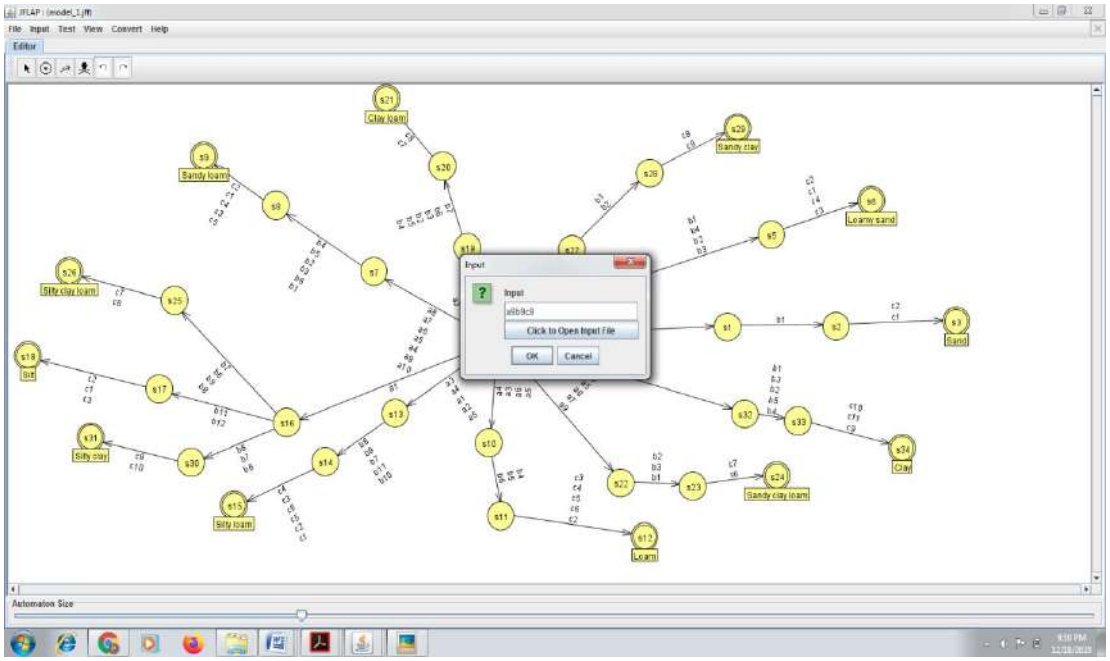


Figure 7: The input pattern *"a9b9c9"* was tested using automated model

The symbol *"a*9*"* of input pattern represents the sand fraction range as 70-80% and the symbol *"b*9*"* of input pattern represents the silt fraction range as 60-73% and also the symbol *"c*9*"* of input pattern represents clay fraction as 40-55%. Suppose if we consider the sample value of sand fraction as 71%, silt fraction as 61% and also clay fraction as 41%, then summation of all there particles size would be 173. For any soil texture sample the summation of sand, silt and clay fraction size must be exactly 100 otherwise the input sample is considered as invalid. The automated model rejected sample input is represented in Figure 8.

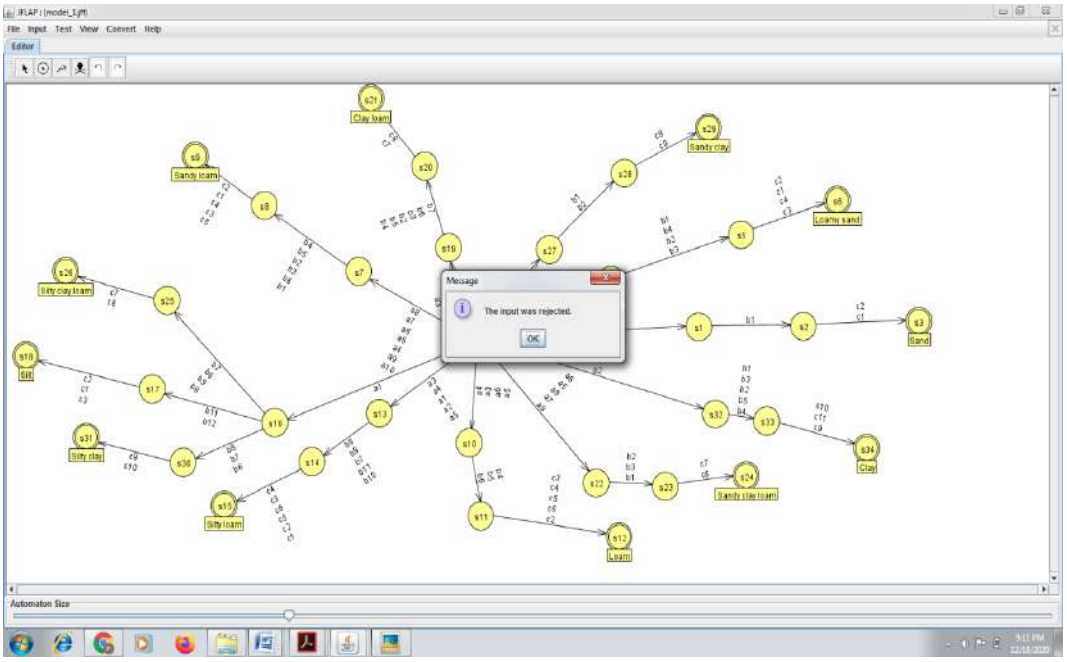


Figure 8: The input pattern *"a9b9c9"* was rejected by automated model

Additionally, multiple soil profile data records can be loaded and predicted at the same time using JFLAP tool and the corresponding details are represented in Figure 9.

