ISBN No.978-81-950383-2-9

# Special Proceedings (25)
# [Based upon the 25th Silver Jubilee (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2023 held at Department of Statistics, University of Jammu, Jammu & Kashmir, India during

# February 15 - 17, 2023]

Society of Statistics, Computer and Applications
https://ssca.org.in/
2023

# Society of Statistics, Computer and Applications

## Council and Office Bearers

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

ISBN No.978-81-950383-2-9

# Special Proceedings (25)
# [Based upon the 25th Silver Jubilee (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2023 held at Department of Statistics, University of Jammu, Jammu & Kashmir, India during February 15 - 17, 2023]

**Editors**

V.K. Gupta
Baidya Nath Mandal
R. Vishnu Vardhan
Ranjit Kumar Paul
Rajender Parsad
Dipak Roy Choudhury

Date of Publication: 15 October 2023

Published By:
Institutional Publisher: Society of Statistics,
Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2023

**Printed by: Galaxy Studio & Graphics**
A-181, Weavers Colony Ashok Vihar, Phase-IV, New Delhi-110052
Mob: +91 98183 52203, +91 95829 41203
Email: galaxystudio08@gmail.com

# CONTENTS

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023

Preface i-ii

1. On the History of Certain Early Developments in Sampling Theory 1-10
*T.J. Rao*

2. Global Hunger Index (GHI) Reminds the Phrase "Lies, Damned Lies and Statistics" 11-19
*Padam Singh*

3. How Gainfully Can the Additional Units Be Used? 21-30
*Opendra Salam Singh, Gurumayum Sandweep Sharma and Bikas K. Sinha*

4. A Nonparametric Structure Based on Hierarchical Dirichlet Processes for Studying Gene-Gene and Gene-Environment Interactions 31-41
*Durba Bhattacharya and Sourabh Bhattacharya*

5. A comparison of CAN and UMVU Estimators in Inliers-Prone Distributions 43-50
*K. Muralidharan*

6. Overcoming Challenges Associated with Early Bayesian State Estimation of Planted Acres in the United States 51-78
*Balgobin Nandram*

7. A New Development of Threshold Selection in Extreme Value Theory 79-92
*K.M. Sakthivel and V. Nandhini*

8. Statistical Inference for Frechet Distribution Based on Dual Generalized Order Statistics 93-106
*Bhagwati Devi, Qazi J. Azhad, Ankita Sharma and Ayush Tripathi*

9. Long Memory in Volatility: Application of Fractionally Integrated GARCH Model 107-118
*Debopam Rakshit and Ranjit Kumar Paul*

10. Assessment of Child Mortality and its Socioeconomic and Demographic Determinants- Evidences from the Latest National Family Health Survey of India 119-129
*Anuradha Rajkonwar Chetiya and Vishal Deo*

| | | |
|---|---|---|
| 11. | Statistical and Artificial Neural Network Approaches for the Classification of Rice Genotypes based on Morphological Characters<br>*Shavi Gupta, Manish Sharma, S.E.H. Rizvi, R.K. Salgotra and S.K. Gupta* | 131-137 |
| 12. | On Zero-inflated Generalized Alternative Hyper-Poisson Distribution and its Properties<br>*C. Satheesh Kumar and Rakhi Ramachandran* | 139-150 |
| 13. | First collision time of Three independent random walks<br>*Arighna Dey, Kumarjit Saha and Anish Sarkar* | 151-168 |
| 14. | Analysis of Large Medical Databases: Addressing the Clinical Relevance of Statistically Significant Findings<br>*Rachana Lele, Anand Seth, Sameer Patel, Jianmin Pan, Marepalli B. Rao, and Shesh N. Rai* | 169-203 |
| 15. | Investigating Regional Income-Based Shared Frailty among Women with Breast Cancer<br>*Vishal Deo, S. Pyne, Meghana Ray and M. Vishnu Vardhan Rao* | 205-215 |

# PREFACE

The Society of Statistics, Computer and Applications (SSCA) was established in 1998 with the aim of creating a platform for advancing and disseminating research in statistics while incorporating information technology. It sought to engage both theoretical and applied statisticians with a strong interest in the innovative applications of statistics across various fields, including agriculture, biological sciences, medical sciences, financial statistics, and industrial statistics. Over the years, the Society has undertaken numerous activities and played a pivotal role in promoting the development of theoretical and applied research in statistical sciences.

One of SSCA's primary initiatives has been the organization of annual national and international conferences across the country. Additionally, SSCA publishes an open-access journal called "Statistics and Applications," which is available on the Society's website (www.ssca.org.in). The journal offers free access to full-length papers, which can be viewed and downloaded at no cost. In addition to regular volumes, the journal also brings out special volumes focusing on emerging thematic areas of global or national significance.

The twenty-fifth Annual Conference of SSCA took place from 15th to 17th February 2023, at the Department of Statistics, University of Jammu, Jammu. The conference was themed "Significance of Statistical Sciences in Emerging Scenario (SSSES 2023)" and served as a platform for renowned international and national statisticians to present significant research findings. The conference featured several noteworthy events, including a pre-conference workshop and various technical sessions. These sessions encompassed the M.N. Das Memorial Lecture and a dedicated session on Financial Statistics, where distinguished statisticians and leading practitioners in the field shared their insights on various finance-related topics. Moreover, the conference included three endowment lectures: the B.K. Kale Memorial Endowment Lecture, J.K. Ghosh Memorial Endowment Lecture, and Bikas Kumar Sinha Endowment Lecture. These lectures were delivered by speakers closely associated with, collaborated with, or were students of the respective honourees. Additionally, there was the V.K. Gupta Endowment Award Lecture for Achievements in Statistical Sciences and Practice, which was presented by Bikas Kumar Sinha. The special proceedings of these sessions have been assigned the ISBN #: 978-81-950383-2-9.

The Executive Council of SSCA made the decision to publish the "Special Proceedings" of the conference, which would feature selected talks, including those presented in the Financial Statistics session. The Guest Editors, nominated by the Society's Executive Council - V.K. Gupta, Baidya Nath Mandal, R. Vishnu Vardhan, Ranjit Kumar Paul, Rajender Parsad, and Dipak Roy Choudhury - were responsible for curating these special proceedings. They invited authors based on their presentations during the conference to submit full papers for consideration in the Special Proceedings.

Distinguished speakers who were shortlisted for contributing to the special proceedings were invited to submit their research papers. After a rigorous review process, 15 research papers were accepted for publication and included in the special proceedings. We owe special debt to all the reviewers for their significant contribution in completing the review process within a short time frame. Our heartfelt appreciation goes to all the members and office bearers of SSCA's Executive Council for their support. We are also grateful to Ms. Jyoti Gangwani for meticulously formatting the papers. Furthermore, we warrant our deepest thanks to Prof. Umesh Rai, Vice Chancellor, University of Jammu, and his dedicated team, particularly Dr. Rahul Gupta, Dr. Parmil Kumar, Dr. Pawan Kumar, Dr. V K Shivgotra, and Dr. Sunil Kumar. Each of them, in his own way deserves major credit for the efforts in organizing this academically enriching conference of SSCA and making their facilities available for the conference.

We expect the readers to benefit from the contents of the special proceedings.

**Guest Editors**

*V.K. Gupta*
*Baidya Nath Mandal*
*R. Vishnu Vardhan*
*Ranjit Kumar Paul*
*Rajender Parsad*
*Dipak Roy Choudhury*

New Delhi
September 2023

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 1–10

# On the History of Certain Early Developments in Sampling Theory

**T. J. Rao**
*Retired Professor, Indian Statistical Institute, Calcutta*

Received: 30 May 2023; Revised: 06 June 2023; Accepted: 08 June 2023

## Abstract

In this paper based on the Bikas Sinha Endowment Lecture, we shall first discuss Rao-Blackwellization of some early estimators obtained in finite population sampling theory along with historical aspects. We note that there are certain lapses with respect to priorities and credits in the literature. Next, we shall briefly sketch the role played by Bikas Sinha related to applications during early days.

*Key words:* Rao-Blackwellization; Probability proportional to size sampling; Call backs; Randomized response technique; Environmental statistics.

**AMS Subject Classification**: 62D05

## 1. Introduction

In this paper, we begin with the technique of *Rao-Blackwellization* in finite population sampling theory, a subject which both Prof. Bikas Sinha (BKS) and I are interested in. Rao-Blackwellization, a term credited to C. R. Rao based on his 1945 'breakthrough' paper published in the *Bulletin of the Calcutta Mathematical Society*, provided improved estimators in conventional as well as adaptive, link-tracing, size-biased sampling techniques. Furthermore, Rao-Blackwellization found applications in statistics and a host of other disciplines including sports, namely Rao-Blackwellized Field Goal percentage estimator (RB-FG%) and possibly social networks such as WhatsApp (RB-WA).

We shall first discuss applications related to improving of estimators in finite population sampling theory relating to Probability Proportional to Size Without Replacement (PPSWOR) selection. We note that proper credit is not given to certain publications and point out certain lapses with respect to (wrt) priorities and credits in the sampling literature.

## 2. Selection with ppswor scheme

Following Basu (1958), wherein he showed that the 'order statistic' (sample units in ascending order of their labels) is a sufficient statistic, Pathak(1961) while discussing sampling from finite populations, noticed that *'any estimator which is not a function of the order statistic'*, can be uniformly improved upon by the use of Rao - Blackwellization technique. BKS along with Sen (Sinha and Sen,1989) goes beyond variance comparisons and generalizes to convex loss functions. In his book on *Finite Population Sampling* with Hedayat (1991), BKS devotes the maximum number of 58 pages for the chapter on PPS sampling. A large part

Corresponding Author: T.J. Rao
Email: tjrao7@gmail.com

*This paper is based on the Bikas Kumar Sinha Endowment Lecture "My Significant Interaction with Bikas Sinha" delivered by the author on 15 February, 2023 during the conference.*

of his work on sampling (solo and with co-authors) was on PPS sampling among others. We shall discuss the case of sampling of 2 units from a finite population of size *N* with the study variate *Y* taking values $Y_i$ and known auxiliary variate *X* related to *Y*, taking values $X_i$ on the units $U_i$, *i = 1,2,....,N.* Let $T_Y$ and $T_X$ denote the population totals of *Y* and *X* respectively.

Let $P_i = X_i / T_X$.

## 2.1. A recap of unbiased estimators of $T_Y$

International Statistical Institute held its biennial session in Delhi in 1951 from 5-11 December. A short session was held in Calcutta from 16 to 18 December along with other international societies. A. C. Das of Indian Statistical Institute presented a paper on successive sampling. As a passing note in this paper, Das (1951) discussed PPSWOR scheme as well at the end.

Thus, if $(i, j)$ are the labels of units selected by PPSWOR in that order, then Das's (1951) estimators for the first and second draws of sample selection are, respectively,

$$t_{1\ Da} = {}^{y_i}\!\big/_{p_i} \qquad (1)$$

and

$$t_{2\ Da} = \frac{1}{(N-1)p_i}\frac{y_j}{p_j}(1-p_i) \quad . \qquad (2)$$

After a gap of 5 years, Des Raj (1956) obtained ordered estimators (for *n* = 2):

$t_1 = y_i / p_i$ for first draw,
$t_2 = y_i + [y_j/\{p_j/(1-p_i)\}]$, based on second draw

and

$$\bar{t} = \{y_i\,(1+p_i)/p_i\} + \{y_j\,(1-p_i)/p_j\}/2 \text{ based on order } (i, j). \qquad (3)$$

$t_i's$ are defined similarly, *i* = 1,2,…,*n* and

$$t_n = y_1 + y_2 + \cdots + y_{n-1} + [y_n/(p_n/(1\text{-}p_1 - p_2 - \ldots. - p_{n-1})]\text{ or equivalently,}$$

$$t_n = y_1 + y_2 +..+ y_n + [y_n/\{x_n/(T_X - x_1 - x_2 - \ldots - x_{n-1} - x_n)\}].$$

By independence of estimators, it is easily seen that

$$\hat{V}(\bar{t}) = \textstyle\sum_1^n (t_i - \bar{t})^2 / n(n-1)$$

is a non-negative unbiased estimator of variance.

## 2.2. Other negative variance estimators

Towards the end of the paper, Das gave an unbiased estimate of variance as well. This estimator received criticism since it *can take negative values.* Horvitz and Thompson (HT, 1952) gave a general homogeneous linear unbiased estimator for $T_Y$, which had *nice* properties. They also gave an unbiased estimator of variance of their estimator, but it *also takes negative values*. A year earlier, Narain (1951), *independently* obtained the same estimator and published in the *Journal of Indian Society of Agricultural Statistics (JISAS)* but was not mentioned by several authors. Thus, credit goes to Narain and HT, and J. N. K. Rao (1999, 2005) rightly called it as NHT estimator. In a discussion of Rao's 1999 paper, J.K. Ghosh observed that it was "*renamed NHT honouring another pioneer Narain".*

The very next year, Yates and Grundy (YG) in 1953 published an alternative variance estimator which *also takes negative values* (less often than HT's). It is interesting to note that Sen (1953) also published the same estimator as YG's in *JISAS.* It is now termed as SYG estimator, thus crediting all the three Sen, Yates and Grundy by researchers and teachers. However, a careful reading of YG's paper points out that, *unaware of* HT*,* Yates *also obtained the* HT estimator and *later* Grundy(G) *joined to give the alternative variance estimator.* Perhaps one should rename NHT estimator as NYHT estimator and SYG variance estimator as SG estimator!!

In view of the above discussion, we note that other variance estimators *also take negative values* and Das's estimator is much criticised. We note here that while for Des Raj's estimators, we need the previous *Y* values to be added to obtain the estimate at a particular draw, for Das's estimator at a particular draw, one need not know the *Y* values of the previous draws. This property comes in handy when one or more *Y* values of the previous draws are unavailable due to non-response, non-cooperation, 'not at home's *etc*. In such a situation one has to depend on Das's only and Des Raj's estimator is of no help.

## 2.3. Basu's concept of 'Face Validity'

More formally, this property can be stated as follows:

'*for estimating population total based on an ordered estimator, it is sufficient to have the Y value at the draw of selection only and the Y-values based on previous draws are not necessary.'*

Borrowing a phrase from Basu (1971), who defines the property of 'face validity', we term the above property as 'order validity'.

Basu (1971) looks at the population total as

$$T_Y = S + S^*,$$

where $S$ is the observed total of $Y$'s and $S^*$ is the unobserved total.

Having observed the $Y$-values and knowing $S$, it is now required to estimate $S^*$.

Now, suppose that the $n$ observed values $Y_i / X_i$ are nearly equal, but $y_n / x_n$ is the largest.

For this situation Desraj's estimator is

$$t_n = y_1 + y_2 + \ldots\ldots + y_{n-1} + ( y_n / x_n ) (T_X - x_1 - x_2 + \ldots\ldots + x_{n-1} )$$

equivalently,

$$t_n = y_1 + y_2 + .. + y_n + [y_n / \{x_n / (T_X - x_1 - x_2 - .. - x_{n-1} - x_n )\}].$$

Basu questions estimating $S^*$ by only one $Y$ value, namely

$$y_n / \{ x_n / (T_X - x_1 - x_2 - .. - x_{n-1} - x_n )\}].$$

Hence, Basu claims that it is not unbiasedness, but is hard to define property of 'Face Validity' of an estimate. He claims

$$t_n = \sum_1^n y_i + \{(\sum_1^n y_i / \sum_1^n p_i) \} (1 - \sum_1^n p_i),$$

which uses all $n$ $Y$-values has a greater face validity.

Note that $t_n$ is nothing but the familiar ratio estimator

$$\hat{Y}_R = \sum_1^n y_i / \sum_1^n x_i )T_X.$$

Following Basu's arguments, one could suggest a concept like face validity as:

*'An estimator is said to be 'order-valid' if 'the estimator based on the result of a particular draw does not depend on the Y-values of the previous draws.'*

However, this estimator may be inefficient, but in the presence of missing values due to non-response, not-at-home's *etc*., such an estimator may be relevant.

## 2.4. Lahiri-Murthy unordered estimator

Murthy (1957) concentrated on Des Raj's ordered estimators and discussed how to obtain an unordered (symmetrized) estimator. In a short section of this paper, titled 'unordering of Das's estimators', he briefly mentions Das's estimator, but unorders for another sampling scheme, and not for PPSWOR under consideration. He did not treat Das's estimator the way he did for Des Raj's as described below:

Recalling that for the ordered sample $(i, j)$ the estimator is (3), namely

$\bar{t}_{ij} = [\{y_i ( 1 + p_i ) / p_i\} + \{ y_j (1- p_i) / p_j \}]/2$ based on order $(i, j)$ and

$\bar{t}_{ji} = [\{y_j ( 1 + p_j ) / p_j\} + \{ y_i (1 - p_j) / p_i \}]/2$ based on order $(j, i)$,

Murthy obtained an estimator combining these two by the respective probabilities of the sample as weights, namely $p_i\, p_j / (1- p_i)$ and $p_j\, p_i / (1- p_j)$, which gave the Unordered (symmetric) Des Raj Estimator:

$$\bar{t}_M = [\{(1 - p_j) (y_i / p_i)\} + \{(1 - p_i) (y_j / p_j)\}] / (2 - p_i - p_j),$$

which is Murthy's (3.17) of his 1957 paper.

A point to be noted here is that Halmos (1946) also mentions symmetrized unbiased estimators. For the last 70 years this is referred to as Murthy's (1957) unordered estimator. In a footnote of his 1957 paper (p. 384), Murthy mentions:

"*Lahiri conjectured that Desraj's estimators can be improved by weighting the different ordered estimators by their respective probabilities and in fact suggested the estimator given by (3.17)*".

So, it may be called Lahiri-Murthy unordered estimator:

$$\bar{t}_{LM} = \quad [\{(1-p_j)\,(y_i\,/\,p_i)\} + \{(1-p_i)\,(y_j\,/\,p_j)\}]\,/\,(2-p_i-p_j)$$

giving credit to Lahiri as well.

Symmetrizing Das's in the same way, we get an interesting symmetrized

$$\hat{Y}_{symm.Das} = (\hat{Y}_{Symm.Desraj} + \hat{Y}_{Midzuno.Lahiri})/2,$$

where $$\hat{Y}_{Midzuno.Lahiri} = \sum_1^n y_i \Big/ \sum_1^n p_i$$

and $\hat{Y}$ is an estimator of $T_Y$. The readers may like to see Rao(2021b) for details.

### 2.5. Further unorderings

For obtaining nonnegative SYG (or SG) variance estimators, Brewer (1963) and Durbin (1967) gave simple $\pi ps$ sampling selection procedures based on ordered samples of size 2. Brewer's method consists of selecting the first unit with probability proportional to $p_i(1-p_i)/(1-2p_i)$ and second unit with probability $p_j/(1-p_i)$, $j \neq i$. This gives $\pi_i = 2p_i$ and SYG variance estimator non-negative.

For the same purpose, Durbin's method selects the first unit with probability $p_i$ and the second unit with probability proportional to $p_j[\{1/(1-2p_i)\} + \{1/(1-2p_j)\}], j \neq i$.

We now observe that the selection of units here is based on an order and we unorder these following the above methodology {see Rao (2021b)}. Thus, wherever order is involved in selection of sample, the estimators can be unordered using proper methodology by Rao-Blackwellization.

## 3. Nonresponse

So far, we have discussed situations that involved reduction of sampling errors. We shall now move on to the case of non-sampling errors of which non-response due to 'not at homes' and refusal to answer sensitive questions are major contributors. For the first category of 'not at homes', the technique of 'call-backs', while for the other one, 'randomized response technique (RRT) were proposed.

### 3.1. To call back or not to call back

For this, what is known as Politz-Simmons technique (PST) in the literature, is used to estimate parameters using data on first call itself, thus avoiding 'call-backs,' by asking respondents during the interview a question about their *availability at home (or, otherwise) at the same time during the preceding five week nights.*

However, during the discussion of the paper read by Yates(1946) at the Royal Statistical Society (RSS) Meeting, Hartley(1946) proposed an 'ingenious' and 'decidedly cheaper' alternative to call-backs . Hartley mentions: "*Details of this scheme were given to the War-time Social Survey, but I understand that, owing to pressure of work, an opportunity of trying has, as yet, not arisen".* Soon after, Politz and Simmons (PS,1949) published their work popularly known as Politz-Simmons technique in the *Journal of American Statistical Association* which is on similar lines to the proposed method of Hartley. PS (1949) while acknowledging the work of Hartley, say: *"It has recently been brought to the authors' attention that a somewhat similar plan was proposed independently by H.O. Hartley before the Royal Statistical Society...."*

In the present day context of ever-changing and emerging socioeconomic scenario of the society, it is to be noted that this question itself has become highly sensitive for the respondents who thereby may evade to answer this question truthfully. Rao *et al.* (2016) have applied Warner's (RRT) in this situation and developed a nontrivial randomized response Hartley, Politz, Simmons(HPS) technique.

`

## 4. Role of BKS in other early contributions

Hailing from, the then, East Bengal, environmentally rich and ecologically diverse background, it is but natural that BKS turned his attention to 'Environment' and other specialised areas of Statistics (see Rao, 2021a). BKS was appointed as 'Expert on Mission' for United Nations (UN) Statistics Training Programme in 1991based on his early contributions and this led to his serving as a consultant to the United States Environmental Protection Agency (USEPA) in 1993.

At home, he was also appointed a Member of the apex body, National Statistical Commission, Government of India (2006-2009). His expertise involved in social and environment statistics. Other early contributions of BKS include 'Official Statistics in neighbouring developing countries in the Indian sub-continent' (Rao and Sinha (2011). Collaborating with his colleagues JK Ghosh *et al.* (1999), a detailed account of 'Evolution of Statistics in India' was presented. Faculty and research scholars of Sociology and other applied statistics unis of ISI took BKS's and the author's help in organising their surveys rigorously. As an example we cite the design and implementation of an innovative survey of Annual Book Exhibition held in Calcutta Maidan, popularly known as '*boi mela'* wherein random time points are chosen.

## 5. Rao-Blackwellized WhatsApp

In the earlier sections, we have discussed the application of Rao-Blackwellization for improving the estimators in sampling theory. We have also mentioned its application in sports to obtain improved estimates of Field Goal Percentages (RB-FG%) in Basketball by Daly Grafstein and Bornn (2018). Their interesting analysis could be applied to ratings of sports persons in tennis, cricket and a host of others as well.

The new concept we proposed deals with 'message clustering' and 'smart response utility' while using WhatsApp (WA). Every day, we are flooded with WA messages on our smart phones. Not all users of WA have time to go through all the messages and take suitable action.

The new concept is based on an 'APP' to be constructed which compresses the data and disregards repetitions. An abridged message which is 'sufficient' is composed. For example, 'Good Morning', 'Have a Good day', Happy xxxx (day of the week), *etc*. can be treated as observations repeated with replacement. The App so constructed will have an AI/ML mechanism that recognises the equivalents and exhibits just one or two short lines editing meaningfully and then lists all the users that sent these particular messages, thus solving the problem of 'message clustering'. Now, an individual can quickly choose from the list, to whom the abridged (meaningful) reply can be sent (ignoring some senders) or a single 'Thank you all', if appropriate, thus enabling 'smart response utility'.

In view of the compression and reduction of data and the ability to present 'sufficient' information, we called it the Rao-Blackwellized WhatsApp (RW-WA). In a strict sense, this concept is not like the research of Daly-Grafstein and Bornn (2018). The new App so constructed reduces redundancy, saves time and effort and could even be made premium.