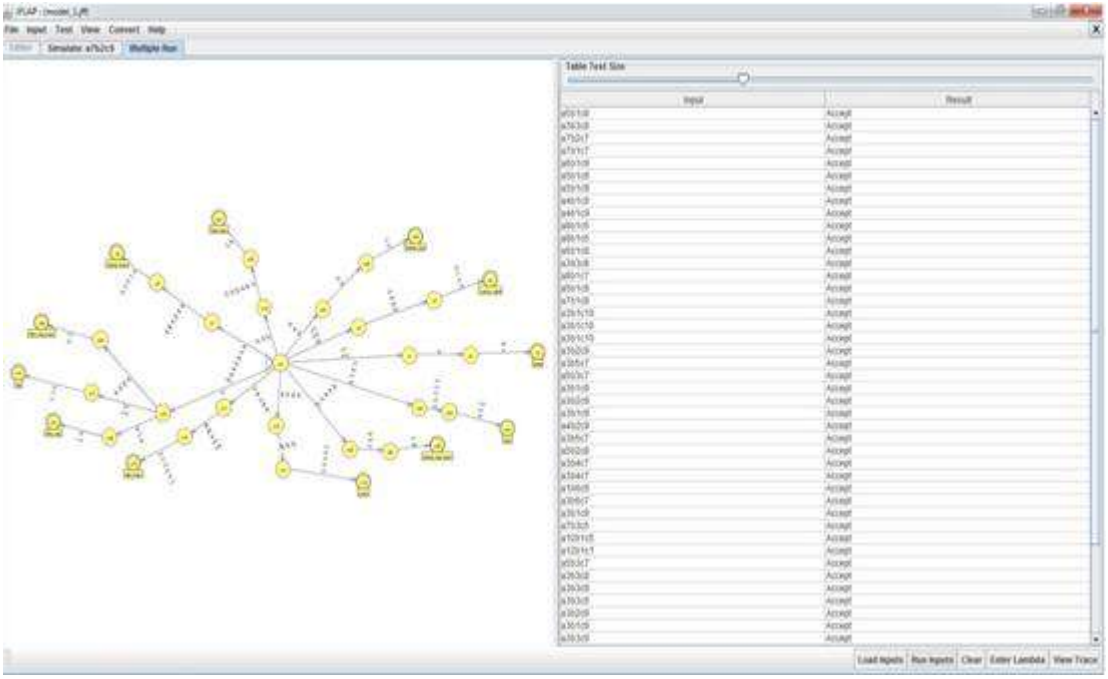


Figure 9: Validation results of automated NDFSM soil texture classification model

4. Conclusion

In this paper the USDA Triangle soil texture classification model is automated using the proposed Non-Deterministic Finite State Machine (NDFSM). The experimental results of NDFSM model exhibited the logical equivalence with USDA triangle model during the testing and validation phase over soil texture classification. The NDFSM soil texture classification model was validated using laboratory tested soil profile dataset. For all the validated patterns the predicted texture of NDFSM model was same as USDA triangle soil texture classification. The proposed automated model simplifies the job of soil texture class prediction.

Acknowledgments

The authors express their gratefulness to the anonymous reviewer for giving some useful suggestions that led to considerable improvement in the presentation of the results. The authors also express their thanks to the Chief Editor for his suggestions on restructuring the contents for better expression.

References

Cho, Y., Sudduth, K. A., and Chung, S. O. (2016). Soil physical property estimation from soil strength and apparent electrical conductivity sensor data. *Biosystems Engineering*,**152**, 68–78.

Crittenden, S. J., and de Goede, R. G. M. (2016). Integrating soil physical and biological properties in contrasting tillage systems in organic and conventional farming. *European Journal of Soil Biology*,**77**, 26–33.

Doran, J. W., Jones, A. J., Arshad, M. A., and Gilley, J. E. (1999). Determinants of soil quality and health. *Soil quality and soil erosion*, **36.**

Ghosh, P. K., Palsaniya, D. R., and Kumar, T. K. (2017). Resource Conservation Technologies for Sustainable Soil Health Management. *Adaptive Soil Management : From Theory to Practices*, 161–187.

Groenendyk, D. G., Ferré, T. P., Thorp, K. R., and Rice, A. K. (2015). Hydrologic-Process-Based Soil Texture Classifications for Improved Visualization of Landscape Function. *PLOS ONE* **10**, 0131299–0131299.

Hartemink, A. E. (2015). The use of soil classification in journal papers between 1975 and 2014. *Geoderma Regional*, **5**, 127–139.

Jha, S. K., and Ahmad, Z. (2018). Soil microbial dynamics prediction using machine learning regression methods.

Jović, B., Ćirić, V., Kovačević, M., Šeremešić, S., and Kordić, B. (2019). Empirical equation for preliminary assessment of soil texture. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **206**, 506–511.

Karlen, D. L., Ditzler, C. A., and Andrews, S. S. (2003). Soil quality: why and how?

Moncada, M. P., Penning, L. H., Timm, L. C., Gabriels, D., and Cornelis, W. M. (2017). Visual examination of changes in soil structural quality due to land use. *Soil and Tillage Research*, **173**, 83–91.

Peigné, J., Vian, J., Payet, V., and Saby, N. P. A. (2017). Soil & Tillage Research Soil fertility after 10 years of conservation tillage in organic farming. *Soil & Tillage Research*, **175**, 194–204.

Rajan, K., Natarajan, A., Anilkumar, K., Gowda, R., and Haris, A. (2014). Assessment of some soil physical indicators in severely eroded lands of southern Karnataka. *Indian J. Soil Conserv*,**42**, 154–163.

Rajan, K., Natarajan, A., Thilagam, V. K., Kumar, K. A., Dinesh, D., Alam, N. M., and Gowda, C. R. (2016). Clay dispersion induced by changes in some soil properties in undulating salt-affected landscapes of southern Karnataka. India. *Current Science*, 874–883.

Rodger, S. H., and Gramond, E. (1998). JFLAP: An Aid to Studying Theorems in Automata Theory. *ACM SIGCSE Bulletin inroads*, **30**, 302–302.

Schoenholtz, S. H., Miegroet, H. Van, and Burger, J. A. (2000). A review of chemical and physical properties as indicators of forest soil quality : challenges and opportunities.

Sirsat, M. S., Cernadas, E., Fernández-Delgado, M., and Khan, R. (2017). Classification of agricultural soil parameters in India. *Computers and Electronics in Agriculture*, **135**, 269–279.

Wu, W., Li, A. D., He, X. H., Ma, R., Liu, H. B., and Lv, J. K. (2018). A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China.*Computers and Electronics in Agriculture*, **144**, 86–93.

Zeng, R., Zhang, G. L., Li, D. C., Rossiter, D. G., and Zhao, Y. G. (2016). How well can VNIR spectroscopy distinguish soil classes? *Biosystems Engineering*, **152**, 117–125.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 2, 2021 (New Series), pp 267 – 279

Frequentist Predictive Inference for Wind Direction Data Under *l*-modal Circular Normal Model Through Sufficiency Approach

Sahana Bhattacharjee and Kishore Kumar Das
Department of Statistics, Gauhati University, Guwahati-781014, Assam, India

Received: 03 May 2020; Revised: 05 January 2021; Accepted 08 January 2021

Abstract

In this paper, the estimation of the future density of wind direction conditioned on the past and present wind direction data using the Sufficiency Approach of Predictive Inference under the *l*-modal Circular Normal model, followed by the equal tail area predictive interval estimation has been done. Point predictive estimator of future observation, termed as the frequentist predictive point estimator under the circular loss function has been obtained Finally, some basic properties of the estimator have been explored.

Keywords: Sufficiency approach of predictive inference; *l*-modal circular normal model; Equal tail area predictive interval estimation; Circular loss function; frequentist predictive point estimator.

1. Introduction

The prediction of the pattern of future occurrences, based on the occurrences in past, is an important aim of statistics and according to some authors; it is the sole aim of this subject. The object of interest to be predicted can be a single value, a set of values or a function of these. The literature boasts of a number of techniques to obtain predictive likelihoods and density functions. Cox and Hinkley (1974) had initially conditioned the data on a minimal sufficient statistic of the parameter and it was Butler (1986) who later on incorporated the future observation in the data and suggested the expression for conditional predictive likelihood based on the minimal sufficient statistic. This constitutes the Sufficiency Approach. In the Bayesian Approach of predictive density estimation, the conditional distribution of the future observation(s) given the past data is obtained simply by marginalization of the joint distribution of the future observations and the population parameter(s) with respect to the parameter. For this purpose, at the outset, the prior densities of the parameters are assumed to be known. The Profile or Maximum Likelihood method consists in predicting the density of future observation from the maximum likelihood function based on the maximum likelihood estimate of both the given and future observation.

After having predicted the future observations(s), one might be interested in carrying out both the point and interval estimation based on the predictive density, followed by evaluating the error or loss incurred in predicting the true value of the observation by its estimator. The loss incurred can be quantified using a loss function. In prediction problems, as stated by Hennig and Kutlukaya (2007), the quality of a predictor is judged with the help

Corresponding Author: Sahana Bhattacharjee
sahana.bhattacharjee@hotmail.com

of loss function, which depends on the observed value and predicted value of the observation. Jammalamadaka and Sen Gupta (1998) had derived the predictive density estimation of the future observation given present and previous data and then carried out the predictive Highest Posterior Density (HPD) interval estimation under von Mises model in circular case and von Mises-Fisher model in spherical case. In predictive analysis concerning circular data, the usual linear loss functions are not well-defined. Gelfand and Ghosh (1998) had proposed the Squared Predicted Errors (SPE) loss function and had used it to choose the best fitting model to circular data by minimizing the posterior predictive loss. Ravindran and Ghosh (2012) had proposed the Absolute Predicted Errors (APE) loss function and used it to choose the best fitting circular model by minimizing this loss with respect to the posterior predictive density. Under the circular loss defined by Sen Gupta and Maitra (1998), the same authors had studied the best equivariance and admissibility property of the maximum likelihood estimator of the mean direction for a single von Mises distribution and also for that of the several independently distributed circular normal distributions. In the linear statistics literature, the Bayes estimator of the parameters of different distributions is obtained by minimizing the posterior loss/predictive loss under different loss functions. Several properties of these loss functions have also been explored. However, in the circular statistics predictive inference literature, the estimation of parameters by minimization of predictive density has not been attempted yet. Another interesting prospect that still remains unexplored is studying the properties of these estimators. Keeping in view these points, the objectives of this paper have been decided upon.

This paper attempts to predict the future density of wind direction conditioned on the past and present wind direction data using the Sufficiency Approach of Predictive Inference under the *l*-modal Circular Normal model and then carry out the equal tail area predictive interval estimation. Further, the predictive point estimation of the future observation under the circular loss function and for the same model has been carried out. Finally, some basic properties of the estimators obtained under the circular loss function are studied.

For achieving the objectives of this paper, daily data on wind direction for Dibrugarh Meteorological station located in Assam, measured during morning for the Monsoon season (June-September) during the years 2012 and 2013 has been procured from the Regional Meteorological Center, Guwahati.

2. Predictive Density Estimation of the Future Observation Through Sufficiency Approach Under the *l*-modal Circular Normal Model

Suppose that $\theta_1, \theta_2, \dots, \theta_n$ is a sample from *l*-modal Circular Normal distribution, whose p.d.f is given by

$$f(\alpha; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} exp\{\kappa \cos l(\theta - \mu)\} \qquad 0 < \theta, \mu < 2\pi;\ \mu < \frac{2\pi}{l}$$

Here, μ and κ represent the mean direction and concentration parameter of the population respectively and the parameter *l* stands for the number of modes of the distribution (Rao and Sengupta, 2001).

Upon computation, while fixing the value of *l* as the number of modes in the sample as it appears in the corresponding histogram, the maximum likelihood estimators of μ and κ have been found to be

$\hat{\mu} = \frac{1}{l} arctan\left\{\frac{\sum_{i=1}^{n} \sin(l\theta_i)}{\sum_{i=1}^{n} \cos(l\theta_i)}\right\}$ and $\hat{\kappa} = A^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} \cos l(\theta - \mu)\right]$

$A^{-1}(.)$ being the inverse function of the ratio of the first and zeroth order Bessel functions of the first kind, both of which are evaluated at a specified non-negative real number.