## 6. Certain lapses in literature and credits

In Sections 2 and 3, we have already mentioned about the credits that were missed out in sampling literature. We shall add here a few more (though not complete) with respect to the early results. The following anecdote may be of interest to the readers who are unaware of the history of the term 'Rao-Blackwellization':

C. R. Rao (1945) established this result and published in the *Bulletin of Calcutta Mathematical Society.*

A couple of years later, Blackwell (1947) obtained the same result in *Annals of Mathematical Statistics.*

Five years later, Scheffe' and Lehmann called it Rao-Blackwell Theorem.

In a 1953 conference, when Berkson named it Blackwellization. C. R. Rao pointed out that he published it in 1945 itself. Berkson replied "Raoization is difficult to say," but later termed it Rao- Blackwellization.

D.V. Lindley, in a book review referred to Blackwell only. When C.R. Rao wrote to him, he replied saying "you have not mentioned it in the introduction of the paper… C. R. Rao replied saying he is unaware that "introduction is written for the benefit of those who only read introduction and not the paper."

In the seminal paper read at the Royal Statistical Society meeting, Neyman derived the optimum allocation of sample size to strata in 1934. It was pointed out to him by Donavan Thompson that Tschprow had already established this result in 1923 Metron paper. Neyman recognized the priority and gave credit to Tchuprow. Thus, one may term this allocation as Tchuprow- Neyman allocation.

Hansen and Hurwitz introduced PPS sampling in their 1942 AMS paper. Mahalanobis in his 1937 paper discusses cumulative totals method for selection with varying probabilities.

Madow and Madow in 1944 discuss systematic sample, while for the selection of sample Anthropometric survey of United Provinces, Mahalanobis, Majumdar and C. R. Rao (1941) used a systematic sample. In the introduction, Mahalanobis points out that for detailed subclassifications, the ultimate sample size would be small giving large errors, a concept echoed in small area estimation.

Later while analysing Bengal Anthropometric data, C. R, Rao recognises that standard tests need to be applied cautiously since the data is based on multi-stage stratification heralding 'Analysis of Complex Surveys'.

Olkin's 1958 Biometrika paper on Multivariate regression estimators was envisaged by B. Ghosh in 1947 in Bulletin of Calcutta Statistical Association.

Murthy in 1964 rediscovers product method of estimation which was attempted using polykays by Robson in 1957 itself.

What we call as Midzuno-Sen (1952) sampling scheme is attributed to Midzuno's student Ikeda (1951), Haj'ek(1949) and Lahiri (1951), now popularly referred to as Lahiri-Midzuno-Sen (LMS) scheme.

Royall's 1970 predictive approach of Biometrika is also attributed to Brewer (1963) for introducing this concept.

It is not clear how one does not find a reference to Kumarappa's 1931 detailed survey of Matar taluka of Gujarat on the advice of Gandhi, which is a medium sized multi subject survey submitted for the attention of the British Raj, while discussing Mahalanobis's surveys of NSS (1950 onwards).

(For details and full references, please see T. J. Rao (2016), *On the History of Certain Early Key Concepts in Sampling Theory and Practice*).

## Acknowledgements

The author wishes to thank Prof. Bimal Sinha, Prof. Vinod Gupta and the organisers of the Conference for inviting him to deliver the Bikas Sinha Endowment Lecture. On this occasion we all join in wishing Prof. Bikas Sinha Happy and Healthy years ahead, academically and personally.

## References

Basu, D. (1958). On sampling with and without replacement. *Sankhya*, **20**, 287-294.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In: Godambe, V. P. and Sprott, D. A. (Eds.). *Foundations of Statistical Inference*, Holt, Rinehart and Winston, 203-242.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, **18**, 105-110.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, **10**, 213-233.

Daly-Grafstein, D. and Bornn, L. (2019). Rao-Blackwellizing field goal percentage. *Journal of Quantitative Analysis of Sports,* **15**, 85–95.

Das, A. C. (1951). On two phase sampling and sampling with varying probabilities. *Bulletin of International Statistical Institute,* **33**, 105-112.

Des Raj (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association,* **51**, 269-284.

Durbin, J. (1967). Design of multistage surveys for estimation of sampling error. *Applied Statistics,* **16**, 152-164.

Ghosh, J. K., Maiti, P., Rao, T. J., and Sinha, B. K. (1999). Evolution of statistics in India. *International Statistical Review,* **61**, 13-34.

Halmos, P. R. (1946). The theory of unbiased estimation. *Annals of Mathematical Statistics,* **17**, 34-43.

Hartley, H. O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society,* **109**, 37.

Hedayat, A. S. and Sinha, B. K. (1991). *Finite Population Sampling,* Wiley, New York.

Horvitz, D. G. and Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* **47**, 663-685.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya,* **18**, 379-390.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169-174.

Pathak, P. K. (1961). Use of 'Order-Statistic' in sampling without replacement. *Sankhya, A,* **23**, 409-414.

Politz, A. N. and Simmons, W. R. (1949). An attempt to get the "not at homes" into the sample without call backs. *Journal of the American Statistical Association,* **44**, 9-31.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, **37**, 81-91.

Rao, J. N. K. (1999). Some current trends in sample surveys theory and methods. *Sankhya, B*, **61,** 1-57.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology,* **31,** 117-138.

Rao, T. J. (2016). On the history of certain early key concepts in sampling theory and practice. *Research Report,* **RR 2016-09**, C. R. Rao AIMSCS, Hyderabad, 1-30.

Rao, T. J. (2021a). Environmental statistics-A brief introduction. *Felicitation Volume of International Journal of Statistical Sciences in honour of Bimal Sinha and Bikas Sinha on their 75th Birthday.*

Rao, T. J. (2021b). Unordering of estimators in sampling theory: Revisited. *Journal of Statistical Theory and Practice,* **15**, 2021.

Rao, T. J. and Sinha, B. K. (2011). A brief history of statistics and its development in the Indian sub-continent. *International Journal of Statistical Sciences*, **11** (*Special Issue in memory of Prof. P. C. Mahalanobis).*

Rao, T. J., Sarkar, J., and Sinha, B. K. (2016). Randomized response and new thoughts on Politz-Simmons Technique. In *Handbook of Statistics,* **34** (Eds. C. R. Rao, A. Chaudhuri and T. C. Christophides), Chapter 15, 233-251.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics,* **5**, 119-127.

Yates, F. (1946). A review of recent statistical developments in sampling and sample surveys. *Journal of the Royal Statistical Society*, **109**, 12-43.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, **15**, 253-261.

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 11–19

# Global Hunger Index (GHI) Reminds the Phrase "Lies, Damned Lies and Statistics"

**Padam Singh**
*Medanta Institute of Education and Research, Medanta – The Medicity, Gurugram – Haryana*

Received: 31 May 2023; Revised: 08 June 2023; Accepted: 10 June 2023

**Abstract**

Poor ranking of India under Global Hunger Index (GHI) has raised concerns as it is contrary to the fact that India ranks fifth in the World Economy and fourth among leading agriculture producing countries. Researchers, planners, policy makers and government has taken a serious view of this. The indicators used under GHI do not measure hunger *per-se*, and, therefore, refereeing GHI as measure of Hunger is a misnomer. Zero Hunger being the priority goal under the Sustainable Development Goals (SDGs), it is imperative to develop a robust and country specific acceptable measure of hunger to track the progress. This is a big challenge for statisticians and other subject specialists.

*Key Words*: Child mortality; Hunger; Stunting; Over nutrition; Under nutrition; Wasting.

## 0. Prologue – Life and work of Late M.N. Das

*I was honoured to have delivered Dr. M. N. Das Memorial Lecture on his Birth Centenary during the conference of Society of Statistics, Computer and Applications (SSCA) on 15$^{th}$ February 2023 at Department of Statistics, University of Jammu, J&K. I was student of Dr. Das during my M.Sc. in Agricultural Statistics at the Institute of Agricultural Research Statistics (IARS) 1966–68. I was fascinated by his simple way of teaching the construction of confounded factorial designs in the course Design of Experiment-I. In spite of taking sampling theory as specialized subject during my M.Sc., I opted for Design of Experiments-II taught by Dr. Das wherein I learned Balanced Incomplete Block Design and Partially Balanced Incomplete Blocked Designs which helped me in developing some πPS sampling through these designs.*

*I owe my entire career to Dr. M. N. Das. When I joined the Institute as a Statistical Investigator in 1970, I was posted in the computer division, but he transferred me to the training division which provided me the opportunity of teaching the students of various courses. This helped me in developing skills for becoming an effective teacher. Also my first promotion from Statistical Investigator to Junior Statistician in 1972 was made by him during his tenure as Director, IARS.*

*Dr. Das was blessed with the power of intuition. Many of his path breaking contributions, especially in designs for factorial experiments, augmented designs, designs for fitting response surfaces, and statistical computing by writing his own software programs, had*

Corresponding Author: Padam Singh
Email: dr.padamsingh2013@gmail.com

*a strong intuitive appeal rather than complex algebraic manipulations. Taking lead from such a researcher, I too developed several newer sampling designs merely through the power of intuition.*

## 1. Background

Global Hunger Index (GHI) is being disseminated annually since 2006. It was initially published by International Food Policy Research Institute (IFPRI) and Welt Hunger Hilfe. In 2007, the Irish NGO Concern Worldwide also became a co-publisher. Presently it is released by Concern Worldwide and Welt Hunger Hilfe.

Table 1 presents the ranking of India for the last 6 years:

**Table 1: India's rank in GHI for last 6 years**

| Year | Rank Of India in GHI | Position from Bottom |
|---|---|---|
| **2017** | 100 Out of 119 countries | 20th |
| **2018** | 103 Out Of 119 countries | 17th |
| **2019** | 102 Out Of 117 countries | 16th |
| **2020** | 94 Out Of 107 countries | 14th |
| **2021** | 101 Out Of 116 countries | 16th |
| **2022** | 107 Out Of 121 countries | 15th |

Table 2 below presents relative ranking of India vis-à-vis neighbouring countries:

**Table 2: Rank of India and neighboring countries in GHI**

| Sr .No. | Country | Rank 2021 | Rank 2022 |
|---|---|---|---|
| 1 | India | 101 | 107 |
| 2 | Pakistan | 92 | 99 |
| 3 | Bangladesh | 76 | 84 |
| 4 | Nepal | 76 | 81 |
| 5 | Sri Lanka | 65 | 64 |

Surprisingly, India ranks below Sri Lanka, Pakistan, Bangladesh and Nepal.

This ranking is contrary to the fact that India ranks 5th in the World Economy and 4th among top Agricultural producing countries in the world.

The poor ranking of India has been a matter of concern. Planners, policy makers and noted columnists have argued that GHI is a misleading Hunger Index and that this faulty measure is creating a flawed narrative against India.

Among prominent researchers Messet (2011) pointed out that GHI has a problem of multiple counts, Hirotsugu (2015) observed that hunger measurement is complex methodological challenge which should not be crudely addressed by such an oversimplified concept and definitions as in GHI and Nigam (2016, 2018, 2019) argued that GHI has high upward bias because while hunger leads to stated syndromes but hunger alone is not the only reason for these.

In view of these issues, in 2019, the Indian Council of Medical Research (ICMR), Department of Health Research of the Ministry of Health and Family Welfare, Government of India, constituted an Expert Committee to critically review the Global Hunger Index.

Based on the report of the committee, a white paper entitled “Global Hunger Index does not really measure hunger – An Indian Perspective”, has been published in Indian Journal of Medical Research (IJMR).

During a meeting on the Global Hunger Index (GHI) held under the Chairpersonship of Dr. Rajiv Kumar, Honourable Vice Chairman, NITI Aayog on 12th November, 2021, it was clearly brought out that Global Hunger Index in its present form is a “misnomer” and does not measure “hunger” correctly due to the choice of indicators and its methodological issues.

This paper presents an overview and critical appraisal of indicators and data used in measuring hunger under GHI.

## 2. Definition of hunger

As per Oxford dictionary, Hunger is the state of not having enough food to eat. FAO defines hunger as “....an uncomfortable or painful physical sensation caused by insufficient consumption of dietary energy. The World Food Program (WFP) treats hunger as not having enough to eat to meet energy requirements. In common parlance, hunger is perceived as people eating inadequately due to poor access to food including lack of purchasing power.

Following couplet (Doha) from Kabir Das, a Hindi poet is very relevant in the context

साईं इतना दीजिए, जामे कुटुंब समाए मैं भी भूखा न रहूं, साधु न भूखा जाए।

कबीर दस जी कहते हैं कि परमात्मा तुम मुझे इतना दो कि जिसमें बस मेरा गुजरा चल जाए, मैं खुद भी अपना पेट पाल सकूं और आने वाले मेहमानों को भी भोजन करा सकूं।

## 3. Indicators used in GHI

The four indicators used in GHI are as under:

**i.** PUN: Proportion of the population with insufficient calories intake *i.e.*, Percentage of population consuming less than Minimum Dietary Energy Requirement of 1800 Kcal/Capita per day Cal (%).

**ii.** CST: Prevalence of stunting in children under five years old (low height for age) (%).

Children whose height-for-age (*Z*-score) is below minus two standard deviations from the median of the reference population are considered short for their age (stunted), or chronically undernourished.

**iii.** CWA: Prevalence of wasting in children under five years old (low weight for height) (%).

Children whose weight-for-height (*Z*-score) is below minus two standard deviations from the median of the reference population are considered thin (wasted), or acutely undernourished.

**iv.** CM: Proportion of children dying before attaining the age of five years (%).

## 4. Appropriateness of indicators in measuring hunger

The appropriateness of indicators in measuring hunger is discussed in what follows:

### 4.1. Is everybody consuming less than 1800kCal/capita/day (MDER) hungry?

As per FAO, those consuming less than MDER (minimum dietary energy requirement) of 1800 kCal/capita/day are categorised as “Undernourished”. If this is so there should not be any symptoms of “Over nutrition” among those consuming less than MDER. In order to examine this, the results of National Nutrition Monitoring Bureau (NNMB) survey on prevalence of overweight, obesity, hypertension and diabetes among those consuming less than 1800 calories/capita/day are presented in Table 3.

Evidently, a significantly sizable proportion of symptoms of over nutrition indicators suggest that all those consuming less than 1800 calorie are not undernourished or hungry. In fact Those who are obese, diabetic and hypertensive might be consuming less than 1800 calories by choice under doctor’s or nutritionist’s advice.

**Table 3: Prevalence of symptoms of over nutrition among those consuming less than 1800 Kcal per day**

| Particulars | | Urban 2016 | | Rural 2012 | |
|---|---|---|---|---|---|
| | | Male (%) | Female (%) | Male (%) | Female (%) |
| SBP ≥140 and/or DBP ≥90 | Hypertension | 33.1 | 22.5 | 22.2 | 20.3 |
| Blood Sugar (mg/dl) | Pre diabetic (110-126) | 10 | 10.4 | 8.7 | 8.8 |
| | Diabetic (≥126) | 14.1 | 10.5 | 7.3 | 6 |
| BMI WHO Classification | Overweight(25-29.9) | 28.1 | 30.4 | 8.3 | 11.2 |
| | Obese (≥30) | 5.7 | 15.9 | 0.9 | 2.5 |
| BMI Asian Classification | Overweight (23-27.49) | 36.1 | 33.3 | 15.2 | 16.8 |
| | Obese (≥27.5) | 16.6 | 28.5 | 3.1 | 6.2 |

The latest information on symptoms of over nutrition as per NFHS 5 are given in Table 4.

**Table 4: Symptom of over nutrition - NFHS 5**

| Symptoms | Men | | | Women | | |
|---|---|---|---|---|---|---|
| | Rural | Urban | Total | Rural | Urban | Total |
| **Overweight or Obese** | 19.3 | 29.8 | 22.9 | 19.7 | 33.2 | 24 |
| **Diabetic** | 14.5 | 17.9 | 15.6 | 12.3 | 16.3 | 13.5 |
| **Hypertension** | 22.7 | 26.6 | 24 | 20.2 | 23.6 | 21.3 |

Thus, a little less than one-fourth are overweight or obese. An equal proportion are hypertensive. Another 15% are diabetics. A large proportion among these might be consuming less calories by choice under doctors' advice.

Thus in measuring hungry out of all consuming less than 1800 calorie those who consume less by choice under doctor's advice should be discounted.

Further a significant proportion of adolescents and working population presently consume packaged food, fast food, soft drinks *etc*. Probably these are not properly captured in NSSO and NNMB surveys.

Further, the MDER of 1800 calories per capita per day seems to be on higher for India. If these issues are addressed the proportions of populations consuming less than MDER due to lack of purchasing power might be less than even one-fourth of the current estimated level.

### 4.2. Are stunting and wasting manifestation of hunger?

Global Hunger Index (GHI) considers stunting and wasting as its constituents. The inclusion of these indicators in GHI has implicit assumption that those who are hungry are likely to be short-statured and lighter. If stunting and wasting are manifestation of hunger, then there should not be stunted and wasted children among relatively rich who do not have problem of purchasing power.

Table 5 presents status of stunting and wasting for top quintiles based on latest NFHS 5.

**Table 5: Stunting and wasting among children according to wealth quintiles**

| Wealth Quintile | Stunting | Wasting |
|---|---|---|
| **Fourth (top 61/% – 80%)** | 28.1 | 17.7 |
| **Highest (top 81/% – 100%)** | 22.9 | 16.2 |

It is seen that significant proportion among top two quintiles have stunted and wasted children. Therefore, hunger is not the cause of stunting and wasting. Further the difference in height and weight between individuals are influenced by parental, genetic/biological factors and environmental factors rather than nutrition alone. Thus, stunting and wasting are not the manifestations of hunger.

### 4.3. Does hunger contribute to child mortality?

The million death study in India, covered 1176.6 thousand deaths during 2010 – 2015. Of these 57.0% deaths were neonatal and 43.0% as post neonatal. Based on this study, the details on major causes of death are summarized in Table 6.

**Table 6: Leading cause of under five deaths**

| Cause | Number of Death | Percent (%) |
|---|---|---|
| **Prematurity or low birth weight** | 370 | 31.4% |
| **Pneumonia** | 108 | 9.2% |
| **Neonatal infections** | 103 | 8.8% |
| **Diarrhoea** | 82 | 7.0% |
| **Injuries** | 82 | 7.0% |
| **Birth asphyxia or trauma** | 57 | 4.8% |

According to this, in India, pre-term birth resulted in 31.4% deaths. Other causes of under five deaths were pneumonia 9.2%; neonatal infections 8.8%; Diarrhoea 7.0%; injuries 7.0% and Birth asphyxia or trauma 4.8%. Thus hunger as the major cause of child mortality is not supported by the cause of death statistics. Importantly, no family will risk the child to die because of hunger.

## 5. Quality of data used

It is important to discuss the quality of data used in computation of GHI. In view of this, data used for calculation of GHI for India and other neighbouring countries are discussed as under:

**Indicator 1: Proportion of undernourishment in population (%)**

| Year | India | Nepal | Pakistan | Bangladesh | Sri Lanka |
|---|---|---|---|---|---|
| **2015** | 15.2 | 7.8 | 22 | 16.4 | 22 |
| **2016** | 15.2 | 7.8 | 22 | 16.4 | 22 |
| **2017** | 14.5 | 8.1 | 19.9 | 15.1 | 22.1 |
| **2018** | 14.8 | 9.5 | 20.5 | 15.2 | 10.9 |
| **2019** | 14.5 | 8.7 | 20.3 | 14.7 | 9 |

From this data (Indicator 1), it is seen that the level of under nourishment for Nepal is surprisingly about half of that of India, which is unbelievable. Further, for Sri Lanka, it shows a sudden drop of more than 10 percentage points in a year (2017-2018), whereas for other countries, the levels are almost stagnant.

**Indicator 2: Prevalence of stunting in children under five years (%)**

| Year | India | Nepal | Pakistan | Bangladesh | Sri Lanka |
|---|---|---|---|---|---|
| **2015** | 38.8 | 37.4 | 45.0 | 36.1 | 14.7 |
| **2016** | 38.7 | 37.4 | 45.0 | 36.4 | 14.7 |
| **2017** | 38.4 | 37.4 | 45.0 | 36.1 | 14.7 |
| **2018** | 38.4 | 35.8 | 45.0 | 36.1 | 17.3 |
| **2019** | 37.9 | 36.0 | 37.6 | 36.2 | 17.3 |

For stunting (Indictor 2), the values for India, Bangladesh, and Nepal are stagnant but there is a drop of 7 percentage points for Pakistan in a year (2018 *vs* 2019). On the other hand, there is an increase of 2.6% points for Sri Lanka in a year (2017 – 2018).

For wasting (Indicator 3) there is an increase of 6 percentage points for India (2016 *vs* 2017) and decrease of similar magnitude for Sri Lanka (2017 *vs* 2018) in a year. Further there is decline of 3.4% points for Pakistan (2018 *vs* 2019) and 2.0% points for Nepal (2017 *vs* 2018).

**Indicator 3: Prevalence of wasting in children under five years (%)**

| Year | India | Nepal | Pakistan | Bangladesh | Sri Lanka |
|---|---|---|---|---|---|
| **2015** | 15.0 | 11.3 | 10.5 | 14.3 | 21.4 |
| **2016** | 15.1 | 11.3 | 10.5 | 14.3 | 21.4 |
| **2017** | 21.0 | 11.3 | 10.5 | 14.3 | 21.4 |
| **2018** | 21.0 | 9.7 | 10.5 | 14.3 | 15.1 |
| **2019** | 20.8 | 9.6 | 7.1 | 14.4 | 15.1 |

**Indicator 4: Under-five mortality rate (%)**

| Year | India | Nepal | Pakistan | Bangladesh | Sri Lanka |
|---|---|---|---|---|---|
| **2015** | **5.3** | **4.0** | **8.6** | **4.1** | **1.0** |
| **2016** | **4.8** | **3.6** | **8.1** | **3.8** | **1.0** |
| **2017** | **4.8** | **3.6** | **8.1** | **3.8** | **1.0** |
| **2018** | **4.3** | **3.5** | **7.9** | **3.4** | **0.9** |
| **2019** | **3.9** | **3.4** | **7.5** | **3.2** | **0.9** |

Though the level of under 5 mortality (Indicator 4) showed large variations across countries, there is a steady decline in under 5 mortality across all countries.

The change in indicators of these magnitude in a year is not acceptable. All these variations in data could be due to change in methodology of data collection during those years in different countries.

The data for India on under nourishment is available only for 2012 and that for stunting wasting and child mortality for 2015. It is understood that for subsequent years data were collected using Gallup Surveys. Therefore these data lack credibility. Use of such data for computation of Index and ranking of country raises serious concerns.

## 6. Conclusions and way forward

The index intended to assess the hunger status for entire population is giving undue excessive weightage to under five children. Moreover, the indicators of undernourishment, stunting, wasting and child mortality do not measure hunger per se and thus, referring GHI as Hunger Index is misnomer. Importantly, the data used on these indicators lack credibility.

In view of this, Global Hunger Index (GHI) reminds us of the phrase "Lies, Damned Lies and Statistics".

This ill-conceived measure of hunger for ranking of countries should not be accepted.

As per the definition of hunger, we have to use a measure which captures "People eating inadequately due to poor access to food and lack of purchasing power."

The only indicator relevant in the context is the population consuming less than the minimum dietary energy requirement (MDER). In this regard the MDER of 1800 calories per capita per day needs to be revisited for India. In the surveys capturing this information through "household consumption expenditure" or "dietary surveys" in addition to collecting information on dietary intake should also collect the information on obesity, overweight, hypertension and diabetes. This information is needed to discount for those consuming less due to choice and ultimately net out those consuming less than MDER due to lack of purchasing power.