Now, for a known l, $(C_{l,n}, S_{l,n})$ is a minimal sufficient statistic for (μ_0, κ) (Rao and SenGupta, 2001, pp.209), where $C_{l,n} = \sum_{i=1}^{n} \cos l\theta_i$ and $S_{l,n} = \sum_{i=1}^{n} \sin l\theta_i$. Based on the $(n+1)$ observations, for which $(C_{l,n+1}, S_{l,n+1})$ is the sufficient statistic, the conditional probability of the future observation θ_{n+1} given $\theta_1, \theta_2, \dots, \theta_n$ is given by

$$Pr(\theta_{n+1}|\theta_1, \theta_2, \dots, \theta_n) = \frac{Pr(\theta_1, \theta_2, \dots, \theta_n, \theta_{n+1})}{Pr(C_{l,n+1}, S_{l,n+1})}$$

$$= \frac{\left[\frac{1}{2\pi I_0(\kappa)} exp\{\kappa \cos l(\theta_i - \mu)\}\right]^{n+1}}{\left[\frac{1}{2\pi I_0(\kappa)} exp\{\kappa(C_{l,n} \sin\mu + S_{l,n} \cos\mu)\}\right] l\psi_n(r_l)}$$

$$= \frac{1}{(2\pi)^n \psi_{n+1}\left(\sqrt{{C_{l,n+1}}^2 + {S_{l,n+1}}^2}\right)} \quad (1)$$

where $C_{l,n+1} = \sum_{i=1}^{n+1} \cos l\theta_i$, $S_{l,n+1} = \sum_{i=1}^{n+1} \sin l\theta_i$ and

$$\psi_n(r) = \int_0^\infty J_0(rt) J_0^n(t) t dt, \qquad 0 \le r \le n$$

$J_0(z)$ being the Bessel function of zeroth order.

Again,

$C_{l,n+1} = C_{l,n} + \cos\theta_{l,n+1}$, $S_{l,n+1} = S_{l,n} + \sin\theta_{l,n+1}$, $C_{l,n} = R_{l,n}\cos\bar{\theta}_{l,n}$, $S_{l,n} = R_{l,n}\sin\bar{\theta}_{l,n}$, so that $R_{l,n} = \sqrt{{C_{l,n}}^2 + {S_{l,n}}^2}$ and $\bar{\theta}_{l,n} = \arctan\left(\frac{S_{l,n}}{C_{l,n}}\right)$.

Therefore, it follows from equation (1) that the predictive density of θ_{n+1} given $\theta_1, \theta_2, \dots, \theta_n$ is

$$g(\theta_{n+1}|\theta_1, \theta_2, \dots, \theta_n) = \frac{1}{(2\pi)^n \psi_{n+1}\left(\sqrt{{R_{l,n}}^2 + 1 + 2R_{l,n}\cos(\theta_{l,n+1} - \bar{\theta}_{l,n})}\right)}$$

$$\propto \frac{1}{\psi_{n+1}\left(\sqrt{{R_{l,n}}^2 + 1 + 2R_{l,n}\cos(\theta_{l,n+1} - \bar{\theta}_{l,n})}\right)}$$

(Rao and Sen Gupta, 2001, pp. 209).

By Rayleigh's approximation for large *n* of the length of the sample resultant length (Lord Rayleigh, 1880), it can be seen that

$$\psi_n(r) \approx \frac{2}{n} exp\left(-\frac{r^2}{n}\right)$$

Then it follows that

$$g(\theta_{n+1}|\theta_1, \theta_2, \dots, \theta_n) \propto \frac{1}{exp\left(\left(-\frac{2R_{l,n}}{n+1}\right)\cos(\theta_{l,n+1} - \bar{\theta}_{l,n})\right)}$$
$$\propto exp\left(\left(\frac{2R_{l,n}}{n+1}\right)\cos(\theta_{l,n+1} - \bar{\theta}_{l,n})\right)$$

which is the p.d.f of a von Mises distribution with center at $\bar{\theta}_{l,n}$ and concentration parameter $\frac{2R_{l,n}}{n+1}$, *i.e.*, $\hat{\mu}_1 = \bar{\theta}_{l,n}$ and $\hat{\kappa}_1 = \frac{2R_{l,n}}{n+1}$.

Thus,

$$g(\theta_{n+1}|\theta_1, \theta_2, \dots, \theta_n) \sim VM\left(\bar{\theta}_{l,n}, \frac{2R_{l,n}}{n+1}\right)$$

We see that the predictive distribution is symmetric and unimodal in nature, the mode being at $\bar{\theta}_{l,n}$. In the following section, we discuss the predictive interval estimation.

3. Predictive Interval Estimation

Let $f(\theta, \mu, \kappa)$ be the predictive density of θ_{n+1} given $\theta_1, \theta_2, \dots, \theta_n$. A $100(1-\alpha)\%$ Predictive Interval for θ is given by $[\theta_L, \theta_U]$ where θ_L and θ_U are such that

$$\int_{\theta_L}^{\theta_U} f(\theta, \mu, \kappa) d\theta = 1 - \alpha \qquad (2)$$

In addition to (2), if the area under the predictive density to the left of θ_L is equal to the area under the predictive density to the right of θ_U, *i.e.* if

$$\int_0^{\theta_L} f(\theta, \mu, \kappa) d\theta = \int_{\theta_U}^{2\pi} f(\theta, \mu, \kappa) d\theta = \frac{\alpha}{2} \qquad (3)$$

then the corresponding predictive interval is termed as $100(1-\alpha)\%$ equal tail area predictive interval.

It can further be seen that the $100(1-\alpha)\%$ equal tail area predictive lower and upper limits, *viz.* θ_L and θ_U are nothing but the $\left(\frac{100\alpha}{2}\right)$th and $\left(100 - \frac{100\alpha}{2}\right)$th percentiles of the predictive distribution respectively, since the density is symmetrical.

4. Predictive Risk Function and Predictive Loss in Predictive Density Estimation of Circular Random Variable

Analogous to the posterior expected loss in the Bayesian parametric inference literature; we have the concept of induced loss in the Predictive inference literature.
Suppose $L(y, a)$ is the loss function associated with predicting the true value $y \in Y$ of a future observation (or set of observations) by $a \in Y$, where Y is the set of future

observation(s). Further assume that $g(y|\tilde{x})$ is the predictive density conditioned on the past and present observations $\tilde{x}$.
Then the expected loss with respect to the predictive density $g(y|\tilde{x})$ is given by

$$L(a) = \int_y L(y,a)g(y|\tilde{x})dy$$

In Bayesian predictive context, Aitchison and Dunsmore (1975) had termed this expected predictive loss as the "Induced Loss" and the value of *a* that minimized the induced loss had been called as the Bayes point predictor of the future observation *y*.
In the Frequentist predictive inference literature, the expected loss w.r.t the predictive density may be termed as the predictive risk function and our aim would consist in minimizing the predictive risk function or equivalently, to find an optimum value of *a* for which the predictive risk function will be the minimum. This predictor may be termed as the frequentist predictive point estimator.

Here, we are dealing with the predictive density estimation of a circular random variable which lies in the range $(0,2\pi)$. So, special loss functions need to be designed which consider the periodicity property of the circular r.v. The loss function is essentially a non-negative function as the loss incurred is positive if the predicted value is different from the true value and zero, otherwise. In other words, the loss function should be an increasing function of the absolute difference between the true value and its predicted value. The circular loss function is hereby considered and the frequentist predictive point estimator of the future observation *y* under this loss function for the *l*-modal Circular Normal predictive density has been worked out.

The circular loss function is defined in the literature as follows:

$$L(y,a) = 1 - \cos(a-y); 0 < a, y < 2\pi$$

The circular loss function can be seen to be a mapping from the set $[0,\pi]$ to $[0,2]$.

The predictive risk function of θ_{n+1} under circular loss function is found to be

$$1 - A(\hat{\kappa}_1)\cos(a - \hat{\mu}_1)$$

Solution of the equation $\frac{d}{da}L(a) = 0$ yields the stationary value of *a* to be

$$a = n\pi + \hat{\mu}_1 \quad , \qquad n = 0,1$$

We further see that the value a_0 of *a* for which $\frac{d^2}{da^2}L(a)|_{a=a_0} > 0$ is attained and hence, becomes the frequentist predictive point estimator of θ_{n+1} given $\theta_1, \theta_2, \dots, \theta_n$ under the circular loss function is found to be

$$n\pi + \hat{\mu}_1 \quad , \qquad n = 0$$

or

$$\hat{\mu}_1 = \bar{\theta}_{l,n} \tag{4}$$

We, thus, see that the frequentist predictive point estimate under the circular loss function coincides with the mean direction of the observations $l\theta_1, l\theta_2, \ldots, l\theta_n$.

5. Properties of the frequentist predictive point estimate under the circular loss function

The frequentist predictive point estimates under the circular loss function is equal to the sample circular mean direction of the observations $l\theta_1, l\theta_2, \ldots, l\theta_n$. It has the following properties:

a) The conditional sampling distribution of $\bar{\theta}_{l,n}$ given the resultant length $R = r_{l,n}$, is given by

$$f\left(\bar{\theta}_{l,n} \,|R = r_{l,n}\right) \sim VM\left(l\mu, \kappa r_{l,n}\right)$$

b) $\bar{\theta}_{l,n}$ is an unbiased estimate of $l\mu$.

The proofs of both these properties are deferred to the Appendix A.

6. Result and Analysis

6.1. *l*-modal circular normal distribution as density of the past data on wind direction

Figure 1 displays the histogram of the daily wind direction data collected from Dibrugarh Meteorological station measured during morning for the Monsoon season (June-September) during the years 2012 and 2013:

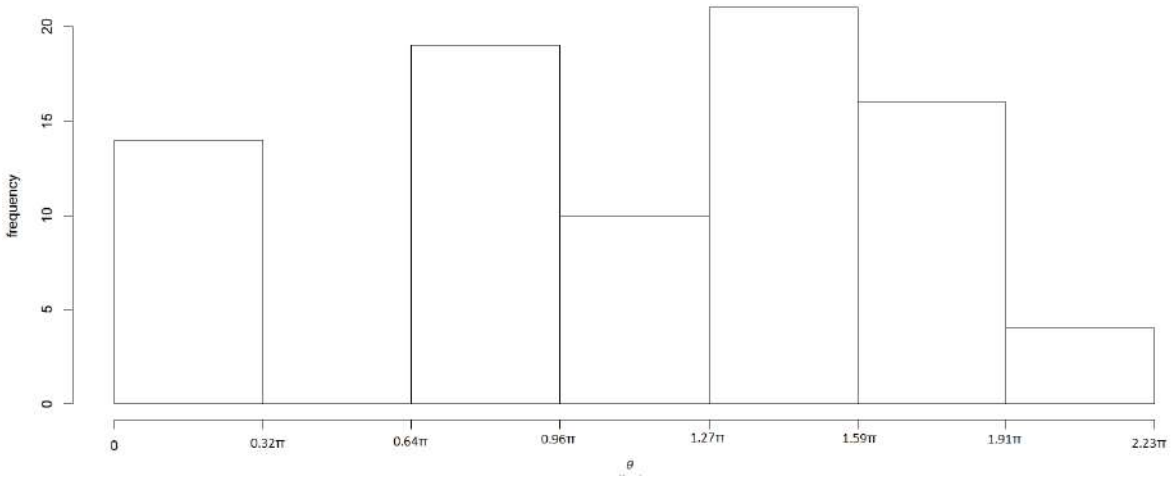


Figure 1: Histogram of the wind direction data collected from Dibrugarh Meteorological station measured during morning for the monsoon season during the years 2012 and 2013

The histogram of the wind direction data under consideration is showing the data to have 3 equidistant modes. The maximum likelihood estimates of the parameters of the *l*-modal Circular Normal distribution are

$\hat{\mu} = 0.1745$ (measured in radians), $\hat{\kappa} = 0.0127$, $\hat{l} = 3$.

The goodness-of-fit test that has been employed here is based on Watson's U^2 test of circular uniformity (Mardia and Jupp, 2000). The critical value of the test statistic at 1% level of significance has been found to be 0.267 whereas the observed value is 0.2397. Thus, the *l*-modal Circular Normal distribution is found to be a good fit to the data on wind direction for Dibrugarh Meteorological station measured during morning for the Monsoon season (June-September) during the years 2012 and 2013. We, therefore, carry out the predictive density estimation of the future observation through sufficiency approach under the *l*-modal Circular Normal model.