The measurement of hunger being a complex methodological issue, there is a need to develop a robust and acceptable country specific measure of hunger. It is a challenge for all concerned.

## Acknowledgement

Ms. Gargi Singh, Biostatistician, Medanta Institute of Education and Research has assisted me in preparation of the manuscript. The support received is gratefully acknowledged.

## References

Concern Worldwide and Welt Hinger Hilfe (2022). *Global Hunger Index Food Systems Transformation and Local Governance.*

International Food Policy Research Institute (2006). A Global Hunger Index: Measurement concept, ranking of countries, and trends. Food Consumption and Nutrition Division Discussion Paper 212. Washington, DC: IFPRI.

International Institute of Population Science (2022). National Family Health Survey India 2019-2021: India fact sheet. Mumbai: MoHFW, Government of India.

Masset, E. (2011). A review of hunger indices and methods to monitor country commitment to fighting hunger. *Food Policy*, **36**, S102-108.

Million Death Study Collaborators (2017). Changes in cause-specific neonatal and 1-59-month child mortality in India from 2000 to 2015: A nationally representative survey. *Lancet*, **390**, 1972-1980.

National Nutrition Monitoring Bureau (2022). Technical report No. 26. Diet and nutritional status of rural population, prevalence of hypertension and diabetes among adults and infant and young child feeding practices: Report of third repeat survey. Hyderabad: NIN- Indian Council of Medical Research.

Nigam, A. K. (2018). Global hunger index revisited. *Journal of the Indian Society of Agricultural Statistics*; **72**, 225-30.

Nigam, A. K. (2019). Improving global hunger index. *Agricultural Research*, **8**, 132-139.

Nigam, A. K., Srivastava, R, Tiwari, P. P., Saxena, R., and Shukla, S. (2016). Hunger in gram panchayats of Banda district (U.P.): A micro-level study. *Journal of the Indian Society of Agricultural Statistics*, **70**, 41-50.

Singh, P., Kurpad, A. V., Verma, D., Nigam, A. K., Sachdev, H. S., Pandey, A., Hemalatha, R., Deb, S., Khanna, K., Awasthi, S., and Toteja, G. S. (2021). Global hunger index does not really measure hunger-An Indian perspective. *Indian Journal of Medical Research*, **154**, 455-60.

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 21–30

# How Gainfully can the Additional Units be Used?

**Opendra Salam Singh[1], Gurumayum Sandweep Sharma[1] and Bikas K. Sinha[2]**
[1]*Manipur University, Imphal, Manipur, India*
[2]*Former Professor, Indian Statistical Institute, Kolkata, India*

Received: 01 June 2023; Revised: 08 June 2023; Accepted: 11 June 2023

## Preamble

Bikas K. Sinha [BKS] is the recipient of "VK Gupta Endowment Award for Achievements in Statistical Thinking and Practice - 2023". The Society of Statistics and Computer Applications [SSCA], New Delhi, gave this award — upon receiving recommendation from its Executive Council [EC]. While receiving the award, BKS made an online presentation during SSCA Annual Conference held in Jammu during February 15-17, 2023. This paper originated from that presentation. BKS is happy to induct Opendra Salam Singh and Gurumayum Sandweep Sharma of the Department of Statistics, Manipur University, Imphal as his collaborators. As BKS says "I have chosen to speak on a topic which is simple to state and comprehend. Yet, the technicalities are quite involved." Simply stated, it goes almost like a proverb: "Larger the sample size, more is the precision"! AND we all know that this is indeed true for [SRSWOR (*N*, *n*), Sample Mean] Strategy. What about [SRSWR (*N*, *n*), Sample Mean] strategy? We must qualify sample mean under srswr: mean based on all units including repeats or mean based on distinct units only? Under both the situations, the claim is valid in some sense. Research scholars may engage themselves for a clear proof.

We intend to discuss some features of this problem of variance reduction *via* enhanced resources in terms of possession of additional population units at a later stage.

*Key Words*: Sampling designs; Sampling strategies; Unbiased estimators [UEs]; Homogeneous UEs [HUEs]; Linear UEs [LUEs]; HLUEs; SRS WR/WOR schemes; Horvitz-Thompson estimator [HTE]; First and second order inclusion probabilities; Connected sampling designs; Additional units; Improved sampling strategies; Lanke's estimator.

## 1. Introduction

We start with a Sampling Strategy based on a Fixed-Size ($n$) [abbreviated as FS($n$)] Sampling Design and an HLUE e($s(n)|Y$) of a Finite Population Mean $\bar{Y}$ corresponding to a study character $Y$. Once the sampling design has been chosen and implemented, and a sample $s(n)$ has been chosen and, further, data collection has been completed, we are told about Enhanced Resource in the sense of $k$ additional units! The enhanced sample size now becomes $(n + k)$, once an additional sampling design of fixed size ($k$) [$FS(k)$] defined over the complement of $s(n)$ is adopted. It leads to the revised HLUE $e(s(n + k)|Y)$ based on the union of the two samples of which $s(n)$ is already at hand.

Corresponding Author: Bikas K. Sinha
Email: bikasksinha118@gmail.com

One pertinent question to be asked is: Whatever be the choice of the initial HLUE $e(s(n)|Y)$ and the choice of the additional sampling design $FS(k)$ [on the complement of $s(n)$], does there exist a suitably defined HLUE $\mathrm{e}(s(n + k)|Y)$ which provides uniformly smaller variance than $e(s(n)|Y)$?

Clearly, in the case of *SRSWOR* $(N, n)$, followed by *SRSWOR* $(N - n, k)$, we end up with *SRSWOR* $(N, n + k)$, and hence choice of the corresponding sample means is well understood for the domination result to hold for every $k \geq 1$.

However, in case of *SRSWR* $(N, n)$, the follow-up sampling operation could be

(i) *SRSWR* $(N, k)$ or, (ii) *SRSWR* $(N - v(n); k)$

where $v(n)$ refers to the number of distinct units selected under *SRSWR* $(N, n)$. In a way, under (ii), therefore, the sample is selected under *SRSWR* out of the complement of the units already selected under *SRSWR* $(N, n)$. Whereas the combination under (i) refers to two independent draws from the whole population, under (ii), the two sets of samples are necessarily disjoint. However, within each, units drawn are not necessarily distinct, as we take recourse to *WR* sampling.

The question we ask is: For a given choice of $e(s(n)|Y)$, what is the choice of $e(s(n)$ U $s(k)|Y)$ for variance reduction? Here, $s(n) \cup s(k)$ must be understood in the most general sense.

This is apparently not an easy problem to address. There are two choices for $e(s(n)|Y)$ under *SRSWR* $(N, n)$: mean based on all units, and mean based on distinct units [notation $v(n)$]. Note that *SRSWR* $(N - v(n), k)$ excludes the distinct units selected in the first round. So, data analysis is conditional not only on $v(n)$, but also on the actual units selected under $s(n)$. Naturally, these are excluded during the second draw. We leave it for research scholars to ponder over this non-standard inference problem.

Bagchi and Sinha [2022] have addressed a different version of this problem. We do not intend to enter into this matter.

## 2. Data analysis under FS(.) designs

Consider $FS(.)$ sampling designs – both Initial $D$ $(N; n)$ and Extension $D$ $(N - n;$ k). Denote by $[D\ (N, n), e(s(n)|Y)]$ the initial sampling strategy for unbiased estimation of a finite population total or mean.

Let $D^*$ $(N - n, k|s(n))$ be the follow-up sampling design $FS$ $(N - n, k)$, conditional on exclusion of $s(n)$.

We ask the question: Given $e(s(n)|Y)$, how would one define $e^*(s(n + k)|Y|s(n))$ – once totally new additional $k$ units are available *via* $s(k)$ - following $FS$ $(N - n, k)$, defined over compliment of $s(n)$, for every $s(n)$ with $P\ (s(n)) > 0$?

Naturally, we desire:

(i) $E^{*}\,[e^{*}\,(s(n+k)|Y\,|s(n))] = E\,[e(s(n)|Y\,)]$ (1)

(ii) $V^{*}\,[e^{*}\,(s(n+k)|Y|s(n))] \leq V\,[e(s(n)|Y)]$ (2)

uniformly in $Ys$, where $E^{*} = E1E2$ and $V^{*} = V1E2 + E1V2$, in usual notations.

This specific problem has been resolved by Lanke (1975) who provided an explicit expression for $e^{*}(s(n+k)|Y)$'s in terms of the $e(s(n)|Y)$'s, provided that $s(n)$ is a subset of $e(s(n+k))$.

We take up this exercise in the sequel.

## 2.1. Lanke's formula

Lanke (1975) considered extending an arbitrary sampling strategy $[D(N, n), e(s(n)|Y)]$ to another sampling strategy $[D(N, m{=}n+k), e(s(n+k)|Y)]$ *via* $Q(N-n, k)$ so that $[D(N, m = n+k), e(s(n+k)|Y)]$ is better than $[D\,(N, n), e(s(n)|Y)]$, irrespective of the choice of $Q(N-n, k)$.

Lanke proposed the estimator $e(s(n+k)|Y)$ through the relation

$$e(s(n+k)|Y)[P(s(n+k))] = \sum_{s(n)\in s(n+k)} e(s(n)|Y)[P\,(\,s(n))Q(s(n+k)-s(n)] \quad (3)$$

Here summation is over all $s(n)$ [subsets of $s(n+k)$].

Further,

$$P\,[s(n{+}k)] = \sum_{s(n)\in s(n+k)}[P\,(s(n))Q(s(n+k)-s(n))] \quad (4)$$

summation being over all $s(n)$ [subsets of $s(n+k)$].

It transpires that Lanke basically applied Rao-Blackwellization technique *i.e.*, averaging technique to produce estimator(s) with reduced sum of squares. We display the technical details below. Upon squaring both sides of (3) and rewriting the same, we obtain

$$e^{2}(s(n+k)|Y)[P\,(s(n+k))] =$$

$$[\,\sum_{s(n)\in s(n+k)} e(s(n)|Y)[P(s(n))Q(s(n+k)-s(n))]]^{2}/P(s(n+k)) \quad (5)$$

By appealing to C-S inequality [elaborated below], we derive from (5):

$$e^{2}(s(n+k)|Y)[P(s(n+k))] \leq [\,\sum_{s(n)\in s(n+k)} e^{2}(s(n)|Y)[P(s(n))Q(s(n+k)-s(n))]] \quad (6)$$

which further simplifies to

$$[\,\sum_{s(n)\in s(n+k)} e^{2}(s(n)|Y)[P(s(n))] \quad (7)$$

Hence the domination result follows in appropriate subgroups and hence on the whole.

### 2.2. Illustrative example: Lanke's formula

Here we take an example to demonstrate the domination result. We start with $N = 10$, $n = 5$, $k = 2$. Let us adopt the initial sampling design in the form:

**Table 1: Initial sampling design of fixed size *n* = 5**

| Sl. No. | *P* (…) |
|---|---|
| 1. | P (1, 2, 3, 4, 5) = 0.075 |
| 2. | P (1, 3, 5, 8, 10) = 0.105 |
| 3. | P (1, 4, 6, 7, 9) = 0.165 |
| 4. | P (4, 6, 7, 8, 10) = 0.135 |
| 5. | P (2, 3, 6, 9, 10) = 0.145 |
| 6. | P (3, 4, 7, 8, 10) = 0.175 |
| 7. | P (5, 6, 7, 8, 9) = 0.180 |
| 8. | P (2, 4, 6, 7, 8) = 0.020 |

The extended design for $k = 2$ must be defined for every sample $s(n)$ [listed above] on its compliment with reference to the whole set of $N = 10$ units. Note that the design shown above is already connected in the sense of positive probability attached to all pairwise units *i.e.*, $P(i, j) > 0$ for all pairs. So, the choice of complimentary samples for the extended design is very simple and we need not restrict to any conditions except that these are complimentary in nature! Of course, the sample size $k = 2$ has to be kept in mind. We take up the following Example of extended design to this effect.

**Table 2: Initial description of *s*(*n*) and extension design using *s*(*k*)**

| Sl. No. | Initial Design | Extension Design |
|---|---|---|
| 1. | (1, 2, 3, 4, 5) | *P* (6, 7) = 0.4; *P* (6, 9) = 0.3; *P* (8, 10) = 0.3∗ |
| 2. | (1, 3, 5, 8, 10) | *P* (2, 4) = 0.7∗; P (4, 7) = 0.3 ∗ ∗ |
| 3. | (1, 4, 6, 7, 9) | *P* (2, 3) = 0.5; *P* (5, 8) = 0.5 ∗ ∗ ∗ ∗ |
| 4. | (4, 6, 7, 8, 10) | *P* (3, 9) = 0.3 ∗ ∗∗; *P* (3, 5) = 0.7 |
| 5. | (2, 3, 6, 9, 10) | *P* (4, 5) = 0.6; *P* (5, 7) = 0.3; *P* (5, 8) = 0.1 |
| 6. | (3, 4, 7, 8, 10) | *P* (1,5) = 0.4∗∗; *P* (2, 9) = 0.4; *P* (6, 9) = 0.2∗∗∗ |
| 7. | (5, 6, 7, 8, 9) | *P* (1, 4) = 0.3∗∗∗∗; *P* (2, 10) = 0.5; *P* (3, 4) = 0.2 |
| 8. | (2, 4, 6, 7, 8) | *P* (3, 9) = 1.00 |

**Remark 1**: Note that for the last design [Sl. No. 8], the extension design is degenerate.

Composition of samples based on extended sampling design is shown below:

| | |
|---|---|
| (1, 2, 3, 4, 5, 6, 7) | (8.1) |
| (1, 2, 3, 4, 5, 6, 9) | (8.2) |
| (1, 2, 3, 4, 5, 8, 10)* | (8.3) |
| (1, 2, 3, 4, 5, 8, 10)* | (8.4) |
| (1, 3, 4, 5, 7, 8, 10)** | (8.5) |

| | |
|---|---|
| (1, 2, 3, 4, 6, 7, 9) | (8.6) |
| (1, 4, 5, 6, 7, 8, 9)*** | (8.7) |
| (3, 4, 6, 7, 8, 9, 10)**** | (8.8) |
| (3, 4, 5, 6, 7, 8, 10) | (8.9) |
| (2, 3, 4, 5, 6, 9, 10) | (8.10) |
| (2, 3, 5, 6, 7, 9, 10) | (8.11) |
| (2, 3, 5, 6, 8, 9, 10) | (8.12) |
| (1, 3, 4, 5, 7, 8, 10)** | (8.13) |
| (2, 3, 4, 7, 8, 9, 10) | (8.14) |
| (3, 4, 6, 7, 8, 9, 10)**** | (8.15) |
| (1, 4, 5, 6, 7, 8, 9)*** | (8.16) |
| (2, 5, 6, 7, 8, 9, 10) | (8.17) |
| (3, 4, 5, 6, 7, 8, 9) | (8.18) |
| (2, 3, 4, 6, 7, 8, 9) | (8.19) |

**Remark 2:** There are altogether 19 extended samples formed through the extension formula. However, not all are distinct. For example, the sample (1, 2, 3, 4, 5, 8, 10)* is formed of (i) (1, 2, 3, 4, 5) combined with (8, 10) as well as of (ii) (1, 3, 5, 8, 10) combined with (2, 4). Lanke argued that once the extended sample is available through the extension formula, both the subsets (i) and (ii) are available and they produce $e(s(n)|Y)$ based on initial sample $s(n)$ under both (i) and (ii). Then he suggested the formula shown above in (3) for combining the two estimators. In this example, for the extended sample (1, 2, 3, 4, 5, 8,10)*, the formula yields

$$e((1, 2, 3, 4, 5, 8, 10)*) = [e((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((8, 10)|s(n))+$$

$$e((1, 3, 5, 8, 10))P(1, 3, 5, 8, 10)P((2, 4)|s(n))]/$$

$$[P(1, 2, 3, 4, 5)P((8, 10)|s(n)) + P(1, 3, 5, 8, 10)P((2, 4)|s(n))] \quad (9)$$

In effect, the estimator based on the extended sample is a convex combination of the two initial estimators listed in (i) and (ii) and these are both available whenever the extended sample (1, 2, 3, 4, 5, 8, 10) is realized. It may be noted that for the extended sample (1, 2, 3, 4, 5, 8, 10), $P$ (1, 2, 3, 4, 5, 8, 10) is given by the denominator above in (9).

Towards variance, or equivalently, sum of squares [SS] computation, we find: $e^2((1, 2, 3, 4, 5, 8, 10))P((1, 2, 3, 4, 5, 8, 10)) = [e((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((8, 10)|s(n))+ e((1, 3, 5, 8, 10))P(1, 3, 5, 8, 10)\mathrm{P}((2, 4)|\mathrm{s}(\mathrm{n}))]^2/[\mathrm{P}((1, 2, 3, 4, 5, 8, 10))]$. (10)

Set

$$a1 = e(1, 2, 3, 4, 5)[P((1, 2, 3, 4, 5))P((8, 10)|s(n))]^{1/2};$$

$$b1 = [P((1, 2, 3, 4, 5))P((8, 10)|s(n))]^{1/2} \quad (11)$$

$$a2 = e(1, 3, 5, 8, 10)[P((1, 3, 5, 8, 10))P((2, 4)|\mathrm{s}(n))]^{1/2};$$

$$b2 = [P((1, 3, 5, 8, 10))P((2, 4)|s(n))]^{1/2} \quad (12)$$

By C-S inequality, we know that

$$[a1b1 + a2b2]^2 \leq [a1^2 + a2^2][b1^2 + \mathrm{b}2^2] \tag{13}$$

which leads to

$$\text{RHS of (10)} \leq [e^2((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((8, 10)|s(n)) \\ + e^2((1, 3, 5, 8, 10))P(1, 3, 5, 8, 10)P((2, 4)|s(n))]. \tag{14}$$

Once all the samples are utilized like in the above, we can go back to computation of the upper bound of the sum of squares [*SS*] of the extended estimator $e(s(n + k)|Y)$. This yields, for example, terms like

$$e^2((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((8, 10)|s(n)); \tag{15}$$

$$e^2((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((6, 7)|s(n)); \tag{16}$$

$$e^2((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)P((6, 9)|s(n)). \tag{17}$$

These three expressions add to $e^2((1, 2, 3, 4, 5))P(1, 2, 3, 4, 5)$, upon obvious simplification. Likewise, we carry on similar computations of the *SS* for the estimators based on extended samples and upon application of C-S inequality, we end up with upper bounds as *SS* based on the samples in the initial sampling design.

**Remark 3**: It is interesting to note that the three extension designs [(8.1), (8.2) and (8.3)*] arising out of a single initial sample do provide three different estimators for the population parameter. After that, the *SS* for each estimator is examined in the light of the C-S inequality. Taken together, we find that the *SS* for the estimator based on extension design is less than or equal to that of the original estimator. We provide below in Section 4 necessary details to encourage the interested teachers and researchers follow the technicalities in settling the claim.

## 3. Behavior of Horvitz-Thompson estimator

In as early as 1967, Prabhu-Ajgaonkar discussed the possibility of an hlue based on a sample of size n to outperform an hlue based on an extended design of size $(n + 1)$ - the two estimators belonging to the same class of hlues. The sampling design was chosen to be the Midzuno Sampling Scheme for a sample of size $n = 2$ and it was to be extended to the Generalized Midzuno Sampling Scheme for a sample of size $n = 3$. However, he actually worked out the case of $n = 1$ against $n = 2$ and that was not at all appealing. Our attempt to work for $n = 2$ to $n = 3$ did not go through with the set-up adopted by him.

Starting with an arbitrarily specified initial $FS(N, n)$ design and extending it to a $FS(N, n + k)$ design by increasing the sample size from $n$ to $n + k < N$, confining to the use of the *HTE* in both the designs, one may not succeed in uniformly improving over the *HTE* based on the original design. A quick and tricky proof goes like this. Let $(\pi_i(n)|FS(N, n))$; $i = 1,2,...,N)$ denote the first order inclusion probabilities based on the original design so that $\sum(\pi_i(n)|FS(N, n)) = n$. Set $Y_i = K\pi_i(n)/n$, $i = 1,2, ...,N$, where $K$ is an arbitrary positive constant.

Then $HTE[s(n)] = \sum_{i\in s(n)} Y_i/\pi_i(n) = K$ for every $s(n)$ with $P(s(n)) > 0$. Hence, $V(HTE) = 0$ at the stated values of $Y_i$'s. In the same vein, evaluated at the same $Y$-values,

$$HTE(s(n+k)) = \sum_{i\in s(n+k)} Y_i/\pi_i(n+k) \tag{18}$$

$$= [K/n]\sum_{i\in s(n+k)} \pi_i(n)/\pi_i(n+k) \tag{19}$$

Therefore, unless $\pi_i(n)/\pi_i(n + k)$ is the same for all $i = 1,2,...,N$, the second estimator has a strictly positive variance. Therefore, uniform domination is not possible using *HTE* in both the situations.

For the case where the extended sampling design $Q(N - n, k|s(n))$ is SRSWOR, Sinha (1980) presented simple conditions on the first and second order inclusion probabilities of the original sampling design $FS(N, n)$ so that $HTE(s(n + t)|\mathrm{Y})$ is better than $HTE(s(n + t - 1)|Y)$ simultaneously for all $t = 1,2, ....,k$ for any arbitrary choice of $k < N - n$.

Sengupta (1982) extensively studied the properties of Lanke's estimator for various choices of $e(s(n)|Y)$ [based on $FS(N, n)$] and its extensions. In particular, he observed that (i) Lanke's estimator, even though it improves over the estimator $e(s(n)|Y)$, may itself turn out to be inadmissible, and (ii) if the estimator $e(s(n)|Y)$ is the sample mean (or *HTE*) then there may not exist an extended sampling design such that Lanke's estimator based on $e^*$ is again the sample mean (or *HTE*). He also showed that when $e(s(n)|Y)$ is the sample mean and the extended sampling design is $SRSWOR(N - n, k)$, Lanke's estimator will again be the sample mean if and only if the initial sampling design $FS(N, n)$ is itself *SRSWOR*.

Some other features of uses of additional resources are discussed in Sengupta *et al.* (1987). Another interesting and related paper on finding admissible estimators is Patel and Dharmadhikari (1977).

## 4. Variance comparison and effect of sample size

With reference to the example taken up above, we will examine the effect of sample sizes $n$ versus $n + k$ by computing 'Efficiency per Unit Observation'. Note that in general terms, efficiency is defined as the reciprocal of variance and efficiency per unit observation is to be computed as reciprocal of

$$n \times V[e(s(n)|Y)] \text{ as against } (n + k) \times V[e(s(n + k)|Y)]. \tag{20}$$

We fix the population $Y$-values as

[1,2,3, .....,10] with a total of 55 and mean of 5.5.

We now opt for the *HTE* [for the population total] based on the original design. In Table 3, we display all the initial samples and the *HTE*-values based on them. Also, we show the corresponding probabilities.