6.2. Predictive density estimation

From the data, we have found

$$\bar{\theta}_{l,n} = 0.5235 \text{ and } \frac{2R_{l,n}}{n+1} = 0.0126$$

Thus,

$$g(\theta_{n+1}|\theta_1, \theta_2, \dots, \theta_n) \sim VM(0.5235, 0.0126)$$

i.e., the distribution of θ_{n+1} given $\theta_1, \theta_2, \dots, \theta_n$ is von Mises with parameters 0.5235 and 0.0126.

6.3. Predictive Interval Estimation

A 95% equal tail area predictive interval for θ_{n+1} is given by $[\theta_L, \theta_U]$ where θ_L=2.5th Percentile and θ_U=97.5th Percentile of $VM(0.5235, 0.0126)$ distribution.

Solving (3) for $\alpha = 0.05$ yields

$$\theta_L = 0.157 \text{ and } \theta_U = 6.126$$

Thus, the 95% equal tail area predictive interval for θ_{n+1} is given by $[0.157, 6.126]$. The interpretation of the above statement is "There is 95% chance that the future observation θ_{n+1} would lie between 0.157 and 6.126".

Similarly, a 90% equal tail area predictive interval for θ_{n+1}, is represented by $[\theta_L', \theta_U']$ where θ_L'=5th Percentile and θ_U'=95th Percentile of $VM(0.5235, 0.0126)$ distribution.

Solving (3) for $\alpha = 0.10$ gives

$$\theta_L' = 0.282 \text{ and } \theta_U' = 6.001$$

Thus, the 99% equal tail area predictive interval for θ_{n+1} is given by $[0.282, 6.001]$. This means there is a 99% chance that the future observation θ_{n+1} would lie within the values 0.282 and 6.001.

6.4. Predictive risk function and frequentist predictive point estimator of the future observation under the circular loss function

In this section, we determine the frequentist predictive point estimator of the future observation under the circular loss function.

It can be seen from expression (4) that the frequentist predictive point estimator of the future observation θ_{n+1} under the Circular loss function (measured in radians) is

$$\hat{\theta}_{n+1} = 0.5235$$

7. Discussion

Through this paper, the future density of wind direction prevailing at Dibrugarh Meteorological station located in Assam, measured during morning for the Monsoon season (June-September) during the years 2012 and 2013, conditioned on the past and present wind direction data using the Sufficiency Approach of Predictive Inference under the *l*-modal Circular Normal model has been derived and then the equal tail area predictive interval estimation of the future observation has been carried out. The predictive point estimator of the future observation under circular loss function has been obtained, which has been termed as the frequentist predictive point estimator. Lastly, the properties of the frequentist predictive point estimator have been explored and it has been found that it follows von Mises or Circular Normal distribution.

As a future scope of the present study, the frequentist predictive point estimator under the different circular distributions can be studied assuming several loss functions and compare their relative efficiencies. Having obtained these estimators, one can then attempt to explore the properties of these estimators.

Acknowledgement

The authors are immensely grateful to the Department of Science and Technology, New Delhi, Govt. of India for providing financial assistance to Miss Sahana Bhattacharjee through the Innovation in Science Pursuit for Inspired Research (INSPIRE) programme. They would also like to thank the reviewer(s) and Editor whose constructive suggestions have helped improve the quality of the original manuscript.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press, New York (ISBN: 9780511569647).

Butler, R. (1986). Predictive likelihood inference with applications (with discussion). *Journal of the Royal Statistical Society*, **48**, 1-38.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London (ISBN: 9780412161605).

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss function. *Biometrika*, **85**, 1-11.

Hennig, C. and Kutlukaya, M. (2007). Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal*, **5**(**1**), 19-39.

Jammalamadaka, S. R. and SenGupta, A. (1998). Predictive inference for directional data. *Statistics & Probability Letters.*, **40**, 247-257.

Kent, J. T., Mardia, K. V. and Rao, J. S. (1979). A characterization of uniform distribution on the circle. *The Annals of Statistics*, **7**, 197-209.

Lord Rayleigh, F. R. S. (1880). On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 5th Series*, **10**(**60**), 73-78.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons Ltd, Chichester (ISBN: 9780471953333).

Jammalamadaka, S. R. and Sengupta, A. (2001). *Topics in Circular Statistics*. World Scientific Publishing Co. Pte. Ltd, Singapore (ISBN: 9789812779267).

Ravindran, P. and Ghosh, S. K. (2012). Bayesian Analysis of Circular Data Using Wrapped Distributions. *Journal of Statistical Theory and Practice*, **5**, 547-561.

Sen Gupta, A. and Maitra, R. (1998). On best equivariance and admissibility of simultaneous MLE for mean direction vectors of several Langevin distributions. *Annals of the Institute of Statistical Mathematics*, **50**, 715-727.

APPENDIX A

A.1: If $\theta_1, \theta_2, \dots, \theta_n$ is a random sample from *l*-modal Circular Normal distribution with mean direction μ and concentration parameter κ, the conditional sampling distribution of $\bar{\theta}_{l,n}$, the mean direction of $l\theta_1, l\theta_2, \dots, l\theta_n$ given the resultant length $R = r_{l,n}$, is given by

$$f\left(\bar{\theta}_{l,n} \,|R_{l,n} = r_{l,n}\right) \sim VM\left(l\mu, \kappa r_{l,n}\right)$$

Proof: Given the random sample $\theta_1, \theta_2, \dots, \theta_n$ from *l*-modal Circular Normal distribution with mean direction μ and concentration parameter κ, the joint density of the observations $(\theta_1, \theta_2, \dots, \theta_n)$ is given by

$$f_{\kappa,l}(\theta_1, \theta_2, \dots, \theta_n) = \frac{1}{I_0^n(\kappa)(2\pi)^n} e^{\kappa \sum_{i=1}^n \cos(l\theta_i - l\mu)}$$

$$= \frac{1}{I_0^n(\kappa)(2\pi)^n} e^{\kappa \sum_{i=1}^n (\cos l\theta_i \cos l\mu + \sin l\theta_i \sin l\mu)}$$

$$= \frac{e^{(\kappa \cos l\mu)(\sum_{i=1}^n \cos l\theta_i) + (\kappa \sin l\mu)(\sum_{i=1}^n \sin l\theta_i)}}{I_0^n(\kappa)(2\pi)^n}$$

$$= \left\{\frac{e^{(\kappa \cos l\mu)(\sum_{i=1}^n \cos l\theta_i) + (\kappa \sin l\mu)(\sum_{i=1}^n \sin l\theta_i)}}{I_0^n(\kappa)}\right\} \frac{1}{(2\pi)^n}$$

$$= \left\{\frac{e^{(\kappa \cos l\mu)(\sum_{i=1}^n \cos l\theta_i) + (\kappa \sin l\mu)(\sum_{i=1}^n \sin l\theta_i)}}{I_0^n(\kappa)}\right\} f_0(\theta_1, \theta_2, \dots, \theta_n) \qquad (5)$$

$f_0(\theta_1, \theta_2, \dots, \theta_n)$ being the joint density of a random sample from Circular Uniform distribution, whose concentration parameter $\kappa = 0$.

It was established by Kent *et al*. (1979) that for circular uniform samples, the resultant length $R_{l,n}$ and $\bar{\theta}_{l,n}$ and hence, $(\bar{\theta}_{l,n} - \mu) = \beta$ is independently distributed and $R_{l,n}, \beta$ have the following respective distributions:

$$f_0(r_{l,n}) = r_{l,n}\psi_n(r_{l,n}), \quad f_0(\beta) = \frac{1}{2\pi}$$

Again, we know that the sample mean direction is rotationally equivariant. So, $(\bar{\theta}_{l,n} - l\mu) = \beta'$ (say) is also uniformly distributed in the range $(0,2\pi)$.

Following Rao and Sen Gupta (2001), pp. 71, it can be seen that the pdf of $\cos\beta' = c'$ (say) is

$$f_0(c') = \frac{1}{\pi\sqrt{1 - c'^2}}$$

Consequently, the joint distribution of the resultant length $r_{l,n}$ and c' for a sample from circular uniform distribution is

$$f_0(r_{l,n}, c') = \frac{r_{l,n}\psi_n(r_{l,n})}{\pi\sqrt{1 - c'^2}}$$

The joint density of the resultant length $R_{l,n}$ and the quantity $\cos(\bar{\theta}_{l,n} - l\mu) = c'$ for the *l*-modal Circular Normal (μ, κ) distribution can be obtained from the joint density in expression (5) by integrating over the samples that have given values of $R_{l,n}$ and c'.

Let $\mathcal{A} = \{(\theta_1, \theta_2, \dots, \theta_n): R_{l,n} = \sqrt{(\sum_{i=1}^n \cos\theta_i)^2 + (\sum_{i=1}^n \sin\theta_i)^2} = r_{l,n}, c' = \cos\beta'\}$.

Thus, the joint density of $(r_{l,n}, c')$ for the *l*-modal Circular Normal distribution is given by

$$f_{\kappa,l}(r_{l,n}, c') = \int_{\mathcal{A}} f_{\kappa,l}(\theta_1, \theta_2, \dots, \theta_n)d\theta_1 d\theta_2 \dots d\theta_n$$

$$= \frac{e^{(\kappa\cos l\mu)(\sum_{i=1}^n \cos l\theta_i) + (\kappa\sin l\mu)(\sum_{i=1}^n \sin l\theta_i)}}{I_0^n(\kappa)} \int_{\mathcal{A}} f_0(\theta_1, \theta_2, \dots, \theta_n)d\theta_1 d\theta_2 \dots d\theta_n$$

$$= \frac{e^{(\kappa\cos l\mu)(\sum_{i=1}^n \cos l\theta_i) + (\kappa\sin l\mu)(\sum_{i=1}^n \sin l\theta_i)}}{I_0^n(\kappa)} f_0(r_{l,n}, c')$$

$$= \frac{e^{(\kappa\cos l\mu)(r_{l,n}\cos\bar{\theta}_{l,n}) + (\kappa\sin l\mu)(r_{l,n}\sin\bar{\theta}_{l,n})}}{I_0^n(\kappa)} \frac{r_{l,n}\psi_n(r_{l,n})}{\pi\sqrt{1 - c'^2}}$$

$$= \frac{e^{(\kappa r_{l,n})\cos(\bar{\theta}_{l,n} - l\mu)}}{I_0^n(\kappa)} \frac{r_{l,n}\psi_n(r_{l,n})}{\pi\sqrt{1 - c'^2}}$$

$$= \frac{e^{(\kappa r_{l,n})c'}}{I_0^n(\kappa)} \frac{r_{l,n}\psi_n(r_{l,n})}{\pi\sqrt{1 - c'^2}} \tag{6}$$

Integrating expression (6) over the range $(-1,1)$ of c', we get the marginal density of the resultant length $R_{l,n}$ of a sample from the *l*-modal Circular Normal distribution as

$$f_{\kappa,l}(r_{l,n}) = \int_{-1}^{1} f_{\kappa,l}(r_{l,n}, c')dc'$$

$$= \int_{-1}^{1} \frac{e^{(\kappa r_{l,n})c'}}{I_0^n(\kappa)} \frac{r_{l,n}\psi_n(r_{l,n})}{\pi\sqrt{1-c'^2}} dc'$$

$$= \frac{r_{l,n}\psi_n(r_{l,n})}{I_0^n(\kappa)} \int_{-1}^{1} \frac{e^{(\kappa r_{l,n})c'}}{\pi\sqrt{1-c'^2}} dc'$$

$$= \frac{r_{l,n}\psi_n(r_{l,n})}{I_0^n(\kappa)} \int_{0}^{2\pi} \frac{e^{(\kappa r_{l,n})\cos\beta'}}{2\pi}\, d\beta'$$

$$= \frac{I_0(\kappa r_{l,n})}{I_0^n(\kappa)} r_{l,n}\psi_n(r_{l,n}) \tag{7}$$

Again, as $R_{l,n}$ and $\bar{\theta}_{l,n}$ are independently distributed for circular uniform samples, their joint distribution is

$$f_0(r_{l,n}, \bar{\theta}_{l,n}) = r_{l,n}\psi_n(r_{l,n})\frac{1}{2\pi}$$

Suppose $\mathcal{B} = \left\{(\theta_1, \theta_2, \dots, \theta_n): R_{l,n} = \sqrt{(\sum_{i=1}^n \cos\theta_i)^2 + (\sum_{i=1}^n \sin\theta_i)^2} = r_{l,n}, \bar{\theta}_{l,n} = arctan\frac{\sum_{i=1}^n \sin l\theta_i}{\sum_{i=1}^n \cos l\theta_i}\right\}$