**Table 3: $s(n)$ $P(\mathrm{s}(n))$ $e(s(n)|Y)$**

| Sl. No. | $s(n)$ | $P(\mathrm{s}(n))$ | $e(s(n)\mid Y)$ |
|---|---|---|---|
| 1. | (1,2,3,4,5) | 0.075 | 38.3934 |
| 2. | (1,3,5,8,10) | 0.105 | 53.6526 |
| 3. | (1,4,6,7,9) | 0.165 | 48.2112 |
| 4. | (4,6,7,8,10) | 0.135 | 57.8106 |
| 5. | (2,3,6,9,10) | 0.145 | 59.8600 |
| 6. | (3,4,7,8,10) | 0.175 | 54.5083 |
| 7. | (5,6,7,8,9) | 0.180 | 64.9370 |
| 8. | (2, 4, 6, 7, 8) | 0.020 | 48.2868 |

Computations yield for the _HTE_ of the population total based on the initial design:

(1) $E\ [HTE] = 55.1453$

(2) $V(HTE) = 52.6695$

Next, towards computation of Lanke’s estimator, we obtain the following:

First, we show $s(n + k)$, next follows $e(s(n + k))$, lastly we show $P(s(n + k))$.

| | | | |
|---|---|---|---|
| 1 | (1,2,3,4,5,6,7) | _e_(1,2,3,4,5) | 0.0300 |
| 2 | (1,2,3,4,5,6,9) | _e_(1,2,3,4,5) | 0.0225 |
| (1, 2) | _combined_ | _e_(1, 2, 3, 4, 5) = 38.3934 | 0.0525 |
| 3 | (1,2,3,4,5,8,10)* | _e_(1,2,3,4,5) | 0.0225 |
| 4 | (1,2,3,4,5,8,10)* | _e_(1,3,5,8,10) | 0.0735 |
| (3,4) | _combined_ | (1, 2, 3, 4, 5, 8, 10)* [0.0225×_e_(1,2,3,4,5)+ | |
| | 0.0735×_e_(1,3,5,8,10)]/0.0960=50.0762 | | 0.0960 |
| 5 | (1,3,4,5,7,8,10)** | _e_(1,3,5,8,10) | 0.0315 |
| 13 | (1,3,4,5,7,8,10)** | _e_(3,4,7,8,10) | 0.0700 |
| (5,13) | _combined_ | (1,3,4,5,7,8,10)** [0.0315×_e_(1,3,5,8,10)+ | |
| | 0.070×_e_(3,4,7,8,10)]/0.1015=54.2427 | | 0.1015 |
| 6 | (1,2,3,4,6,7,9)) | _e_(1,4,6,7,9) = 48.2112 | 0.0825 |
| 7 | (1,4,5,6,7,8,9)*** | _e_(1,4,6,7,9) | 0.0825 |
| 16 | (1,4,56,7,8,9)*** | _e_(5,6,7,8,9) | 0.0540 |
| (7,16) | _combined_ | (1,4,5,6,7,8,9)*** [0.0825×_e_(1,4,6,7,9) + | |
| | 0.0540 × _e_(5, 6, 7, 8, 9)]/0.1365 = 54.8280 | | 0.1365 |
| 8 | (3,4,6,7,8,9,10))**** | _e_(4,6,7,8,10) | 0.0405 |
| 15 | (3,4,6,7,8,9,10)**** | _e_(3,4,7,8,10) | 0.0350 |
| (8,15) | _combined_ (3,4,6,7,8,9,10))**** [0.0405×_e_(4,6,7,8,10)+ | | |
| | 0.035 × _e_(3, 4, 7, 8, 10)]/0.0755 = 56.2797 | | 0.0755 |
| 9 | (3,4,5,6,7,8,10) | _e_(4,6,7,8,10) = 57.8106 | 0.0945 |

| | | | |
|---|---|---|---|
| 10 | (2,3,4,5,6,9,10) | $e$(2,3,6,9,10) | 0.0870 |
| 11 | (2,3,5,6,7,9,10) | $e$(2,3,6,9,10) | 0.0435 |
| 12 | (2,3,5,6,8,9,10) | $e$(2,3,6,9,10) | 0.0145 |
| (10, 11, 12) | *combined* | $e$(2, 3, 6, 9, 10) = 59.86 | 0.1450 |
| 14 | (2,3,4,7,8,9,10) | $e$(3,4,7,8,10) = 54.5083 | 0.0700 |
| 17 | (2,5,6,7,8,9,10) | $e$(5,6,7,8,9) | 0.0900 |
| 18 | (3,4,5,6,7,8,9) | $e$(5,6,7,8,9) | 0.0360 |
| (17, 18) | *combined* | $e$(5, 6, 7, 8, 9) = 64.9370 | 0.1260 |
| 19 | (2,3,4,6,7,8,9) | $e$(2,4,6,7,8) = 48.2868 | 0.0200 |

**Remark 4:** It may be noted that we started with a total of 8 samples for the sample size $n = 5$ and after extension, we ended up with a total of 19 samples. However, for the estimator in the above, we have ended up with a total of 11 samples. Computations yield:

(1) $E[e(n + k)|Y] = 55.1453$

(2) $V(e(n + k)|Y) = 38.2284$

Therefore, Lanke's estimator performs better with the use of additional units. Finally, referring to (4.1), we work out efficiency of the extended estimator by comparing 5 × $V$ (*HTE*) with 7 × $V$ (*extended estimator*). The quantities are respectively 263.3475 and 267.5988. Therefore, according to this criterion, Lanke's extension formula fails to provide a more efficient estimator.

## 5. Concluding remarks

Mr. Sharma [Research Scholar in Statistics] and Dr. Singh [Statistics Faculty] express their gratitude to Prof. K. K. Singh Meitei, Head, Department of Statistics, Manipur University, Imphal, for providing excellent academic atmosphere towards conducting collaborative research and for creating opportunities for Prof. Sinha's multiple visits for collaborative research with the faculty and students of this department.

We raised the issue of effective use of additional resources. In general terms, Lanke's estimator serves this purpose. However, this estimator itself may not be admissible in the class of competing estimators [Sengupta *et al.* (1987)]. Further, though variance reduction is achieved, efficiency per unit observation may not necessarily increase with enhanced resources. This area of research still holds rich rewards for those who wish to venture into the perplex question of profitable use of additional resources.

## References

Bagchi, S. B. and Sinha, B. K. (2022). Some inferential aspects of mixture sampling designs. *Thai Statistician*, **20**, 233 - 239.

Lanke, J. (1975). *Some Contributions to Theory of Survey Sampling*. Ph.D. Thesis, University of Lund, Sweden.

Patel, H. C. and Dharmadhikari, S. W. (1977). On linear invariant unbiased estimators in survey sampling. *Sankhya C*, **39**, 21-27.

Prabhu-Ajgaonkar, S. G. (1967). The effect of increasing sample size on the precision of an estimator. American Statistician, **21**, 26-28.

Sengupta, S. (1982). *Further Studies on Some Strategies in Sampling Finite Populations*. Unpublished Ph.D. Thesis, Calcutta University.

Sengupta, S. N., Sinha, B. K. and Sastry, G. P. (1987). Some inferential aspects of finite population sampling with additional resources. *Journal of Statistical Planning and Inference*, **16**, 203 - 211.

Sinha, B. K. (1980). *On the Concept of Admissible Extensions of Sampling Designs and Some Related Problems*. Technical Report, No. 14/80, Stat-Math. Division, Indian Statistical Institute, Calcutta.

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 31-41

# A Nonparametric Structure Based on Hierarchical Dirichlet Processes for Studying Gene-Gene and Gene-Environment Interactions

**Durba Bhattacharya**[1] **and Sourabh Bhattacharya**[2]
[1]*Department of Statistics, St. Xavier's College (Autonomous), Kolkata*
[2]*Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata.*

Received: 15 June 2023; Revised: 01 July 2023; Accepted: 04 July 2023

## Abstract

It has been established in many studies that genes interact in complex networks among themselves and with various environmental factors to cause diseases. In this article, we discuss how realistic statistical models for case-control genotype data can be developed using nonparametric Bayesian techniques founded on hierarchies of Dirichlet process based mixture models for studying such complex interaction structures. Suitable Bayesian hypothesis testing procedures need to be developed for uncovering the roles of genes, environment and their interactions in case-control studies. Empowered with an efficient TMCMC based parallelisable algorithm, application of our ideas to data simulated under five different setups of disease-gene-environment association as well as a real, Myocardial Infarction (MI) dataset yielded interesting results that not only agrees with the existing works in this area, but also gives some novel insights into the genetic interactions underlying the disease.

*Keywords:* Hierarchical Dirichlet process; Case-control study; Myocardial infarction; Parallel processing; Transformation based MCMC; Gene-gene and gene-environment interaction.

## 1. Introduction

Present day biomedical research is pointing towards the significance of interactions between genes and the environment in causing complex diseases. According to Hunter (2005), considering the contributions of genes and environment to a disease separately, ignoring their interactions, might lead to incorrect estimation of the disease proportion that is explained by these factors. The additive linear models or the logistic model based approaches, (see for example Ahn *et al.* (2013), Wen and Stephens (2014) and Liu, Ma and Amos (2015) resting on Fisher's definition of interaction result in the inclusion of a large number of interaction terms even with a moderate number of genetic and environmental factors. The existing Bayesian techniques like BEAM, EpiBN study interaction by identifying the SNPs that influence the disease risk given particular allele combinations, ignoring the genes as functional units. In a nutshell, none of the existing methods, classical or Bayesian, attempts simultaneous modelling of the uncertainties associated with the genes as the functional units along with the interactions, both at SNP and gene level through unified statistical models.

The fact that the genetic data may arise from a stratified population with an unknown number of subpopulations makes the problem all the more demanding. The Bayesian semiparametric model proposed by Bhattacharya and Bhattacharya (2020 a) takes care of the

Corresponding Author: Durba Bhattacharya
E-mail: durba@sxccal.edu

above mentioned problems by proposing a model based on Dirichlet Processes (DP) and a hierarchical matrix-normal distribution, encapsulating the mechanism of dependence among genes under environmental effects with respect to genotype data arising out of a possibly stratified population.

As the environmental variables may affect the gene-gene interactions of individuals differently, depending on the extent and type of their exposure to the environmental factors, in this article, we introduce a novel Bayesian nonparametric model for gene-gene and gene-environment interactions for case-control genotype data that solves the issues detailed above. Our model represents the individual genotype data as finite mixtures based on Dirichlet processes as before, but instead of the hierarchical matrix normal distribution, we introduce a hierarchy of Dirichlet processes that create appropriate nonparametric dependence among the genes induced by the environment. We develop a novel and highly parallelisable Markov Chain Monte Carlo (MCMC) methodology that combines the efficiencies of modern parallel computing infrastructure, Gibbs steps, retrospective sampling methods, and Transformation based Markov Chain Monte Carlo (TMCMC). Application of our model and methods to five different simulation experiments for the validation purpose yielded quite encouraging results. Application to a real myocardial infarction (MI) case-control type dataset yielded results which broadly agree with the results reported in the literature, and also provided new and interesting insights into the mechanisms of 4 gene-gene and gene-environment interactions.

The rest of our paper is structured as follows. We introduce our HDP-based Bayesian nonparametric gene-gene and gene-environment interaction model in Section 2, and in Section 3 we extend the Bayesian hypothesis testing procedures proposed in Bhattacharya and Bhattacharya (2020 a) to learn about the roles of genes, environmental variables and their interactions in case-control studies, with respect to our current HDP model. In Section 4 we briefly discuss the results of application of our model and methodologies to 5 biologically realistic simulated data sets, the details of which are provided in section S-3 of the supplement in Bhattacharya and Bhattacharya (2020 b). In Section 5 we analyse the real MI dataset using our ideas, demonstrating quite interesting and insightful outcome. Finally, we summarize our work with concluding remarks in Section 6.

## 2. Bayesian nonparametric model based on hierarchies of Dirichlet process for gene-gene and gene-environment interactions

### 2.1. Case-control genotype data

For $s$ = 1, 2 denoting the two chromosomes, let ${x^s}_{ijkr}$ = 1 and ${x^s}_{ijkr}$ = 0 indicate the presence and absence of the minor allele of the $i$-th individual belonging to the $k$-th group, for $k$ = 0, 1, with $k$ = 1 denoting case; at the $r$-th locus of $j$-th gene, where $i$ = 1, . . . , $N_k$; $r$ = 1, . . . , $L_j$ and $j$ = 1, . . . , $J$; let $N$ = $N_1$ + $N_2$. Let $E_i$ denote a set of environmental variables associated with the $i$-th individual. We now proceed to model this case-control genotype and the environmental data using our Bayesian semiparametric model, described in the next few sections.

### 2.2. Mixture models based on Dirichlet processes

Let $x_{ijkr}$= (${x^1}_{ijkr}$ , ${x^2}_{ijkr}$) and $L$ = max ($L_1$, ... , $L_J$). We assume that for every triplet ($i$, $j$, $k$), $X_{ijk}$ = ($x_{ijk1}$, ..., $x_{ijkL}$) have the mixture distribution

$$[X_{ijk}] = \sum_{m=1}^{M} \pi_{mijk} \prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr}) \tag{1}$$

where $f(.\,|p_{mijkr})$ is a Bernoulli mass function given by:

$$f(x_{ijkr}|p_{mijkr}) = {p_{mijkr}}^{{x^1}_{ijkr}+{x^2}_{ijkr}}(1-p_{mijkr})^{2-({x^1}_{ijkr}+{x^2}_{ijkr})} \tag{2}$$

In the above, $M$ denotes the maximum number of mixture components and $p_{mijkr}$ stands for the minor allele frequency at the *r*-th locus of the *j*-th gene for the *i*-th individual of the *k*-th case/control group.

Allocation variables $z_{ijk}$, with probability distribution

$$[z_{ijk} = m] = \pi_{mijk}, \tag{3}$$

for *i* = 1, . . . , $N_k$ and *m* = 1, . . . , *M*, allow representation of (1) as

$$[X_{ijk}|z_{ijk}] = \prod_{r=1}^{L} f(x_{ijkr}|p_{mijkr}) \tag{4}$$

Following Majumdar *et al.* (2013), Bhattacharya and Bhattacharya (2018), we set $\pi_{mijk}$ = 1/*M*, for *m* = 1, . . . , *M*, and for all (*j*, *k*).

Letting $p_{mijk}$ = $(p_{mijk1}, \ldots, p_{mijkL})$, we assume that

$$p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk} \overset{iid}{\sim} G_{ijk}; \tag{5}$$

$$G_{ijk} \sim \text{DP}(\alpha_{G,ik}, G_{0,jk}) \tag{6}$$

where DP($\alpha_{G,ik}$, $G_{0,jk}$) stands for Dirichlet process with expected probability measure $G_{0,jk}$ having precision parameter $\alpha_{G,ik}$, with

$$\log(\alpha_{G,ik}) = \mu_G + {\beta_G}^T E_{ik}, \tag{7}$$

where $E_{ik}$ is a *d*-dimensional vector of continuous environmental variable for the *i*-th individual in the *k*-th group, $\beta_G$ is a *d*-dimensional vector of regression coefficients, and $\mu_G$ is the intercept term. The model can be easily extended to include categorical environmental variables along with the continuous ones.

## 2.3. Hierarchical Dirichlet processes to model the dependence between the genes and case-control status

We further assume that for *k* = 0, 1,

$$G_{0,jk} \overset{iid}{\sim} DP(\alpha_{G_0k}, H_k); j = 1, \ldots, J, \tag{8}$$

$$\text{where } \log(\alpha_{G_0,k}) = \mu_{G_0} + {\beta_{G_0}}^T \overline{E}_k, \tag{9}$$

$$\text{with } \overline{E}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} E_{ik} \tag{10}$$

We postulate the last level of hierarchy as

$$H_k \overset{iid}{\sim} \mathrm{DP}(\alpha_H \widetilde{H});\ k = 0,1 \tag{11}$$

where $\log(\alpha_H) = \mu_H + {\beta_H}^T \bar{\bar{E}}$, (12)

with $\bar{\bar{E}} = \frac{\bar{E}_0 + \bar{E}_1}{2}$ (13)

We specify the base probability measure $\widetilde{H}$ as follows: for $m$ = 1, . . . , $M$, $i$ = 1, . . . , $N_k$, $k$ = 0, 1, and $r$ = 1, . . . , $L$,

$$p_{mijkr} \overset{iid}{\sim} \text{Beta } (\nu_1, \nu_2), \tag{14}$$

Under $\widetilde{H}$, where $\nu_1, \nu_2 > 0$.

Note that our model consists of one more level of hierarchy of Dirichlet processes than considered in the applications of Teh *et al*. (2006), who introduce hierarchical Dirichlet processes (HDP). For detailed discussion on the dependence structure induced by our hdp-based model see Section 3 of Bhattacharya and Bhattacharya (2020 b).

## 3. Detection of the roles of environment, genes and their interactions with respect to our hdp based model

### 3.1. Formulation of the tests and interpretation of their results

To test if genes have any effect on case-control, we formulate the following hypotheses:

$$H_{01}\colon h_{0j} = h_{1j};\ j = 1, \ldots, J, \tag{15}$$

versus

$$H_{11} : \text{not } H_{01}, \tag{16}$$

where $h_{0j}(.) = \frac{1}{M}\sum_{m=1}^{M}\prod_{r=1}^{L_j} f(.\,|{p_{mi_0jk=0}}^r)$ (17)

$$h_{1j}(.) = \frac{1}{M}\sum_{m=1}^{M}\prod_{r=1}^{L_j} f(.\,|{p_{mi_1jk=1}}^r) \tag{18}$$

In the above, for $k$ = 0, 1, $i_{\mathrm{k}}$ is the index such that $P_{Mi_kjk} = \{p_{1i_kjk}, p_{2i_kjk}, \ldots, p_{Mi_kjk}\}$ is an appropriate measure of central tendency (see Section 4.2.1 of Bhattacharya and Bhattacharya (2020 b)) of $\{P_{Mijk} = \{p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk}\}$; $i$ = 1, . . . , $N_k$.

#### 3.1.1. Bayesian test for the significance of the environmental variables

To check if the environmental variables are significant, we shall test the following:

for $l$ = 1, . . . , $d$,

$$H_{02l}\colon \beta_{Gl} = 0 \text{ versus } H_{12l}\colon \beta_{Gl} \neq 0, \tag{19}$$

$$H_{03l}\colon \beta_{G_0l} = 0 \text{ versus } H_{12l}\colon \beta_{G_0l} \neq 0, \tag{20}$$

and $H_{04l}\colon \beta_{Hl} = 0$ versus $H_{14l}\colon \beta_{Hl} \neq 0$. (21)

### 3.1.2. Bayesian test for significance of gene-gene interaction

In order to test for gene-gene interaction, it is necessary to first reasonably define a measure of gene-gene interaction influenced by environmental variables.

For our purpose, we first define

$$\overline{p}_{mijk} = \frac{\sum_{r=1}^{L_j} p_{mijkr}}{L_j} \tag{22}$$

With the above definition, for subject *i* belonging to case-control group *k*, we consider the following covariance

$$C(i, j_1, j_2, k) = cov(\text{logit}(\overline{p}_{z_{ij_1k}ij_1k}), \text{logit}(\overline{p}_{z_{ij_2k}ij_2k}), \tag{23}$$

as quantification of gene-gene dependence that accounts for population memberships of subject *i* with respect to genes $j_1$ and $j_2$, through $z_{ij_1k}$ and $z_{ij_2k}$. While implementing our model using our parallelised MCMC methodology, we simulate $C(i, j_1, j_2, k)$ at each iteration by generating $\{p_{mijkr}: r = 1, \ldots, L_j\}$ as many times as required from the respective full conditionals holding the remaining parameters fixed, and then compute the empirical covariance corresponding to (23) using the generated iid samples conditionally on the remaining parameters to approximate (23).

**Formulation of the Bayesian tests for gene-gene interactions**

To test for subject-wise gene-gene interaction, we consider the following tests:

for $i = 1, \ldots, N_k$, $k = 0, 1$, and for$j_1, j_2 \in \{1, \ldots, J\}$,

$$H_{05ij_1j_2k}: C(i, j_1, j_2, k) = 0 \text{ versus } H_{15ij_1j_2k}: C(i, j_1, j_2, k) \neq 0. \tag{24}$$

For some appropriate divergence measure, *d*, between two distributions, if

$\underset{1 \le j \le J}{max}$ d($h_{0j}$, $h_{1j}$), is significantly small with high posterior probability, then $H_{01}$ is to be accepted. In case $H_{01}$ is rejected, we go forward to perform various tests related to gene-gene and gene-environment interactions, enlisted in Sections 3.1.1. and 3.1.2. above. For interpretations and detailed discussion on the tests see Section 4.1.4 of Bhattacharya and Bhattacharya (2020 b).

## 3.2. Methodologies for implementing the Bayesian tests

### 3.2.1. Hypothesis testing based on clustering modes

Here we exploit the concept of "central" clustering introduced by Mukhopadhyay *et al.* (2011). Briefly, central clustering may be interpreted as a suitable measure of central tendency of a set of clusterings.

For $k = 0, 1$, let $i_k$ denote the index of the central clusterings of $P_{Mijk} = \{p_{1ijk}, p_{2ijk}, \ldots, p_{Mijk}\}$, $i = 1, \ldots, N_k$. We then study the divergence between the two clusterings of $P_{Mi_0jk=0} = \{p_{1i_0jk=0}, p_{2i_0jk=0}, \ldots, p_{Mi_0jk=0}\}$and $P_{Mi_1jk=1} = \{p_{1i_1jk=1}, p_{2i_1jk=1}, \ldots, p_{Mi_1jk=1}\}$ , for $j = 1, \ldots, J$.

Significantly large clustering distance between $P_{Mjk=0}$ and $P_{Mjk=1}$indicates rejection of $H_0$, but insignificant clustering distance does not necessarily provide strong evidence in favour of the null. In this regard, Bhattacharya and Bhattacharya (2018) (see also Bhattacharya and Bhattacharya (2020 a)) argue that the Euclidean distance is an appropriate candidate to be tested for significance before arriving at the final conclusion. Briefly, we first compute the averages $\overline{p}_{mijk}$ = $\sum_{r=1}^{L_j} p_{\mathrm{mi}jkr}/L_j$, then consider their logit transformations logit($\overline{p}_{mijk}$) = log $\{\overline{p}_{mijk}/(1-\overline{p}_{mijk})\}$. Then, we compute the Euclidean distance between the vectors

logit($\overline{P}_{Mi_0jk=0}$) = {logit($\overline{p}_{1i_0jk=0}$), logit($\overline{p}_{2i_0jk=0}$),…, logit($\overline{p}_{\mathrm{M}i_0jk=0}$)} and

logit($\overline{P}_{Mi_1jk=1}$) = {logit($\overline{p}_{1i_1jk=1}$), logit($\overline{p}_{2i_1jk=1}$),…, logit($\overline{p}_{\mathrm{M}i_1jk=1}$)}

We denote the Euclidean distance associated with the $j$-th gene by

$$d_{E,j} = d_{E,j}(\mathrm{logit}(\overline{P}_{Mi_0jk=0}), \mathrm{logit}(\overline{P}_{Mi_1jk=1}))$$

and denote $\underset{1\leq j\leq J}{max}$ $d_{E,j}$ by $d^*{}_{E,j}$.

### 3.2.2. Formal Bayesian hypothesis testing procedure integrating the above developments

In our problem, we need to test the following for reasonably small choices of $\varepsilon$'s:

$H_{0,d^*}$: $d^* < \varepsilon_{d^*}$ versus $H_{1,d^*}$: $d^* \geq \varepsilon_{d^*}$; (25)

$H_{0,d^*{}_E}$: $d^*{}_E < \varepsilon_{d^*{}_E}$ versus $H_{1,d^*{}_E}$: $d^*{}_E \geq \varepsilon_{d^*{}_E}$; (26)

For l = 1,2,...,$d$

$H_{0,\beta_{Gl}}$: $|\beta_{Gl}| < \varepsilon_{Gl}$ versus $H_{1,\beta_{Gl}}$: $|\beta_{Gl}| \geq \varepsilon_{Gl}$ (27)

$H_{0,\beta_{Gl}}$: $|\beta_{G_0l}| < \varepsilon_{G_0l}$ versus $H_{1,\beta_{G_0l}}$: $|\beta_{G_0l}| \geq \varepsilon_{G_0l}$ (28)

$H_{0,\beta_{Hl}}$: $|\beta_{\mathrm{H}l}| < \varepsilon_{Hl}$ versus $H_{1,\beta_{\mathrm{H}l}}$: $|\beta_{\mathrm{H}l}| \geq \varepsilon_{Hl}$ (29)

and, for $i$ = 1, . . . , $N_k$, $k$ = 0, 1, $j_1, j_2 \in \{1, . . . , J\}$,

$H_{0,\mathrm{C(i},j_1,j_2,\mathrm{k)}}$ : $|\mathrm{C(i},j_1,j_2,\mathrm{k)}| < \varepsilon_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}$ versus $H_{1,\beta_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}}$ : $|\beta_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}| \geq \varepsilon_{\mathrm{C(i},j_1,j_2,\mathrm{k)}}$ (30)

If $H_0$ is rejected in (25) or in (26), we could also test if the $j$-th gene is influential by testing, for $j$ = 1, . . . , $J$, $H_{0,\hat{d}_j}$: $\hat{d}_j < \varepsilon_{\hat{d}_j}$ versus $H_{1,\hat{d}_j}$: $\hat{d}_j \geq \varepsilon_{\hat{d}_j}$, where $\hat{d}_j = \hat{d}(\overline{P}_{Mi_0jk=0}, \overline{P}_{Mi_1jk=1})$; we could also test $H_{0,d_{E,j}}$: $d_{E,j} < \varepsilon_{d_{E,j}}$ versus $H_{1,d_{E,j}}$: $d_{E,j} \geq \varepsilon_{d_{E,j}}$. For the null model and choice of $\varepsilon$ see Bhattacharya and Bhattacharya (2020 b).

## 4. Simulation studies

For simulation studies, we first generate realistic biological data for stratified population with known gene-environment interaction from the GENS2 software of Pinelli *et al*. (2012). To this data, we then apply our model and methodologies in an effort to detect gene-environment interaction effects that are present in the data. We consider simulation studies under 5 different true model set-ups: (a) presence of gene-gene and gene-environment interaction; (b) absence of genetic or gene-environmental interaction effect; (c) absence of genetic and gene-gene interaction effects but presence of environmental effect; (d) presence of genetic and gene-gene interaction effects but absence of environmental effect; and (e) independent and additive genetic and environmental effects.

The details of our simulation experiments are provided in Section S-3 of the supplement of Bhattacharya and Bhattacharya (2020 b). Here we briefly summarize the results of our experiments. In case (a), we correctly obtained clear significance of the influence of genetic effects. Also, $\beta_{\mathrm{H}l}$ turned out to be very significant, demonstrating significant overall impact of the environmental variable on gene-gene interaction. The posteriors of the number of sub-populations gave high probabilities to the correct number of sub populations in all the 5 simulation experiments. Quite importantly, we demonstrate in cases (a), (d) and (e) where the genes are relevant, that our HDP model can detect disease predisposing loci (DPL) with more precision compared to the matrix-normal-inverse-Wishart model for gene-environment interactions proposed in Bhattacharya and Bhattacharya (2020A). In case (b) using our ideas in conjunction with significance testing in a simple logistic regression framework, we are correctly able to conclude that the genetic or gene-environmental effects are insignificant.

## 5. Application of hdp based ideas to a real, case-control dataset on myocardial infarction

We now consider application of our model and methods to a case-control dataset on early-onset of myocardial infarction (MI) from MI Gen study, obtained from the dbGaP database http://www.ncbi.nlm.nih.gov/gap.

### 5.1. Data description

The MI Gen data obtained from dbGaP consists of observations on presence/absence of minor alleles at 727478 SNP markers associated with 22 autosomes and the sex chromosomes of 2967 cases of early-onset myocardial infarction, 3075 age and sex matched controls. The average age at the time of MI was 41 years among the male cases and 47 years among the female cases. The data broadly represents a mixture of four sub-populations: Caucasian, Han Chinese, Japanese and Yoruban. Using the Ensembl human genome database (http://www.ensembl.org/) we could categorize 446765 markers out of 727478 with respect to 37233 genes. As in Bhattacharya and Bhattacharya (2020 a) we considered 32 genes covering 1251 loci, for 200 individuals. These 1251 loci include 33 SNPs that are believed to be associated with MI and also those that are believed to be associated with different cardiovascular end points like LDL cholesterol, smoking, blood pressure, body mass, *etc*. Other than the 33 SNPs, the remaining 1218 SNPs are not known to be associated with the disease (see Bhattacharya and Bhattacharya (2020 a)) for the details and the relevant references.

### 5.2. Remarks on model implementation

Our parallel MCMC algorithm detailed in Section S-2 of the supplement of Bhattacharya and Bhattacharya (2020 b), takes about 7 days to generate 30,000 iterations on our VMware consisting of 1 TB RAM, 60 double-threaded, 64-bit physical cores, each running at 2.5 GHz; 50 such cores were available to us. We discard the first 10, 000 iterations as burn-in, using the subsequent 20,000 iterations for our Bayesian inference. Convergence is studied using informal convergence diagnostics such as trace plots. Some instances are provided in Section S-3 of the supplement of Bhattacharya and Bhattacharya (2020 b).

## 5.3. Results of the real data analysis

### 5.3.1. Effect of the sex variable

We obtain $P(|\beta_{Gl}| < \varepsilon_{Gl}$ |Data) $\approx$ 0, $P(\left|\beta_{G_0 l}\right| < \varepsilon_{G_0 l}$|Data) $\approx$ 0 and $P(|\beta_{\mathrm{H}l}| < \varepsilon_{Hl}$|Data) $\approx$ 1. In other words, although $\overline{\overline{E}}$ (here $E$ being the sex variable) is insignificant, both $E_{ik}$ and $\overline{E}_k$ are very significant. Thus, in this study, sex seems to play an important role in influencing gene-gene interaction.

### 5.3.2. Roles of individual genes

With the clustering metric we obtained $P(d^* < \varepsilon_{d^*}) \approx 0.030$ while that with the Euclidean distance we obtained $P({d^*}_E < \varepsilon_{{d^*}_E}$|Data) $\approx 0.540$. That is, the maximum of the gene-wise clustering metrics turns out to be significant, while the maximum of the gene-wise Euclidean metrics is seen to be insignificant. None of the individual genes turned out to be significant, for either the clustering metric or the Euclidean metric. The posterior probabilities of the null hypotheses (of no significant genetic influence) with respect to the clustering metric is shown in Figure 1.

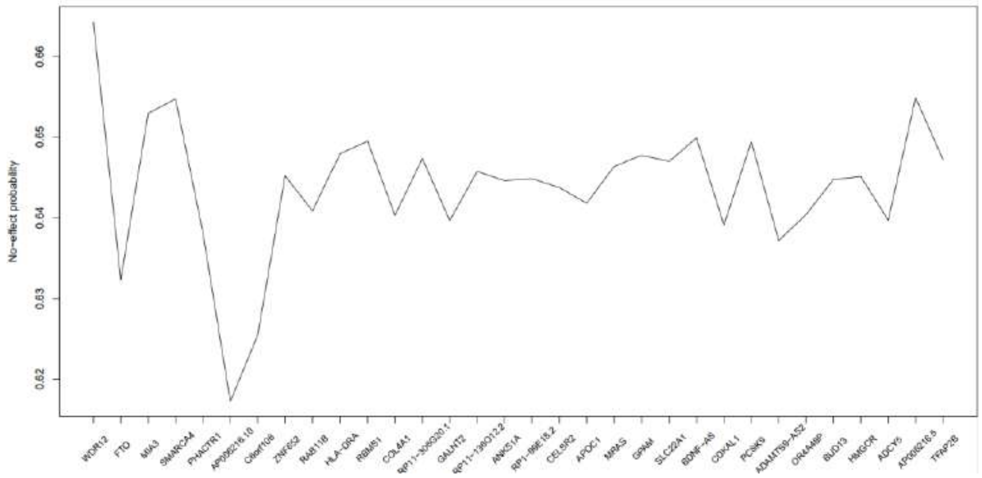
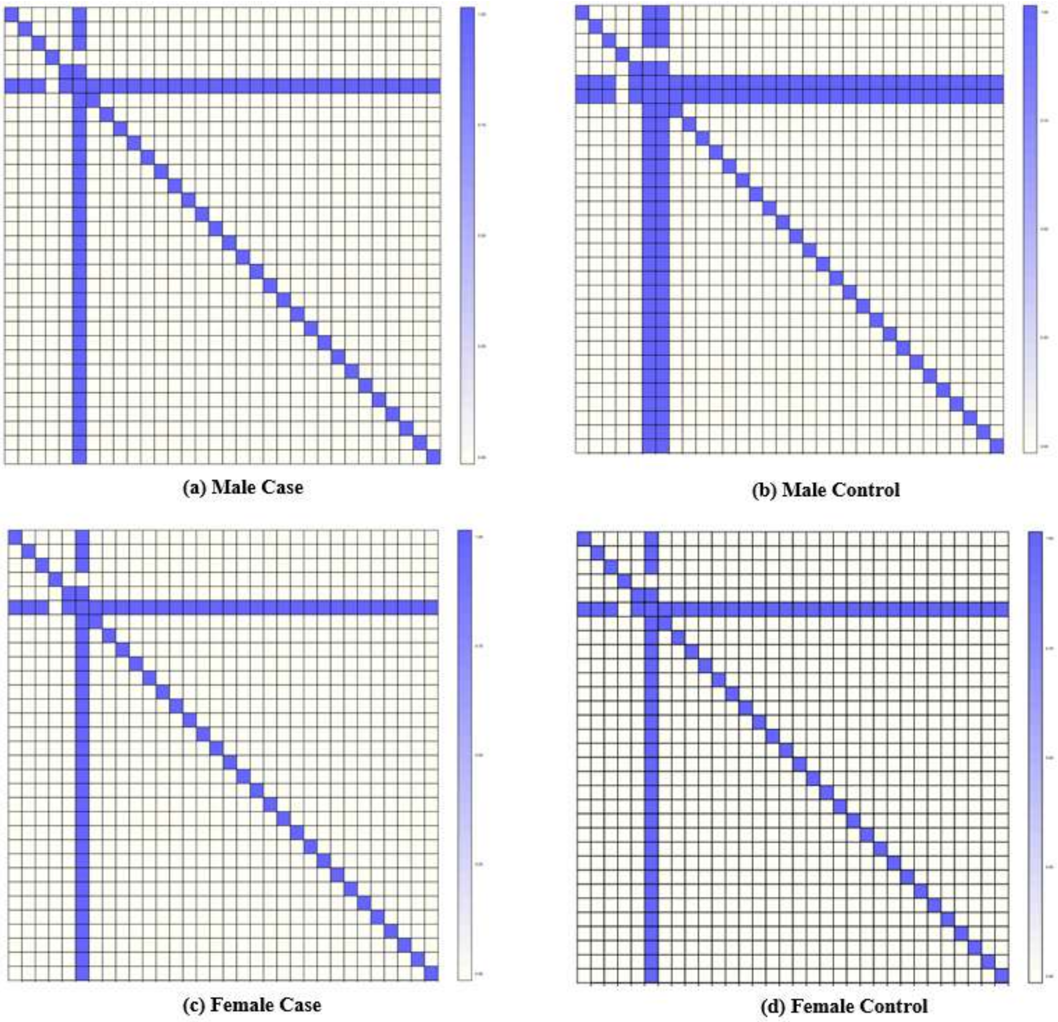


**Figure 1: Posterior probability of no genetic effect with respect to clustering metric**

### 5.3.3. Gene-gene interactions

Figures 2(a) to 2 (d) show the typical gene-gene correlations representative of cases and controls in males and females. The colour intensities correspond to the absolute values of the correlations. Although the correlations are small in all the subjects, the tests of hypotheses reveal some interesting structures. Our tests indicate that for most of the subjects, at least one of the genes AP006216.10 and C6orf106 interact with every other gene. The subjects, for whom no significant genetic interactions involving AP006216.10 and C6orf106 were detected, turned out to be male cases, indicating that the lack of genetic interaction in these males might be associated with MI. On the other hand, the interactions of the genes AP006216.10 and C6orf106 with all the genes seemed to reduce the risk of the disease for the other subjects.

Thus, following this study, the gene-gene interactions need to be investigated further for their possible beneficial effect on the subjects against MI.



(a) Male Case

(b) Male Control

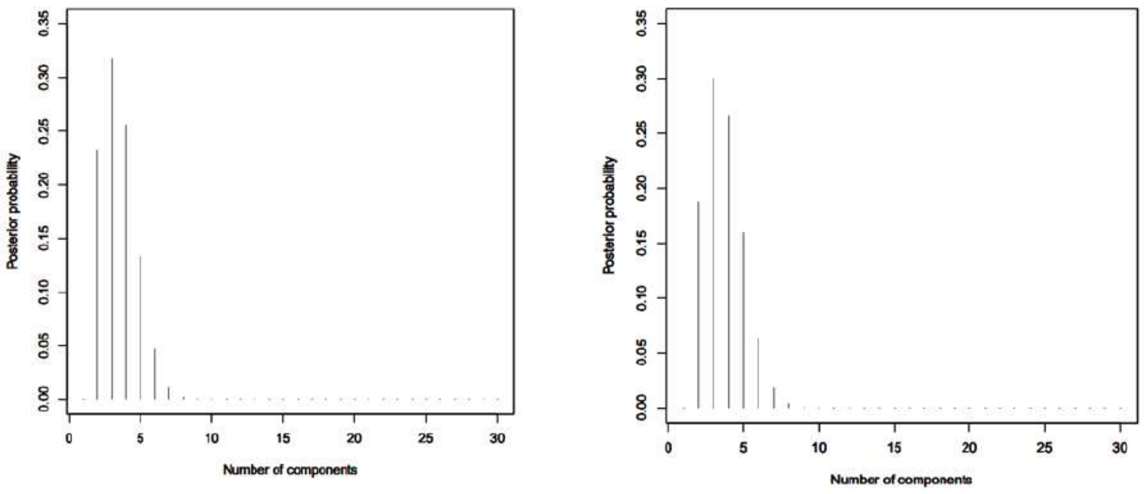
(c) Female Case

(d) Female Control

**Figure 2: Typical gene-gene posterior correlation plot for male cases and controls and female cases and controls**

### 5.3.4. Posteriors of the number of sub-populations

Figure 3 shows the posteriors of the number of sub-populations for the males and females associated with respectively. Observe that the posteriors are quite similar, with the mode at 3 and 4 components receiving the next highest probability. Thus, the 4 sub-populations, irrespective of sex, are well supported by our model.



(a) Males

(b) Females

**Figure 3: Typical posteriors of the number of components for males and females**

## 6. Summary and conclusion

In this paper, we have proposed a novel Bayesian nonparametric gene-gene and gene-environment interaction model based on hierarchies of Dirichlet processes. This model is a significant improvement over the existing work in this area, in the sense of much clear interpretability and accounting for subject-specific gene-gene interactions. We propose a novel parallel MCMC algorithm to implement our model (Sections S-1and S-2 of the supplementary material of Bhattacharya and Bhattacharya (2020 b)), that combines powerful technology with conditionally independent structures inherent within our HDP based model and efficient TMCMC methods. Applications of our ideas to biologically realistic datasets generated under 5 different setups characterized by different combinations and structures associated with gene-gene and gene environment interactions demonstrated encouraging performance of our model and methods. Our analysis of the real MI dataset yielded results that are broadly in agreement with the previous works on the same dataset. For example, in accordance with Bhattacharya and Bhattacharya (2020A) (see also Lucas *et al.* (2012)) we obtained strong impact of the sex variable, weak gene-gene correlations but no significant effect of the individual genes. But special mention must be reserved for our original finding that two genes, AP006216.10 and C6orf106, tend to fight the disease by their positive interaction with the remaining genes. Another interesting discovery that emerged from our analyses is that only in male cases all the gene-gene interactions were insignificant. These two findings seem to confirm the general belief that as compared to females, males are more vulnerable to heart attack.

## References

Ahn, J., Mukherjee, B., Gruber, S. B., and Ghosh, M. (2013). Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *The Annals of Applied Statistics*, **7**, 543–569.

Bhattacharya, D. and Bhattacharya, S. (2020 a). Effects of gene-environment and gene-gene interactions in case-control studies: A novel Bayesian semiparametric approach. *Brazilian Journal of Probability and Statistics*, **34**, 71–89. Also available at “https://arxiv.org/abs/1601.03519”.

Bhattacharya, D. and Bhattacharya, S. (2020 b). A non-gaussian, nonparametric structure for gene-gene and gene-environment interactions in case-control studies based on hierarchies of Dirichlet processes. arXiv:1704.07349v2 [stat.AP].

De Iorio, M., Elliott, L. T., Favaro, S., Adhikari, K., and Teh, Y. W. (2015). Modeling population structure under hierarchical Dirichlet processes. Available at https://arxiv.org/abs/1503.08278.

Hunter, D. J. (2005). Gene environment interactions in human diseases. *Nature Publishing Group*, **6**, 287–298.

Lucas, G., Lluis-Ganella, C., Subirana, I., Masameh, M. D., and Gonzalez, J. R. (2012). Hypothesis-based analysis of gene-gene interaction and risk of myocardial infraction. *Plos One*, **7**, 1–8.

Majumdar, A., Bhattacharya, S., Basu, A., and Ghosh, S. (2013). A novel Bayesian semiparametric algorithm for inferring population structure and adjusting for case-control association tests. *Biometrics*, **69**, 164–173.

Mukhopadhyay, S., Bhattacharya, S., and Dihidar, K. (2011). On Bayesian "Central clustering": Application to landscape classification of western ghats. *Annals of Applied Statistics*, **5**, 1948–1977.

Pinelli, M., Scala, G., Amato, R., Cocozza, S., and Miele, G. (2012). Simulating gene-gene and gene-environment interactions in complex diseases: gene-environment interaction simulator 2. *BMC Bioinformatics*, **13**; https://doi.org/10.1186/1471-2105-13-132.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Wen, X., and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogenous subgroups: from meta-analyses to gene-environment Interactions. *Annals of Applied Statistics*, **8**, 176–203.

Yi, N., Kaklamani, V. G. and Pasche, B. (2011). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Annals of Human Genetics*, **75**, 90–104.

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 43–50

# A Comparison of CAN and UMVU Estimators in Inliers-Prone Distributions

**K. Muralidharan**
*Department of Statistics, Faculty of Science*
*The Maharaja Sayajirao University of Baroda, Vadodara, 390 002 India*

Received: 18 June 2023; Revised: 30 June 2023; Accepted: 05 July 2023

## Abstract

Among various classical estimations procedures, a relatively better estimation is provided by Consistent Asymptotic Estimators (CAN) method. The method of CAN provides estimators for parametric functions of regular and non-regular or degenerate families of distributions. In this article, we present CAN estimators for parametric functions of inlier-prone (a case of degenerate) distribution models. The estimates are also compared numerically.

*Key words*: Consistency; Degenerate family of distributions; Inlier-prone models; Minimum variance unbiased estimators.

## 1. Introduction

In statistical estimation theory, one starts with the data $(x_1, x_2, \dots, x_n)$ of a random variable $X$, which are assumed to be independent and identically distributed with a common probability distribution $f(x, \theta)$ characterized by an unknown population parameter $\theta \in \Omega$, where $\theta$ can be real-valued scalar or vector. The objective is to propose a best inference for $\theta$ or $\psi(\theta)$ which satisfies good statistical properties. If the probability model is uniquely defined, one can suggest suitable estimators for the parameter or parametric functions explicitly. Let $T = T(x_1, x_2, \dots, x_n)$ be an estimator of $\theta$ based on the observed sample values $(x_1, x_2, \dots, x_n)$. By using the techniques of transformation or form the basic principles of distribution theory, one could, at least theoretically, obtain the sampling distribution of the estimator $T$ and thus begin the inference of the population parameter $\theta$.

There are many criteria and procedures available for deciding the best estimator for $\theta$ or $\psi(\theta)$ in Statistics literature. The best estimator in a statistical sense is decided based on a comparison of the variance or mean square error (MSE) of the estimator of one method over the other. For this we assume that $T$, a real-valued statistic, is to be used as an estimator of real parameter $\theta$ based on a random sample of size $n$ from $\{f(x, \theta), \theta \in \Omega\}$, $\Omega \subset R^1$. One of the criteria based on a large sample size is the consistency of an estimator.

**Definition 1:** An estimator $T_n$ is said to be consistent for $\theta$ if $T_n \to \theta$ for each $\theta \in \Omega$ in probability and the convergence in probability is taken under the distribution indexed by $\theta$.

Corresponding Author: K. Muralidharan
Email: lmv_murali@yahoo.com
*This paper is based on the B.K. Kale Memorial Endowment Lecture "A comparison of CAN and UMVU Estimators in inliers-prone distributions" delivered by the author on 15 February 2023 during the conference.*

A very important property of a consistent estimator is the invariance under continuous transformation, a property not enjoyed by an unbiased estimator. Thus, if $\psi(\theta)$ is a continuous function and if $T$ is consistent for $\theta$, then the invariance property says that $\psi(T)$ is consistent for $\psi(\theta)$. Because of the invariance property of consistent estimators, for all practical purposes one need to consider consistent estimators of $\theta$ only for further study of the estimators. The invariance property can be extended to the case of vector valued $T$ and $\theta$ as follows:

**Definition 2:** Let $T$ be jointly consistent for $\theta$ and let $\psi$ be $k$-dimensional continuous functions from $\Omega$ to $R^k$, then $\psi(T)$ is jointly consistent for $\psi(\theta)$ (Kale and Muralidharan, 2015).

To choose between consistent estimators one can compare the MSE's of the estimators, where MSE is defined as $MSE(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$, where $\hat{\theta}$=$T(x)$ is the unbiased estimate of $\theta$. For instance, if $T_1$ and $T_2$ are both consistent for $\theta$ then we would prefer $T_1$ to $T_2$ if $MSE(T_1) \leq MSE(T_2)$, $\forall\theta\epsilon\Omega$. This comparison generally results into the comparison of the sample sizes of the two estimators. Thus, if $T_1$ is preferred over $T_2$ then by Tchebychev inequality it follows that $P[|T_1 - \theta| < \epsilon]$ converges to unity faster than $P[|T_2 - \theta| < \varepsilon] \to 1$ as $n \to \infty, \forall\theta\epsilon\Omega$ and $\varepsilon > 0$. For large $n$, it is easy to show that $a_n(T - \theta) \to N(0, \sigma_T^2(\theta)$ or $T \sim AN\left(0, \frac{\sigma_T^2(\theta)}{a_n^2}\right)$, where $a_n$ is the blow-up factor (Kale and Muralidharnan, 2015). Such an estimator is called Consistent Asymptotic Normal or CAN estimator. As discussed above, if $\psi(\theta)$ is a continues differentiable function then according to invariance property of consistent estimators the CAN estimator for $\psi(\theta)$ is defined as follows:

**Definition 3:** Let $T$ be CAN for $\theta$ so that $T \sim AN\left(\theta, \frac{\sigma_T^2(\theta)}{a_n^2}\right)$ and let $\psi$ be differentiable functions such that $\frac{d\psi}{d\theta}$ is continuous and nonvanishing then $\psi(T)$ is CAN for $\psi(\theta)$ and $\psi(T)\sim$ $AN\left(\psi(\theta), \sigma_T^2(\theta)\left(\frac{d\psi}{d\theta}\right)^2 / a_n^2\right)$ (Kale and Muralidharan, 2015).