The joint density of $(r_{l,n}, \bar{\theta}_{l,n})$ for *l*-modal Circular Normal distribution is

$$f_{\kappa,l}(r_{l,n}, \bar{\theta}_{l,n}) = \int_{\mathcal{B}} f_{\kappa,l}(\theta_1, \theta_2, \dots, \theta_n)d\theta_1 d\theta_2 \dots d\theta_n$$

$$= \frac{e^{(\kappa\cos l\mu)(r_{l,n}\cos\bar{\theta}_{l,n}) + (\kappa\sin l\mu)(r_{l,n}\sin\bar{\theta}_{l,n})}}{I_0^n(\kappa)} f_0(r_{l,n}, \bar{\theta}_{l,n})$$

$$= \frac{e^{(\kappa r_{l,n})\cos(\bar{\theta}_{l,n} - l\mu)}}{I_0^n(\kappa)} r_{l,n}\psi_n(r_{l,n})\frac{1}{2\pi} \tag{8}$$

Therefore, the conditional distribution of $\bar{\theta}_{l,n}$ given $R_{l,n} = r_{l,n}$ is obtained as follows:

$$f(\bar{\theta}_{l,n} | R_{l,n} = r_{l,n}) = \frac{f_{\kappa,l}(r_{l,n}, \bar{\theta}_{l,n})}{f_{\kappa,l}(r_{l,n})}$$

$$= \frac{\frac{e^{(\kappa r_{l,n})\cos(\bar{\theta}_{l,n}-l\mu)}}{I_0^n(\kappa)} r_{l,n}\psi_n(r_{l,n})\frac{1}{2\pi}}{\frac{I_0(\kappa r_{l,n})}{I_0^n(\kappa)} r_{l,n}\psi_n(r_{l,n})}$$

$$= \frac{e^{(\kappa r_{l,n})\cos(\bar{\theta}_{l,n}-l\mu)}}{2\pi I_0(\kappa r_{l,n})}$$

which is the pdf of Von Mises distribution with mean direction $l\mu$ and concentration parameter $\kappa r_{l,n}$.

A.2: The mean direction $\bar{\theta}_{l,n}$ of $l\theta_1, l\theta_2, \dots, l\theta_n$, where $\theta_1, \theta_2, \dots, \theta_n$ is a sample from the *l*-modal Circular Normal distribution with mean direction μ and concentration parameter κ is an unbiased estimator of $l\mu$.

Proof: In the context of circular statistics, an estimate *t* taking values on the unit circle is said to be unbiased for a parameter α of a circular probability distribution (Mardia and Jupp, 2000, pp.83) if

$$\frac{E(\cos t, \sin t)}{\| E(\cos t, \sin t) \|} = (\cos\alpha, \sin\alpha)$$

It follows from proof (1) that the for samples from *l*-modal Circular Normal distribution,

$$\bar{\theta}_{l,n} \sim VM(l\mu, \kappa r_{l,n})$$

Thus,

$$E(\cos\bar{\theta}_{l,n}, \sin\bar{\theta}_{l,n}) = \left(E(\cos\bar{\theta}_{l,n}), E(\sin\bar{\theta}_{l,n})\right)$$

$$= \left(\frac{I_1(\kappa r_{l,n})}{I_0(\kappa r_{l,n})}\cos l\mu, \frac{I_1(\kappa r_{l,n})}{I_0(\kappa r_{l,n})}\sin l\mu\right)$$

$$= \left(A(\kappa r_{l,n})\cos l\mu, A(\kappa r_{l,n})\sin l\mu\right)$$

and

$$\| E(\cos\bar{\theta}_{l,n}, \sin\bar{\theta}_{l,n}) \| = \sqrt{\left(A(\kappa r_{l,n})\cos l\mu\right)^2 + \left(A(\kappa r_{l,n})\sin l\mu\right)^2}$$

$$= \sqrt{\left(A(\kappa r_{l,n})\right)^2}$$

$$= A(\kappa r_{l,n})$$

Finally,

$$\frac{E\left(\cos\bar{\theta}_{l,n}, \sin\bar{\theta}_{l,n}\right)}{\| E\left(\cos\bar{\theta}_{l,n}, \sin\bar{\theta}_{l,n}\right) \|} = \frac{\left(A\left(\kappa r_{l,n}\right)\cos l\mu, A\left(\kappa r_{l,n}\right)\sin l\mu\right)}{A\left(\kappa r_{l,n}\right)}$$

$$= (\cos l\mu, \sin l\mu)$$

or $\bar{\theta}_{l,n}$ is an unbiased estimator of $l\mu$

APPENDIX B

Table B.1: Data set on daily wind direction for Dibrugarh Meteorological station located in Assam, measured (in degrees) during morning for the Monsoon season (June-September) during the years 2012 and 2013

Wind direction measured in degrees			
50	230	150	230
320	230	150	140
150	20	270	320
360	50	230	270
320	180	180	150
140	150	50	50
150	320	180	50
230	320	150	320
20	270	320	320
210	180	270	230
20	270	50	360
150	210	140	20
320	230	20	270
230	150	320	360
150	50	150	150
360	320	320	180
210	270	50	230
210	150	320	320
150	20	150	230
210	150	230	
230	320	230	

Source: Regional Meteorological Center, Guwahati, Assam

For the remaining days, no wind flow was detected and so, the measure of wind direction corresponding to those days were reported as NIL (and are, hence, excluded from the data set).

Publisher
Society of Statistics, Computer and Applications
B-133, Ground Floor, C.R. Park, New Delhi - 110019
Tele: 011 - 40517662
https://ssca.org.in/
statapp1999@gmail.com
2021

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203