We now propose CAN estimators for parametric functions by considering a family of distributions which are degenerated at some random point. This degeneracy may be due to the occurrence of *instantaneous* or *early* failures together known as inliers are usually seen in life testing experiments. In the instantaneous failure cases, the random variable will have discrete probability mass at the origin (that is lifetime will be zero) and some positive lifetimes, and in the early failure case the failure times may be small in relation to other lifetimes. For modeling positive lifetimes, we have used exponential distribution, as it has been widely used as a model in areas ranging from studies on the lifetimes of manufactured items to research involving survival or remission times in chronic diseases. The exponential distribution has the pdf

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \ x \geq 0 \tag{1}$$

The maximum likelihood estimator of $\theta$ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The desirable properties of $\hat{\theta}$ are numerous. In particular $\hat{\theta}$ is exactly distributed as $\left(\frac{\theta}{2n}\right)\chi^2_{(2n)}$ and it is a sufficient, efficient, and minimum variance estimator of $\theta$.

The article is organized as follows: The model presentations along with some distributional results are given in Section 2. Along with the CAN estimation, we also propose uniformly minimum variance unbiased estimate (UMVUE) for various parametric functions in Section 3. The numerical illustration is presented in the last section.

## 2. Inliers-prone model

If the underlying distribution is exponential as given in (1.1), then the inliers-prone model with instantaneous failures is shown as

$$g(x;p,\theta) = \begin{cases} 1-p, & x = 0 \\ \frac{p}{\theta}e^{-\frac{x}{\theta}}, & x > 0 \end{cases} \tag{2}$$

Let $X_1, X_2, \dots, X_n$ be a random sample from (2) then the pdf of $X_i$ is

$$g(x_i;p,\theta) = \begin{cases} (1-p)^{I(x_i)}\left(\frac{p}{\theta}e^{-\frac{x}{\theta}}\right)^{1-I(x_i)} & x_i \geq 0, 0 < p \leq 1, \theta > 0, i = 1,2,\dots,n \\ 0, & o.w. \end{cases}$$

where,

$$I(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{o.w.} \end{cases} \tag{3}$$

Aitchison (1955) had proposed various unbiased functions for parametric function in (3). Kale and Muralidharan (2000) were the first authors to introduce the term inliers in connection with the estimation of $(p, \theta)$ of early failure model with modified failure time distribution (FTD) being (1) with mean $\theta$. A similar problem was attempted by Lai et al. (2007), wherein they have defined nearly instantaneous through the sample configurations, considering Weibull as the underlying FTD. For a detailed review of inliers prone models and their inferences, refer to Muralidharan (2010).

If $p = P(x > 0)$ and further, if we denote $\sum_{i=1}^{n} I(x_i) = n - r$, where $r$ is number of positive observations, then the joint pdf is given by

$$g(\underline{x};p,\theta) = \begin{cases} (1-p)^{n-r}\left(\frac{p}{\theta}\right)^{r} e^{-\frac{1}{\theta}\sum_{i=1}^{n}(1-I(x_i))x_i}, x_i \geq 0, r = 0,1,\dots,n \\ 0, & o.w. \end{cases} \tag{4}$$

The following results are now obvious:

**Result 1:** The joint density function given in (4) is a two-parameter exponential family of distribution.

**Result 2:** $\left(\sum I(x_i), \sum\left(1 - I(x_i)\right)x_i\right)$ are jointly sufficient for $p$ and $\theta$.

**Result 3:** The MLE of $p$ and $\theta$ are respectively given by $\hat{p}_{MLE} = \frac{r}{n}$ and $\hat{\theta}_{MLE} = \frac{1}{r}\sum_{x_i>0} x_i$.

**Result 4:** $\left(\hat{p}_{MLE}, \hat{\theta}_{MLE}\right)' \sim AN^{(2)}\left[(p,\theta)', \ diag\left(\frac{p(1-p)}{n}, \frac{\theta^2}{np}\right)\right]$.

**Result 5:** The parameters $p$ and $\theta$ are orthogonal.

**Result 6:** The true reliability or survival function for the model at time $t$ is given by

$$S(t) = pe^{-\frac{t}{\theta}}, t > 0, \theta > 0$$

**Result 7:** $g_{Z|R}(z;\theta|r) = \begin{cases} \frac{e^{-\frac{z}{\theta}} z^{r-1}}{\Gamma r \ \theta^r}, & z > 0, r > 0 \\ 1, & z = 0, r = 0 \end{cases}$,

where $z = \sum_{i=1}^{n}[1 - I(x_i)]x_i \left(= \sum_{x_i>0} x_i\right)$.

## 3. UMVUE and CAN estimators

It is observed that, obtaining conditional distribution given the sufficient statistics is bit difficult in the above model. Therefore, we use exponential family approach to study the distributional properties.

The equation (4) is written as

$$g(x;p,\theta) = \frac{\left[e^{-\frac{1}{\theta}}\right]^{(1-I(x))d(x)} \left[\frac{\theta(1-p)}{p}\right]^{I(x)}}{\left(\frac{\theta}{p}\right)}$$

$$=[a(x)]^{(1-I(x))}[h(\theta)]^{(1-I(x))d(x)}\left[\frac{g(\theta)(1-p)}{p}\right]^{I(x)}\left(\frac{g(\theta)}{p}\right)^{-1} \quad (5)$$

where $a(x) = 1$, $h(\theta) = e^{-\frac{1}{\theta}}$, $d(x) = x$, $g(\theta) = \theta$. The density in (5) is so obtained is defined with respect to measure $\mu(x)$ which is the sum of Lebesgue measure over $(0, \infty)$ and a singular measure at $\{0\}$, is a well-known form of two parameter exponential family with natural parameters $(\eta_1, \eta_2) = \left(log\left(\frac{\theta(1-p)}{p}\right), log\left(e^{-\frac{1}{\theta}}\right)\right)$ generated by the underlying indexing parameters $(p,\theta)$. Here $\left(I(x), \left(1 - \ I(x)\right)x\right)$ is jointly minimal sufficient for $(p,\theta)$ as $I(x)$ and $(1 - I(x))x$ do not satisfy any linear restriction. Hence the natural parameter space is convex set in $E_2$ containing a two-dimensional rectangle making (5) a full rank family. The statistic $\left(I(x), \left(1 - \ \text{I(x)}\right)x\right)$ is thus complete (Lehmann and Casella, 1998, p 42). Kale and Muralidharan (2000) considered the above mixture and obtained optimal estimating equation for $\theta$ ignoring $p$ in the case of exponential failure time distribution.

Further, if we denote $z = \sum_{i=1}^{n}[1 - I(x_i)]x_i \left(= \sum_{x_{i>0}} x_i\right)$, then the joint density function can be expressed as

$$g(\underline{x}; p, \theta) = \binom{n}{r}(1-p)^{n-r}\left(\frac{p}{\theta}\right)^{r} e^{-\frac{z}{\theta}} \tag{6}$$

Hence $(n - R, Z)$ are jointly complete sufficient for $(p, \theta)$. Also, the variable $(Z|R = r, r > 0)$ is distributed as a Gamma random variable with parameter $(r, \theta\ )$. Since, $n - R$ is binomial which is same as that of $R$ with parameter $(n, p)$, the joint distribution of $(n - R, Z)$ is

$$\begin{aligned} g(z, n-r; p, \theta) &= P(n - R = n - r)\ g(z; \theta|n - r) \\ &= P(R = r)\ g(z; \theta|r) \\ &= \binom{n}{r}(1-p)^{n-r}p^{r}\frac{1}{\Gamma r\ \theta^{r}} z^{r-1}e^{-\frac{z}{\theta}} \end{aligned}$$

$$= \begin{cases} (1-p)^n, & z = 0;\ r = 0 \\ \binom{n}{r}\dfrac{z^r}{\Gamma r}e^{-\frac{z}{\theta}}\left(\dfrac{\theta(1-p)}{p}\right)^{n-r}\left(\dfrac{\theta}{p}\right)^{-n} & , z > 0;\ r > 0 \end{cases}$$

$$= \begin{cases} (1-p)^n, & z = 0;\ r = 0 \\ B(z, r, n)[h(\theta)]^z\left[\frac{g(\theta)(1-p)}{p}\right]^{n-r}\left(\frac{g(\theta)}{p}\right)^{-n}, & z > 0;\ r > 0 \end{cases} \tag{7}$$

where

$$B(z, r, n) = \begin{cases} 1, & z = 0; r = 0 \\ \binom{n}{r}B(z|r), & z > 0; r > 0 \end{cases} \tag{8}$$

is such that $(1-p)^n + \sum_{r=1}^{n}\int_{z>0}\binom{n}{r}B(z|r)\left[e^{-\frac{1}{\theta}}\right]^{z}\left(\frac{\theta(1-p)}{p}\right)^{n-r}\left(\frac{\theta}{p}\right)^{-n}dz = 1$ and $B(z|r) = \frac{z^{r-1}}{\Gamma r}$. Following Roy and Mitra (1957) and Jani and Singh (1995), it is possible to obtain the uniformly minimum variance unbiased estimates (UMVUE) for some parametric functions. Note that, the UMVUE's of parametric function $\phi(p, \theta)$exits if and only if $\phi(p, \theta)$ can be expressed in the form

$$\phi(p, \theta) = \alpha(0,0, n)(1-p)^n + \sum_{r=1}^{n}\int_{z>0}\frac{\alpha(z,r,n)e^{-\frac{z}{\theta}}\left(\frac{\theta(1-p)}{p}\right)^{n-r}}{\left[\frac{\theta}{p}\right]^{n}}dz.$$

Below we consider some estimates for the parametric functions:

**Result 8:** For $m \le n$, the UMVUE of $(1-p)^m$ is $G_m(Z, R, n)$ as given by

$$G_m(z, r, n) = \begin{cases} \dfrac{\binom{n-m}{r}}{\binom{n}{r}}, r = 0,1, \dots, n - m \\ 0, \qquad o.w. \end{cases}$$

**Result 9:** *For* $m = 1$, Result 8 reduces to the UMVUE of $(1 - p)$ as

$$G_1(z,r,n)=\begin{cases}\dfrac{n-r}{n}, & r>0; z>0\\ 1, & r=0, z=0\end{cases}$$

**Result 10:** $\psi(T_1)=(1-\frac{r}{n})^m$ is CAN estimator of $\psi(p)=(1-p)^m$ with asymptotic variance

$$\frac{m^2}{n}p(1-p)^{2m-1}.$$

**Result 11:** For $m\leq\frac{n}{2}$, the UMVUE of the variance of $G_m(Z,R,n)$ is computed as

$$\widehat{var}[G_m(z,r,n)]=\begin{cases}G_m^2(z,r,n)-G_{2m}(z,r,n), r=1,2,\dots,(n-2m)\\ G_m^2(z,r,n), & r=(n-2m+1),\dots,(n-m)\\ 0, & otherwise\end{cases}$$

$$=\begin{cases}\left[\dfrac{\binom{n-m}{r}}{\binom{n}{r}}\right]^2-\dfrac{\binom{n-2m}{r}}{\binom{n}{r}}, r=1,2,\dots,(n-2m)\\ \left[\dfrac{\binom{n-m}{r}}{\binom{n}{r}}\right]^2, & r=(n-2m+1),\dots,(n-m)\\ 0, & o.w.\end{cases}$$

**Result 12:** For $m=1$, the UMVUE of the variance of UMVUE of $(1-p)$ is given by

$$\widehat{var}[G_1(z,r,n)]=\begin{cases}\dfrac{r(n-r)}{n^2(n-1)}, & r=1,2,\dots,(n-1)\\ 0, & o.w.\end{cases}$$

**Result 13:** For $k>0$ the UMVUE of parametric function $(1-p)^n+\left(\frac{p}{\theta}\right)^k[1-(1-p)^{n-k}]$ is given by

$$H_k(z,r,n)=\begin{cases}\dfrac{(r)_k(r-1)_k}{(n)_k z^k}, & r=1,2,\dots,n;\ z>0\\ 1, & r=0;\ z=0\end{cases}$$

where $(a)_k=a(a-1)\dots(a-k+1)$, and $z=\sum_{x_{i>0}}x_i$.

For various values of $k\geq 1$, one can obtain the UMVUE of the parametric function. Unfortunately, it is impossible to find a unbiased estimate for the parameter $\theta$ alone. Aitchison (1955) through the usual classical approach obtain the UMVUE of the parametric function $(1-p)^2\theta^2$ as

$$\varphi(z,r,n)=\begin{cases}\frac{(2n-r-1)z^2}{n(n-1)(r+1)}, & r>0;\ z>0\\ 0, & r=0;\ z=0\end{cases}$$

**Result 15:** $\psi(T_2) = (\sum_{x_i>0} x_i)^m$ is CAN estimator of $\psi(\theta) = \theta^m$ with asymptotic variance $\frac{m^2\theta^{2m}}{np}$.

**Result 16:** For fixed $x$, the UMVUE of pdf $g(x; p, \theta)$ is shown as

$$\phi_x(z, r, n) = \begin{cases} \dfrac{r(r-1)}{nz}\left(1+\dfrac{x}{z}\right)^{r-2}, & 0 < x < z;\ r = 1,2, \dots, n \\ \dfrac{n-r}{n}, & x = 0;\ r = 0,1, \dots, n-1 \\ 0, & o.w. \end{cases}$$

**Result 17:** For $r = n$, that is when all the observations are coming from the density, then the UMVUE of the density $f(x; \theta)$ is simplified as

$$\phi_x(z, r, n) = \begin{cases} \dfrac{n-1}{z}\left(1+\dfrac{x}{z}\right)^{n-2}, & 0 < x < z;\ n > 1 \\ 0, & o.w. \end{cases}$$

**Result 18:** For fixed $x$, the UMVUE of variance of pdf $g(x; p, \theta)$ is obtained as

$$\widehat{var}[\phi_x(z, r, n)]$$

$$= \begin{cases} \left[\dfrac{r(r-1)}{nz}\left(1-\dfrac{x}{z}\right)^{r-2}\right]^2 \\ -\dfrac{r(r-1)^2(r-2)}{n(n-1)z(z-x)}\left(1-\dfrac{x}{z}\right)^{r-2}\left(1-\dfrac{x}{z-x}\right)^{r-3}, & 0 < x < z;\ r = 2 \dots, n \\ \left[\dfrac{r(r-1)}{nz}\left(1-\dfrac{x}{z}\right)^{r-2}\right]^2, & 0 < x < z;\ r = 2, \dots, n \\ \dfrac{r(n-r)}{n^2(n-1)}, & x = 0;\ r = 0,1, \dots, n-1 \\ 0, & o.w. \end{cases}$$

For $r = n$, all the results will reduces to that of the estimates of an exponential distribution, without inliers.

**Result 19:** For fixed $z$ and $r$, the UMVUE of the survival function $S(t) = P(X > t)$, $t \geq 0$ is obtained as

$$\hat{S}(t) = \begin{cases} \dfrac{r}{n}\left(1-\dfrac{t}{z}\right)^{r-1}, & t < z \\ 0, & o.w. \end{cases}$$

**Result 20:** For fixed $z$ and $r$, the *UMVUE* of the variance of $\hat{S}(t)$ is obtained as

$$\widehat{var}\left[\hat{S}(t)\right] = \begin{cases} \left[\dfrac{r}{n}\left(1-\dfrac{t}{z}\right)^{r-1}\right]^2 - \dfrac{r(r-1)}{n(n-1)}\left(1-\dfrac{2\,t}{z}\right)^{r-1} & ,z > 2t \\ \left[\dfrac{r}{n}\left(1-\dfrac{t}{z}\right)^{r-1}\right]^2, & t < z < \ 2t \\ 0, & o.w. \end{cases}$$

For $r = n$, both the above results reduce to the case of an exponential distribution.

**Result 21:** $\psi(T_3) = (r/n)e^{-t/\sum_{x_i>0} x_i}$ is CAN for the survival function $S(t) = P(X > t) = pe^{-t/\theta}$ with asymptotic variance $\frac{pe^{-2t/\theta}}{n\theta^2}$.

Definition 3 can be extended to multiparameter case so that CAN estimator for linear combination of parameters can be made possible. Let $T = (T_1, T_{2,\dots,}T_m)'$ be a vector valued estimator which is consistent for a vector parameter $\theta = (\theta_1, \theta_2, \dots, \theta_m)'$ then $T_i$ is CAN for $\theta_i$ with asymptotic variance $\frac{\lambda_{ii}(\theta)}{n}$ and any linear combination $T' = \sum_{i=1}^{m} l_i T_i$ is CAN for $\sum_{i=1}^{m} l_i \theta_i$ with asymptotic variance $\frac{1}{n} l' \Lambda(\theta) l$, where $\Lambda(\theta)$ is the variance-covariance matrix of vector of parameters $\theta$ (Kale and Muralidharan, 2015).

**Result 22:** Let $\psi(p, \theta) = l_1 p + l_2 \theta$, then the estimator $T' = l_1 \left(\frac{r}{n}\right) + l_2 \sum_{x_i>0} x_i$ is CAN for $\psi(p, \theta)$ with asymptotic variance $\frac{1}{n}(l_1^2 p(1-p)/n + l_2^2 \theta^2/(np))$.

We now investigate the MVU estimation of $\theta \; or \; \psi(\theta)$ based on Cramer-Rao Lower Bound (CRLB) to the variance of an unbiased estimator. Let {*f*(*x*, θ), θ ∈ Ω}, $\Omega c R^1$ be a class of distributions $I_X(\theta)$ is the Fisher Information, then under some regularity conditions (refer to Kale and Muralidharan, 2015) the CRLB for $V(T) \geq \left(\frac{d\psi(\theta)}{d\theta}\right)^2 / I_X(\theta)$. For instance, if $\psi(\theta) = \theta^2$ then the CRLB for $V(T)$ is $\frac{4\theta^4}{np}$. Similarly, the CRLB for $V(T)$ for estimating $\psi(p) = (1-p)^m$ is obtained as $\frac{mp(1-p)^{2m-1}}{n}$.

## References

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**, 901–908.

Jani, P. N. and Singh, A. K. (1995). Minimum variance unbiased estimation in multi-parameter exponential family of distributions. *Metron*, **53**, 93–106.

Kale, B. K. and Muralidharan, K. (2015). *Parametric Inference: An Introduction*. Narosa Publishing House, New Delhi.

Lai, C. D., Khoo, B. C., Muralidharan, K., and Xie, M. (2007). Weibull model allowing nearly instantaneous failures. *Journal of Applied Mathematics and Decision Sciences*, Article ID 90842: 11 pages.

Muralidharan, K. (2000). The UMVUE and Bayes estimate of reliability of mixed failure time distribution. *Communications in Statistics - Simulations and Computations*, **29**, 603 – 619.

Muralidharan, K. (2010). Inliers prone models: A review. *Prob Stat Forum*, **3**, 38–51.

Roy, J. and Mitra, S. K. (1957). Unbiased minimum variance estimation in a class of discrete distributions. *Sankhya*, **18**, 371–378.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 51-78

# Overcoming Challenges Associated with Early Bayesian State Estimation of Planted Acres in the United States

**Balgobin Nandram**
*Department of Mathematical Sciences*
*Worcester Polytechnic Institute, Worcester, MA, 01609, USA*

Received: 15 June 2023; Revised: 22 July 2023; Accepted: 24 July 2023

## Abstract

National surveys in the United States have become expensive with low response rates, and there is an abundance of administrative data (non-probability samples). Government agencies are now beginning to integrate these two sources of data to improve the quality of official statistics. Our application is on agriculture, where the study variable is planted acres and estimates early in the current year are much needed by the USDA's National Agricultural Statistics Service (NASS). A solution of this problem is important for economic, policy and many other reasons. This is a very difficult problem to solve because there are many challenges, including the poor quality of the available early survey data, that must be overcome. We attempt to solve the problem by integrating the probability samples from designed surveys and the non-probability samples, relatively much larger, which come from 'administrative' data or 'historical' data. Keeping in line with NASS's preference, we use Bayesian small area temporal models (a non-spatial model and a spatial model) to infer the early state estimates of planted acres. The Bayesian Fay-Herriot model is manipulated to link the data, and the Gibbs sampler, which is operationalized, is used to fit the two models. We show that the spatio-temporal model provides higher quality state estimates than the non-spatio-temporal model.

*Key words*: Conditional autoregressive (CAR) model; Data integration and data quality; Fay-Herriot model; Gibbs sampler; Non-probability samples; Structural error models.

**AMS Subject Classifications**: 62F15, 62D05, 62D10, 62P12

## 1. Introduction

It is the objective of this paper to show how to estimate planted acres for states early in the season in the United States. These estimates are based on historical data, administrative data and survey data. Estimates of planted acres are so important that for farmers and price analysts, almost every discussion of crop fundamentals begins with planted

Corresponding Author: Balgobin Nandram; Email: balnan@wpi.edu, balgobin.nandram@usda.gov
This paper is based upon the Keynote Address I delivered on 15 February 2023 during the 25th Annual International Conference of the SSCA held at University of Jammu during 15-17 February 2023.

acres (*e.g.*, Kansas Farm Bureau, 2020). It is our purpose to demonstrate how to integrate these data sources with limited access to the actual survey data, which are confidential, and instead we used published data. The key problem is to provide estimates of planted acres in June of any year, when the quality of the actual survey data is poor, and one needs to access other data, most of which are available to the public. We imagine that the current year is 2021 and estimates of planted acres are required in June. The methodology is being developed so that it can be used readily for June of any year just after the survey data are available. However, it is not the purpose of this paper to present methodology or substantial results from modeling real data; rather it is intended to show critical thinking on, and the struggles in the original stages of, this project at the National Agricultural Statistics Service.

We have Farm Service Agency (FSA) data, which are administrative data. These are voluntary to the farmers and are essentially a non-probability sample. In our models, we integrate survey data and non-survey data in the "current year" (2021), collected in June, and all available data over the past decade. We work at state level because county level data are not available in June. We include all possible data sources. It is required that all model estimates must cover (*i.e.*, larger than) FSA planted acres. We also have final results from the Agricultural Statistics Board (ASB). Both FSA and ASB values are historical data (before 2021). FSA values are not available in June of the current year but ASB values are available in March and June; see The National Agricultural Statistics Service (2021 a) for prospective plantings in March and The National Agricultural Statistics Service (2021 b) for acreage in June. We have analyzed planted acres (thousands of acres) for corn, which is our focus, and a similar analysis can be done for other crops such as soybeans.

NASS conducts quarterly Agricultural Production Survey (APS) in an ongoing effort to capture activities throughout the life cycle of the crop. These include planting intentions (March), early estimates of planted acreage (June, with some intentions), and output activities for small grains crops (September) such as buckwheat, flax, oats and rye, and major row crops (December) such as corn, soybeans, cotton, potato. The June Area Survey (JAS) provides an under-coverage adjustment for the list-based samples obtained during the June, September and December APS surveys.

According to Young and Chen (2022), "The NASS conducts more than a hundred national surveys and produces more than 400 reports each year. An annual publication calendar details the day and time each report is to be released, and the NASS has consistently released its reports according to schedule more than 99% of the time." The NASS acreage and production reports are considered by many to be the "final word" because they are **Unbiased** (they are not influenced by either buyers or sellers of commodities); **Timely** (data are provided well in advance of when they will be available from other sources); **Consistent** (the same statistically sound procedures are followed each time, building on a multi-year data-base); and **Transparent** (NASS ensures that all participants have equal access to the information). For further discussions of these notions, see, for example, Kansas Farm Bureau (2020). The Research Development Division (RDD) at NASS is charged to ensure that all procedures are current, and if not, they are revised and new methods are developed.

There are several reasons why early estimates of planted acres are needed in the

United States.

1. June Acreage Report is a very important economic indicator in the United States and the amount of planted acres affect prices later. It follows the well known demand and supply principle in economics.

2. Stake holders and economists need high quality estimates, and it is incumbent on NASS to provide these estimates. NASS is required by law (Agricultural Marketing Act of 1946 and the Census of Agriculture Act of 1997) to produce estimates for several key crops as early as March. Markets are hungry for information.

3. Budgets are allocated to different programs around this time.

4. Many internal programs at NASS depend on quality estimates of planted acres (*e.g.*, cash rental rates), so these numbers must be reported early.

5. Even before March, many farmers are, or are considering marketing portions of their expected production.

6. To make the most informed decision, farmers, agribusinesses and even speculators need as accurate a picture as possible as what market fundamentals are and how those fundamentals are changing as the year progresses. USDA's NASS provides objective information to all market participants at the the same time at no cost.

Now, we give an idea of the order of magnitude of the APS and the JAS. March APS has about 80,000 US farm operators, a survey of farmers conducted in the first two weeks of March to get intentions, selected from a list of farmers that ensure that all operations had a chance to be selected. Note that intentions are not binding, and the farmers could change their minds, and this is a difficulty that is impossible to address in June. Like all NASS surveys, data are collected by mail, internet, telephone, and personal interviews. June Crop Acreage report, which includes two surveys, the APS, a survey of over 70,000 farmers, asking the farmers how many acres they had planted and still intended to plant, and the JAS, which includes over 11,000 individual (one square mile) segments, in which enumerators physically inspect, to see what has been planted (and then ask the farmers what will be planted on any unplanted tracts in the segment). This is a dual-frame survey and the two surveys are combined to complete the June Crop Area estimates.

Every farmer participating in the USDA Farm Service Agency (FSA) programs, such as marketing assistance loans or deficiency payments, must file an FSA-578 Report of Acreage. However, the acreage report deadline is July 15 for FSA (not March or June), and not every farmer gets it on time, and not every county office office gets the data inputted immediately. Also, not every farmer participates in the FSA programs. Consequently, the August FSA reports underestimates planted acres. Therefore, it is still an important constraint that must be incorporated into our model for the June estimates of planted acres; see Office of the Chief Economist (2019). It turns that it is a difficult problem to incorporate the constraint directly into the model, but this is not our purpose in this paper; see Nandram *et al.* (2021, 2023) and Chen, Nandram and Cruze (2022) of work already done at NASS, where the constraints are placed directly into the models.

Next, we discuss the challenges that we must overcome to provide estimates of planted acres with reasonable and satisfactory quality. It is pertinent to list the challenges here.

1. FSA current year values are not available in June. They become available later in August.
2. The dual-frame (APS/JAS) model estimates must be larger than the FSA values.
3. State survey indications do not capture variation very well.
4. There are outlying states (some very large and some very small).
5. With the initialization of modernization and unification at NASS, we want to combine administrative data (non-probability sample) with the surveys (APS/JAS). Historical data (available) are incorporated as the non-probability sample; we have 10 years of ABS/FSA data before the current year.
6. NASS wants model estimates for 48 US states (excluding Alaska and Hawaii). Typically data for corn may be available for all 48 states with missing survey data; soybeans are available from fewer number of states, actually 29.
7. Weather variables (temperature and precipitation) are difficult to use, although they are important. Current work at NASS is now trying to make use of the weather variables.
8. Landsat satellite (imagery) data are of poor quality in June, and they are not useful; in March there are only intentions.
9. Covariates must be incorporated as well; there are missing values here also.
10. Meeting the annual production schedule is difficult.
11. National Academy of Sciences, Engineering and Medicine (2017) recommended that NASS use Bayesian Small Area models. These models are complicated, and Markov chain Monte Carlo methods (ıe.g., Gibbs samplers) are needed to fit them.

A non-probability sample and a probability sample can be combined in several ways. This depends on available data; see Rao (2020) for both design-based and model-based approaches for making valid inferences by integrating data from surveys and other sources. Also, Li, Chen and Wu (2020) presented double robustness with quasi-randomization via propensity scores. Nandram, Choi and Liu (2021) and Nandram and Rao (2021, 2023) provided Bayesian analyzes. But these can be carried out when survey weights are available from the probability sample. In the current work, survey weights are already incorporated into the survey indications for states, and combining the two samples need an alternative approach. We use a measurement error model to combine the two samples; see Fuller (1987) and Berg *et al.* (2021). In our model, there is a linear relation between the FSA values and the ASB values for the historical data (the non-probability sample), and this same relation holds between the current year's FSA values and the true value of planted acres. This permits integration of the two data sources.

This paper has five sections including the current one, Section 1. In Section 2, we describe the available data. In particular, we describe how to estimate the FSA values before June of the current year, 2021. We also show how to impute the missing indications and variances. In Section 3, we describe the temporal models, a non-spatial model and a spatial model, which we use for comparison. We also describe the computations, and model diagnostics. In Section 4, we present the data analysis of the public-used data. Section 5 has concluding remarks. Appendix A contains a short description on how to go down to the level of Agricultural Statistics Districts (ASD) for further analysis. Appendix B has a brief description of how to deal with clustering in the indications. Appendix C contains a list of abbreviations used in the paper.

## 2. Available data and FSA values

In this section, we give a more detailed discussion of the data we must use to exemplify the actual situation. We primarily study corn, but there are other crops of interest such as soybeans, All wheat and All cotton; again see The National Agricultural Statistics Service (2021 a) for prospective plantings in March and The National Agricultural Statistics Service (2021 b) for acreage in June.

### 2.1. General data

We have Farm Service Agency (FSA) and Agricultural Statistics Board (ASB) historical data for the past ten years before 2021, and these are not confidential. Our idea is that the relation between the FSA values and the ASB values should be similar to the relation between the FSA values and the true planted acres in the current year. This is how the non-probability sample (FSA values and ASB data) are used. In March, there are indications on planting intentions, approved by the ASB; the March ASB values are also available for the past ten years. As stated, we have 10 years of FSA values before 2021, but not in 2021, which we need. Note that ASB values are available to the public. We have the Agricultural Production Survey (APS) and June Area Survey (JAS) dual frame survey indications, but these are confidential, they are not available for the public use, and they are not used in this paper. However, approved estimates are available in June for the public, and these are the ones used in this paper for exploratory analysis. There is on-going work on the actual data at NASS.

We also have five covariates, which are Percent farmland irrigated - $x_2$, Population density - $x_3$, Value of cropland - $x_4$, National commodity crop production index (NCCPI), an index of soil quality, - $x_5$, Number of farms - $x_6$. These are publicly available. A simple regression of June 2021 survey indications on the covariates gives an $R^2 \approx 75\%$; $x_4$ and $x_6$ are significant; $x_2$, $x_3$, $x_5$ and $x_4 * x_6$ are not significant. Other variables such as weather (temperature and precipitation) are currently being explored at NASS.

In Figure 1, we show the maps of quintiles of the FSA values and the survey indications of planted acres in the current year. We can see some differences (e.g., OK moves from 3 to 2, OH from 4 to 5, AL from 2 to 3).

## 2.2. Imputing missing data

We show how to impute the FSA values of the current year (2021). Then we show how to impute missing indications and variances. The public use data on planted acres do not come with estimated variances, which are needed in the Fay-Herriot model; see The National Agricultural Statistics Service (2021 a, b).

### 2.2.1. Current year FSA values

We use $T+1=10$ (*i.e.*, $T=9$) years of FSA values and March intentions (put out by ASB) to predict the current year FSA values. We denote the ten years of historical data by

$$(\hat{\theta}_{it}^{(f)}, \hat{\theta}_{it}^{(a)}), i=1,\ldots,\ell, t=-T,\ldots,0.$$

Note that $t=1$ is the current year, the year of interest. Then, we use simple linear regression,

$$\hat{\theta}_{it}^{(f)} = \beta_0 + \beta_1 \hat{\theta}_{it}^{(a)} + e_i, i=1,\ldots,\ell, t=-T,\ldots,0.$$

We fit this model to get the following estimates of the regression coefficients. The 10 speculative states for corn gave $\hat{\beta}_0 = -58.05, \hat{\beta}_1 = .978, R^2 \approx 1$. Although it is not particularly relevant, the 11 speculative states for soybeans gave $\hat{\beta}_0 = 13.03, \hat{\beta}_1 = .986, R^2 \approx 1$. Therefore, the fits are pretty good for both corn and soybeans. However, there are some aberrations for smaller states (38 for corn and 18 for soybeans). Finally, we predict the current year FSA values from

$$\hat{\theta}_{1i} \equiv \hat{\theta}_{i1}^{(f)} = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_{i1}^{(a)}, i=1,\ldots,\ell.$$

The $\hat{\theta}_{1i}$ will be used as part of the observed data in this paper or at NASS.

It is possible to improve this procedure using covariates such as precipitation and temperature (under study at NASS).

### 2.2.2. Missing indications and variances

Fewer than 48 states are observed for corn and fewer than 29 states for soybeans; some states are missing both indications and variances. We use the adjacent neighbors of a specific state without indications and/or variances via an incidence matrix to impute the remaining states for corn. The same can be done for soybeans (currently under experimentation at NASS) and other crops such as All wheat and All cotton.

Let $\mathcal{C}_i$ denote the set of adjacent neighbors of the $i^{th}$ state, and let $n_i$ denote the number of counties in the $i^{th}$ state. Then, if the $i^{th}$ state's indication and/or variances are missing, define

$$\hat{\theta}_i = \frac{\sum_{j \in \mathcal{C}_i} n_j \hat{\theta}_j}{\sum_{j \in \mathcal{C}_i} n_j} \text{ and } \hat{\sigma_i^2} = \left\{ \prod_{j \in \mathcal{C}_i} (\hat{\sigma_j^2})^{n_j} \right\}^{\frac{1}{\sum_{j \in \mathcal{C}_i} n_j}},$$

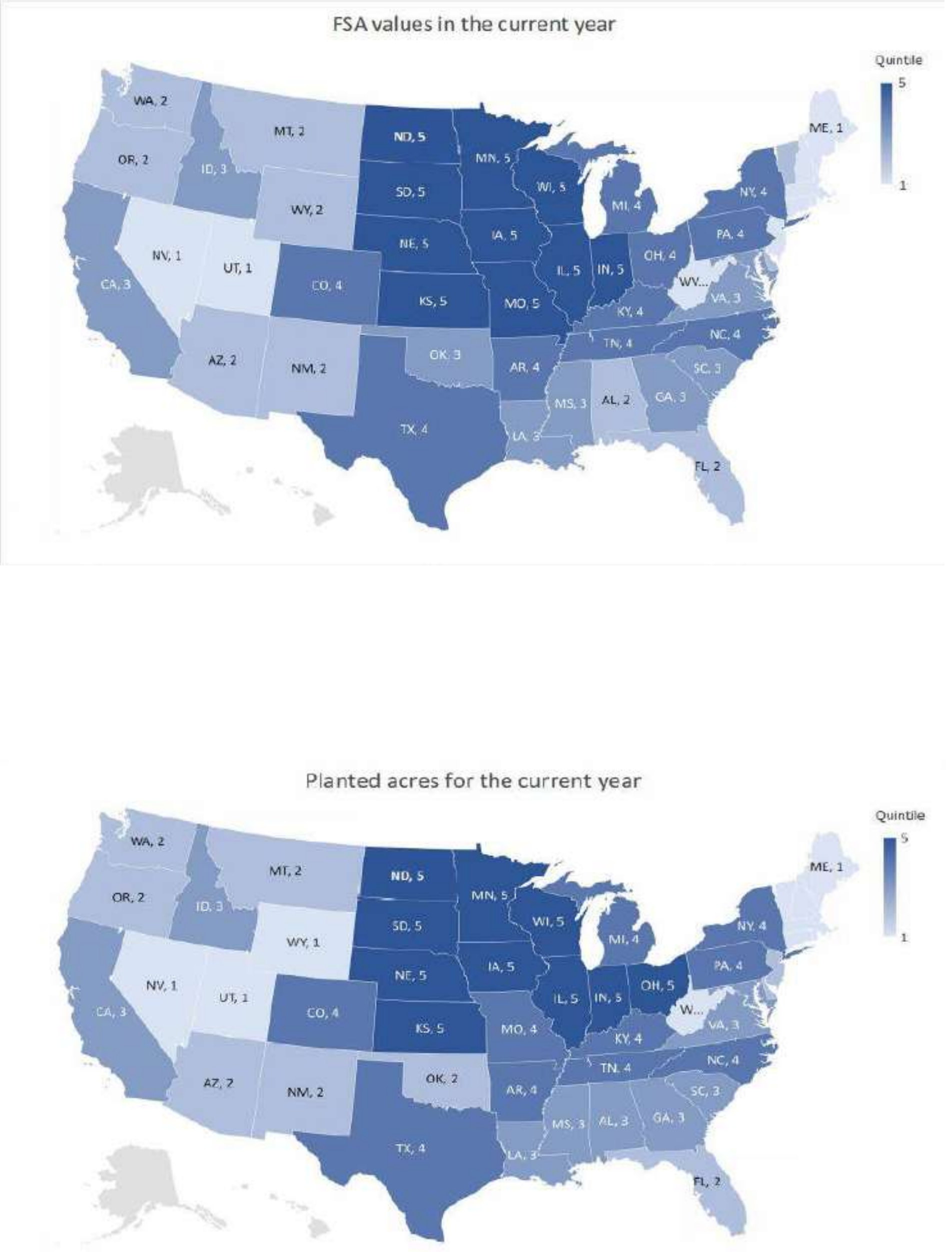


**Figure 1: Quintiles of FSA values and survey indications of planted acres: The quintiles for the FSA values (survey indications) are** 74 (85), 293 (330), 597 (640), 3276 (3350), **and for the FSA values (survey indications), the minimum and maximum values are** 1.89 (2.00) **and** 12323 (13100).

the weighted arithmetic mean and the weighted geometric mean for respectively the indications and variances. These are the 'observed' indications and variances corresponding to the missing states in the public-used data.

Public-used ASB indications do not come with variances, but these are confidential data available at NASS for the survey indications. Because of confidentiality, we cannot use these data in this paper. Besides the variances are too small (optimistic) because of the large amount of data that go into a state indication. When the state indications are obtained, heterogeneity and clustering in the data are not taken into account. The data are weighted (to reflect the survey design) from the operation (farm) level to state level.

Assuming that $\hat{\theta}_i$ is observed, we take

$$\hat{\sigma}_i^2 = CV_i^2 \times \hat{\theta}_i^2, i = 1, \ldots, \ell,$$

where $CV_i$ is the coefficient of variation and the $\hat{\theta}_i$ are the state indications. Here $CV_i$ is also unknown, so we take

$$CV_i = \text{Uniform}(.10, .50), i = 1, \ldots, \ell,$$

because a coefficient of variation of .30 is taken to be a threshold at most government agencies in the United States. Here $\ell = 48$ for corn and 29 for soybeans. This procedure is a bit problematic because it penalizes some large states and some small states appear too good.

An alternative and slightly better procedure is to take $CV_i$ to be inversely proportion to the number of counties, $n_i$, in the $i^{th}$ state, and $\frac{1}{\ell}\sum_{i=1}^{\ell} CV_i = .30$, again a threshold for a reliable estimate in US government agencies. This gives

$$CV_i = \max\left(.10, \frac{0.30}{\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{n_i}}\right), i = 1, \ldots, \ell,$$

where $n_i$ is the number of counties in the $i^{th}$ state, and for flexibility, the $CV_i$ can be kept larger than 0.10 for speculative states. But we have not done so for this paper. It is another difficult problem to specify the coefficient of variations, and clearly more data are needed.

As a summary, we present the data we want to analyze. The true values that we want to estimate are denoted by $\underset{\sim}{\theta} = (\theta_i, i = 1, \ldots, \ell)$, where $\ell$ denotes the number of states. This will vary with different commodities, but as was stated we will deal only with planted acres (thousands of acres) for corn. We denote the data by $D$, where

$$D = \{\underset{\sim}{\hat{\theta}}^{(f)}, \underset{\sim}{\hat{\theta}}^{(a)}, \underset{\sim}{\hat{\theta}}_1, \underset{\sim}{\hat{\theta}}_2, \underset{\sim}{\hat{\sigma}}_2^2\}.$$

The (FSA, ASB) historical values are $(\hat{\theta}_{ti}^{(f)}, \hat{\theta}_{ti}^{(a)}), t = -T, \ldots, 0$ ; the current year FSA values are $\hat{\theta}_{1i}$, obtained by imputation; the current year survey indications and variances are $(\hat{\theta}_{2i}, \hat{\sigma}_{2i}^2)$ in June, obtained from the APS and the JAS; and the covariates are $\underset{\sim}{x}_i, i = 1, \ldots, \ell, c = 6$, including an intercept. We adapt the Fay-Herriot model, and a novel simplification is to introduce the ratios,

$$\kappa_i = \sigma_0^2/\hat{\sigma}_i^2, i = 1, \ldots, \ell, \sigma_0^2 = \text{ GM of } \hat{\sigma}_i^2,$$

where GM stands for geometric mean, and the $\kappa_i$ are assumed known. We will introduce two models, which use all the available data, and they are a non-spatio-temporal (NST) model and spatio-temporal (ST) model; the $\theta_i$ are linked to the covariates, $\underline{x}_i$, via regression with unknown coefficients. There are two additional features: First the models take care of outliers, and second, the constraint, $\theta_i > \hat{\theta}_i, i = 1, \ldots, \ell$, is not part of the models, but it is taken care of in the output analysis. This avoids model complications. For a single model, Bayesian diagnostics are not appropriate if the constraint is included because they check how close the predictive data are to the observed data.

## 3. Bayesian small area models

Small area models are appropriate because the indications from many small states are not reliable. For survey data, there is only one data point for each state. There are also supplemental data for ten years before the current year, and the FSA imputed value for each state of the current year, 2021. In Section 3.1, we present the two models and we briefly describe the computation. In Section 3.2, we present a diagnostic assessment of the two models.

### 3.1. Models and computations

We first describe the non-spatial model. The spatial model is similar except with one adjustment.

For the historical data, we assume

$$\hat{\theta}_{ti}^{(f)} \mid \{\hat{\theta}_{ti}^{(a)}, \alpha_0, \alpha_1, \psi_1, \sigma^2\} \overset{ind}{\sim} \text{Normal}(\alpha_0 + \alpha_1 \hat{\theta}_{ti}^{(a)}, \psi_1 \sigma^2), \tag{1}$$

$t = -T, \ldots, 0, i = 1, \ldots, \ell$, and for the current year's FSA values, we assume

$$\hat{\theta}_{1i} \mid \{\alpha_0, \alpha_1, \theta_i, \sigma^2\} \overset{ind}{\sim} \text{Normal}(\alpha_0 + \alpha_1 \theta_i, \sigma^2). \tag{2}$$

For indications and variances, we assume

$$\hat{\theta}_{2i} \mid \{\theta_i, z_i = 0, p, \sigma^2, \psi_2\} \overset{ind}{\sim} \text{Normal}(\theta_i, \psi_2 \frac{\sigma^2}{\kappa_i}),$$

$$\hat{\theta}_{2i} \mid \{\theta_i, z_i = 1, p, \sigma^2, \psi_2\} \overset{ind}{\sim} \text{Normal}(\theta_i, \frac{\sigma^2}{\kappa_i}) \tag{3}$$

$$z_i \mid p \overset{ind}{\sim} \text{Bernoulli}(p)$$

$$\theta_i \mid \{\underline{\beta}, \sigma^2, \rho\} \overset{ind}{\sim} \text{Normal}(\underline{x}_i' \underline{\beta}, \frac{\rho}{1-\rho} \sigma^2), \tag{4}$$

and

$$\pi(\alpha_0, \alpha_1, p, \rho, \psi_1, \psi_2, \underline{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}\text{Beta}(\frac{\sqrt{\ell}}{2}, \frac{\sqrt{\ell}}{2}). \tag{5}$$

The beta prior is used for stability, and it is motivated by minimum mean square error; see Casella and Berger (2002, *pg.* 332) for example. The mixture model, used to accommodate outliers and robustness to normality, is an extension of the Fay-Herriot model (Fay and Herriot, 1979). Also, see Goyal, Datta and Mandal (2020) for a slightly different formulation of the mixture model.

Because the parameters are weakly identified in the survey part of the model, there is a need to specify bounds for $\alpha_1$ and $\alpha_2$, and we do so using an exploratory data analysis, namely $a_0 < \alpha_1 < a_1$, $b_0 < \alpha_2 < b_1$. We also specify $c_0 < \rho < c_1$. We believe the relation in (1) is tight so we assume $0 < \psi_1 < 1$. We also assume that $0 < \psi_2 < 1$ because outliers should be more variable than non-outliers, and $0 < p < 1/2$ because there should be fewer outliers than non-outliers. These latter assumptions are natural, and all constraints are incorporated into the model when it is fit using the Gibbs sampler. However, for simplicity, the constraint, $\theta_i > \hat{\theta}_{1i}, i = 1, \ldots, \ell$, that the model estimates are larger than FSA values is incorporated into the output analysis, not within the Gibbs sampler.

Note that the non-probability sample and the probability sample are linked by (2), and (1) and (2) have the same regression coefficients.

For the spatial model, we use the conditional auto-regressive (CAR) model,

$$\underline{\theta} \mid \{\underline{\beta}, \sigma^2, \rho\} \overset{ind}{\sim} \text{Normal}\{X\underline{\beta}, \frac{\rho}{1-\rho}\sigma^2(R - \psi_3 W)^{-1}\}, X = (\underline{x}_i'), \tag{6}$$

where $\lambda_1, \ldots, \lambda_\ell$ are eigenvalues of $R^{-1}W$ in increasing order (some negative and some positive). We simply replace (4) by (6) with an extra parameter, $\psi_3$, beyond the less flexible intrinsic ($\psi_3 = 1$) CAR model (Janicki *et al.* 2022). A priori, we assume

$$\pi(\alpha_1, \alpha_2, p, \rho, \psi_1, \psi_2, \psi_3, \underline{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}\text{Beta}(\frac{\sqrt{\ell}}{2}, \frac{\sqrt{\ell}}{2}), 0 < \psi_1, \psi_2 < 1, \frac{1}{\lambda_1} < \psi_3 < \frac{1}{\lambda_\ell}, \tag{7}$$

replacing (5) by (7). The NST model and the ST model are discussed in great detail in Nandram (2022), but this report is confidential. An earlier discussion is given by Berg et al. (2021); many issues in that paper are addressed in the report. This is part of the general measurement error model (*e.g.*, Fuller, 1987).

Let $\Omega = (\alpha_1, \alpha_2, \underline{\beta}, \psi_1, \psi_2)$ for the non-spatial model, $\Omega = (\alpha_1, \alpha_2, \underline{\beta}, \psi_1, \psi_2, \psi_3)$ for the spatial model ($\psi_3$ is not in the nonspatial model), and $D = \{\underline{\theta}^{\hat{(f)}}, \underline{\theta}^{\hat{(a)}}, \underline{\hat{\theta}}_2, \underline{\hat{\theta}}_2, \underline{\hat{\sigma^2}}\}$ denote the data. Then, using Bayes' theorem, the joint posterior density is

$$\pi(\Omega, \underline{z}, p, \underline{\beta}, \underline{\theta}, \sigma^2 \mid D).$$

We state the following steps in the griddy Gibbs sampler.

1. Integrate out $\underline{\theta}$ to get
$$\pi(\Omega, \underline{z}, p, \underline{\beta}, \sigma^2 \mid D).$$

2. Draw $(\underline{z}, p)$ together (collapsing and blocking),
$$\pi(\underline{z}, p \mid \Omega, \underline{\beta}, \sigma^2, D) = \pi(\underline{z} \mid p, \Omega, \underline{\beta}, \sigma^2 \mid D)\pi(p \mid \Omega, \underline{\beta}, \sigma^2 \mid D).$$

3. Draw $(\underline{\beta}, \sigma^2)$ together (collapsing and blocking),
$$\pi(\underline{\beta}, \sigma^2 \mid \Omega, \underline{z}, p, D) = \pi(\sigma^2 \mid \Omega, \underline{z}, p, \underline{d})\pi(\underline{\beta}, \mid \Omega, \underline{z}, p, \sigma^2, D).$$

4. Sample $\pi(\Omega \mid \underline{z}, p, \underline{\beta}, \sigma^2, D)$.

5. Monitor convergence (Geweke test and effective sample size).

6. Sample the Rao-Blackwellized density, $\pi(\underline{\theta} \mid \Omega, \underline{z}, p, \underline{\beta}, \sigma^2, D)$, in the output analysis subject to constraints (model estimates must cover FSA values). These are truncated univariate normal densities for non-spatial model and truncated multivariate normal densities for spatial model.

Markov chain Monte Carlo methods (Gibbs sampler with some collapsing and blocking to improve convergence and better mixing) are used to fit the two models; see Liu (1994) for collapsing and Tan and Hobert (2009) for blocking. In fact, we use the griddy Gibbs sampler (Ritter and Tanner, 1992) in which some CPDs are sampled using the grid method. The constraints are not included in the models to allow them to be as simple as possible, rather they are performed in an output analysis. In the non-spatial model, this is straight forward as we can sample from independent truncated normal densities, but in the spatial model, we need to sample from truncated multivariate normal densities (Ridgeway 2016). The constraint $\theta_i > \hat{\theta}_i, i = 1, \ldots, \ell$, in the output analysis.

In Table 1 we show the good performance of the Gibbs sampler under both models. Specifically, the Geweke tests show that the Gibbs sampler is stationary with all p-values being larger than .05 and the effective sample size (ESS) of each parameter is the nominal value of 1000, except the one for $\rho$ under the ST model, but this is still good. This shows that the two Gibbs samplers are strongly mixing. Also, note that the computational times are also operational at NASS; see the note to Table 1.

### 3.2. Model diagnostics

We use standard Bayesian diagnostics to check the goodness of fit of the two models. We assess the more interesting mixture part of the model (*i.e.*, the survey data).

We start by computing two simple diagnostic measures. Let $PM_i$ and $PSD_i, i = 1, \ldots, \ell$, denote the posterior means and posterior standard deviations from the two models.

Specifically, we have computed

$$ARES = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(\hat{\theta}_{2i} - PM_i)^2}, \quad ASTD = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell} PSD_i^2}.$$

For the non-spatial (spatial) model, we have $ARES = 2388$ (2005) and $ASTD = 366$ (120), showing the spatial model has performed much better than the non-spatial model in terms of these two measures. It is very good for the spatial model that it provides estimates closer to the direct estimates (indications) with smaller posterior standard deviations.

As a further check on the models, we have done a Bayesian cross-validation analysis (*i.e.*, delete one observation and predict it). The idea is the same for both models, but the specific mathematical formulas are different for the non-spatial model and the spatial model. Define

$$f(\hat{\theta}_{2i} \mid \underline{\hat{\theta}}_{(2i)}) = \sum_{h=1}^{M} W_{ih} f(\hat{\theta}_{2i} \mid \underline{\hat{\theta}}_{(2i)}, \Omega^{(h)}), W_{ih} = \frac{\{f(\hat{\theta}_{2i} \mid \Omega^{(h)}\}^{-1}}{\sum_{h=1}^{M}\{f(\hat{\theta}_{2i} \mid \Omega^{(h)}\}^{-1}}, i = 1, \ldots, \ell.$$

The residuals are $r_i = \hat{\theta}_{2i} - \mathrm{E}(\theta_{2i} \mid \underline{\hat{\theta}}_{(2i)}), i = 1, \ldots, \ell$. Then, a dispersion measure (DM, Wang *et al.* 2011), which we have developed, is

$$DM_1 = \frac{1}{\ell}\sum_{i=1}^{\ell} |r_i|,$$

and as this measure is not invariant to scale, we have now modified it to

$$DM_2 = \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{|r_i|}{\mathrm{Std}(\theta_{2i} \mid \underline{\hat{\theta}}_{(2i)})}.$$

We also counted the number, $n_0$ of $r_i > 0$, the number, $n_3$, of $|r_i| \geq 3$ and the number, $n_4$, of $|r_i| \geq 4$. For the non-spatial (spatial) model, we got $DM_1 = 1144\ (110), DM_2 = 5.61\ (0.91), n_0 = 28\ (22), n_3 = 29\ (13), n_4 = 22\ (8)$. The spatial model is much better than the non-spatial model under these measures.

We have also calculated three standard Bayesian diagnostics with respect to the survey indications, $\hat{\theta}_{2i}$, which are the deviance information criterion (DIC), the Bayesian predictive p-value (BPP) and the log-pseudo marginal likelihood (LPML). The DICs are 875 (803), the BPPs are .399 (.594) and the LPMLs are $-417$ ($-419$) for the non-spatial (spatial) model. For the BPP and LPML there is basically no preference. However, the DIC does show that the spatial model is significantly better than the non-spatial model.

Finally, we compute the average absolute relative deviation (AARD) and the square root of the average squared relative deviation (RASRD), where we compare the posterior

**Table 1: Gibbs sampler diagnostics ($p$-values of Geweke test and effective sample sizes)**

| | Non-spatial | | Spatial | |
|---|---|---|---|---|
| | *P-val* | *ESS* | *P-val* | *ESS* |
| $\beta_1$ | .80 | 1000 | .67 | 1000 |
| $\beta_2$ | .77 | 1000 | .25 | 1000 |
| $\beta_3$ | .29 | 1000 | .63 | 1000 |
| $\beta_4$ | .38 | 1000 | .84 | 1000 |
| $\beta_5$ | .48 | 1000 | .66 | 1000 |
| $\beta_6$ | .92 | 1000 | .83 | 1000 |
| $\sigma^2$ | .19 | 1000 | .74 | 1000 |
| $\alpha_1$ | .21 | 1000 | .60 | 1000 |
| $\alpha_2$ | .22 | 1000 | .47 | 1000 |
| $p$ | .75 | 1000 | .06 | 1000 |
| $z$ | .85 | 1000 | .60 | 1000 |
| $\rho$ | .44 | 1000 | .97 | 884 |
| $\psi_1$ | .18 | 1000 | .63 | 1000 |
| $\psi_2$ | .61 | 1000 | .19 | 1000 |
| $\psi_3$ | – | – | .09 | 1000 |

NOTE: For the non-spatial model, the Gibbs sampler is run $55,000$ times, with a "burn in" of $5,000$ and we pick every $50^{th}$ one and this takes 3 minutes; for the spatial model, the Gibbs sampler is run $75,000$ times, with a "burn in" of $15,000$ and we pick every $60^{th}$ one and this takes 49 minutes. Here $z$ is the number of outliers.

means of planted acres to last years ASB values as

$$AARD = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{|PM_i - ASB_i|}{ASB_i}, \quad RASRD = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}\left\{\frac{PM_i - ASB_i}{ASB_i}\right\}^2}.$$

We expect the current year's ASB values, which are unknown, to be similar to last year's. In Table 2 we show that the ST model does better than the NST model; the numbers under the ST model are smaller than those under the NST model. Specifically, the spatio-temporal (ST) model has smaller AARD and RASRD values than under the non-spatio-temporal (NST) model with or without the constraints.

**Table 2: Average absolute (squared) relative deviation by model and constraint**

| Constraint | *AARD* | | *RASRD* | |
|---|---|---|---|---|
| | *NST* | *ST* | *NST* | *ST* |
| No | 0.240 | 0.209 | 0.349 | 0.315 |
| Yes | 0.272 | 0.229 | 0.391 | 0.365 |

NOTE: NST: Non-spatio-temporal; ST: Spatio-Temporal

## 4. Data analysis

In this section, for corn we compare the NST model and the ST model under the constraint that model planted acres must be larger than the FSA values. First, we look at the important hyper-parameters to show their importance in the models. Second, we look at the model estimates of the planted acres. In the summaries, we use posterior mean (PM), posterior standard deviation (PSD), posterior coefficient of variation (PCV) and 95% highest posterior density interval (HPDI) for the true state planted acres ($i.e.$, $\theta_i, i = 1, \ldots, \ell$). We consider only corn with $\ell = 48$ states. We also use maps and graphs to make more detailed comparisons.

### 4.1. Posterior inference of hyper-parameters

We look at posterior inference of some of the nuisance parameters. For example, the regression parameters contain important information; see Table 3.

Now, we discuss the results in Table 3. First, the Percent farmland irrigated has a negative effect on planted acres. Most of the speculative states for corn, except Nebraska, have little irrigation systems; California and the southern states have a lot of irrigation systems but less corn production. The value of cropland has a positive effect on planted acres, as it should. NCCPI has a positive effect on planted acres for corn. This must be true because better soil should lead to higher planted acres. This is also a good showing for the ST model, as under the NST model, while there is a large probability that $\beta_6$ is positive, the 95% HPDI contains 0. However, the Number of farms has a negative effect on planted acres. One possible explanation is the following. As the number of farms go up, one would expect smaller farms. In smaller farms, one would expect a larger variety of commodities, not fully dominated by corn.

We note that $\sigma^2$ is estimated very well under the ST model. It has a PCV of 2.21 under the NST model, but under the ST model, the PCV is 0.07, a huge improvement. The 95% HPDI for $\alpha_1$ is $(-23.68, 14.14)$ under the ST model, and it is good that $\alpha_1$ is not significant. Also, the 95% HPDI for $\alpha_2$ is $(.998, 1.009)$ under the ST model, and it is good that one is in it. (This is not true for the NST model.) This is important because it shows the power of the historical data. Here $\alpha_1$ and $\alpha_2$ are not identifiable in the models if there were no historical data. Another important point is that $\psi_1$ is closed to one in the ST model, but not so close under the NST model. Finally, the features of $p$, $z$ and $\rho$ are almost the same under both models. It is good that $\rho$ and $\psi_3$ are large under the ST model because it

**Table 3: Posterior summaries of hyper-parameters**

| | Non-spatial | | | | Spatial | | | |
|---|---|---|---|---|---|---|---|---|
| | *PM* | *PSD* | *PCV* | *HPDI* | *PM* | *PSD* | *PCV* | *HPDI* |
| $\beta_1$ | 2556.49 | 438.95 | 0.17 | (1750.30, 3336.09) | 2483.14 | 299.13 | 0.12 | (1868.99, 3044.76) |
| $\beta_2$ | -34.44 | 14.91 | -0.43 | (-60.13, -8.30) | -21.42 | 8.37 | -0.39 | (-37.19, -6.19) |
| $\beta_3$ | 0.56 | 0.49 | 0.89 | (-0.38, 1.41) | 0.01 | 0.36 | 19.63 | (-0.71, 0.72) |
| $\beta_4$ | 2771.19 | 207.03 | 0.07 | (2346.28, 3120.29) | 2932.2 | 105.1 | 0.04 | (2730.42, 3148.56) |
| $\beta_5$ | 3.24 | 3.48 | 1.07 | (-3.45, 9.31) | 6.27 | 2.12 | 0.34 | (2.18, 10.31) |
| $\beta_6$ | -0.03 | 0.004 | -0.15 | (-0.03, -0.02) | -0.03 | 0.002 | -0.08 | (-0.03, -0.02) |
| $\sigma^2$ | 58233 | 128555 | 2.21 | (16121, 167976) | 30922 | 2269 | 0.07 | (26396, 34931) |
| $\alpha_1$ | 3.55 | 1.19 | 0.34 | (2.01, 5.86) | -5.96 | 9.86 | -1.66 | (-23.68, 14.14) |
| $\alpha_2$ | 0.998 | 0 | 0 | (0.997, 0.998) | 1.002 | 0.004 | 0.004 | (0.998, 1.009) |
| $p$ | 0.38 | 0.09 | 0.24 | (0.19, 0.50) | 0.38 | 0.09 | 0.25 | (0.19, 0.50) |
| $z$ | 18.46 | 5.43 | 0.29 | (8.00, 28.00) | 18.38 | 5.54 | 0.3 | (6.00, 27.00) |
| $\rho$ | 0.96 | 0.003 | 0.003 | (0.95, 0.97) | 0.96 | 0.002 | 0.003 | (0.96, 0.97) |
| $\psi_1$ | 0.51 | 0.28 | 0.57 | (0.04, 0.98) | 0.99 | 0.01 | 0.01 | (0.97, 1.00) |
| $\psi_2$ | 0.51 | 0.28 | 0.57 | (0.02, 0.95) | 0.68 | 0.22 | 0.32 | (0.30, 1.00) |
| $\psi_3$ | – | – | – | (–,–) | 0.87 | 0.02 | 0.02 | (0.83, 0.89) |

NOTE: The five covariates are **Percent farmland irrigated**, Population density, **Value of cropland**, **National commodity crop production index (NCCPI)** and **Number of farms**. Here $z$ is the number of outliers. (The bolded covariates are important.)

shows that the CAR model has a significant effect.

## 4.2. Posterior inference for planted acres

In this section we compare the NST model and the ST model when we make posterior inference about planted acres under the constraint that the model planted acres are larger than the FSA planted acres.

In Table 4 we present posterior inference for the first thirteen states (in the order of state abbreviations), including small (*e.g.*, AZ, CT, FL) and some large (*e.g.*, IL, IN, IA) corn producing states. Apart from rounding, the constraints are satisfied in all states. The PMs are mostly similar and the PSDs under the spatial model are mostly smaller than those under the non-spatial model. This makes the PCVs under the spatial model mostly smaller than those under the non-spatial model, and therefore the 95% HPDIs are much shorter. These PCVs are smaller than the corresponding ones for the 'observed' data. Specifically, note that the gains in PCVs for CA, CO, FL and IL with unreliable data (larger CVp2). There are similar patterns for the other states, which are too numerous to list. We will look at all the states in greater detail using several plots (see below).

We now compare the spatial structure of the corn data under the constraint models. We have used the quintiles of the posterior means; note that the quintiles are not the same for the two sets of posterior means. In Figure 2, we show the map of the quintiles of planted acres. We can see some changes in these maps (CA, ID move from 2 to 3; AZ, NM move from 1 to 2; ND moves from 3 to 5; OH moves from 5 to 4, *etc.*). Otherwise, the two maps are mostly similar; however, the quintiles can hide the details, so we will discuss this further.

Figure 4 shows a plot of the posterior coefficients of variation (CVs) of the two models versus those of the observed data for the 48 states. We can see all the points are below the $45^o$ reference line, showing clearly that the two models provide improved reliability. We can also see most of the points corresponding to the non-spatial model are closer to the reference line than those from the spatial model, showing the estimates from the spatial model are more reliable. Those points, where a star and a dot are close together, correspond to the states with very large planted acres such as Iowa.

Figure 3 shows a plot of the posterior coefficients of variation (CVs) of the spatial model versus those of the non-spatial model for the 48 states. We can see all the points, except four (two very close), are below the $45^o$ reference line; the points falling on the reference line correspond to the states with large planted acres. This clearly shows that the spatial model provides improved reliability over the non-spatial model.

For completeness, we also look at the plot of PMs (Figure 5) and PSDs (Figure 6) for the spatial model versus the non-spatial model. For the PMs, it is really good that all of the points, except five of them, are nearly on the $45^o$ straight line through the origin. For the PSDs, it is also good that all of the points, except eight of them (five very close), are below the $45^o$ straight line through the origin. There is one of them in which the PSD is much lower under the ST model.

Integrating the (FSA, ASB) historical data into the models, which accommodate the survey data, appear to be important. Estimating the unknown FSA values in June of the current year is a reasonable thing to do. In general, the spatio-temporal (ST) model is better than the non-spatio-temporal (NST) model. The ST model fits the data better than the NST model. The constraint estimates from the ST model have smaller PCVs than those from the NST model.

These results show that the ST model provides higher precision and is more reliable than the NST model. Also the posterior means of the two models are very similar.

## 5. Concluding remarks

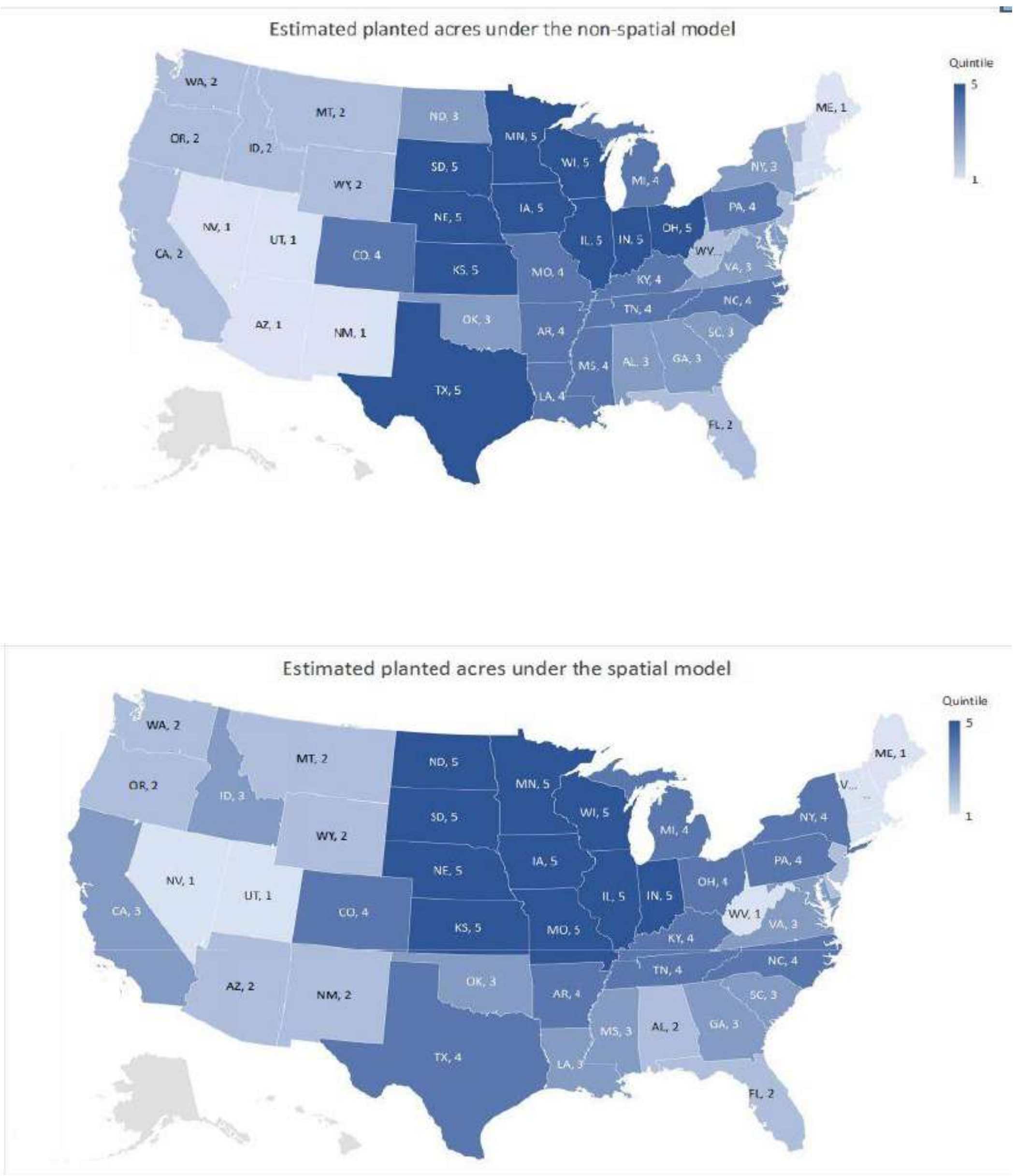
We have shown how to estimate planted acres for US states. This is on-going research and there are rapid changes under way as NASS pursued early estimates of planted acres, as early as June, and this is important for various reasons that we have discussed. As modernization and unification are under way at NASS, data integration is an important activity in this endeavor, and a lot of money and man power are put into it by NASS. Specifically, we have pointed out the struggles to find suitable statistical procedures in the initial stages. We have pointed out many challenges to get early estimates of planted acres and how to overcome some of them. Because of confidentiality, we have not used the real data in this paper, and the results presented may not be appropriate. As clearly described, the real data also have shortcomings, yet this project is extremely important to NASS.

In this paper, we have shown how to integrate a non-probability sample (FSA values) with a probability sample from a dual-frame survey (APS and JAS) to provide early estimates of planted acres for corn. One difficulty encountered is that the model estimates must be

**Table 4: Posterior summaries for planted acres (thousands) under the constraint**

| State | *fp1* | *p2* | *sp2* | *CVp2* | *PM* | *PSD* | *PCV* | *95% HPDI* |
|---|---:|---:|---:|---:|---:|---:|---:|---:|
| a. Non-spatial model | | | | | | | | |
| AL | 293 | 350 | 75 | 0.21 | 376.96 | 53.45 | 0.14 | (293.71, 478.25) |
| AZ | 88 | 95 | 19 | 0.2 | 106.25 | 12.45 | 0.12 | (87.98, 130.31) |
| AR | 733 | 750 | 103 | 0.14 | 818.61 | 62.06 | 0.08 | (733.64, 944.41) |
| CA | 401 | 470 | 161 | 0.34 | 570.15 | 113.02 | 0.2 | (401.17, 784.72) |
| CO | 1418 | 1400 | 646 | 0.46 | 1579.68 | 149.38 | 0.1 | (1418.42, 1879.06) |
| CT | 22 | 26 | 5 | 0.18 | 27.84 | 3.5 | 0.13 | (22.46, 34.41) |
| DE | 166 | 175 | 50 | 0.29 | 206.17 | 30.3 | 0.15 | (166.24, 264.16) |
| FL | 78 | 100 | 48 | 0.48 | 127.1 | 34.25 | 0.27 | (78.27, 194.12) |
| GA | 420 | 460 | 69 | 0.15 | 487.03 | 46.2 | 0.1 | (420.62, 578.79) |
| ID | 342 | 400 | 107 | 0.27 | 455.68 | 72.49 | 0.16 | (346.79, 585.72) |
| IL | 10465 | 11200 | 3511 | 0.31 | 10484.83 | 27.88 | 0 | (10464.58, 10542.13) |
| IN | 4988 | 5400 | 772 | 0.14 | 5028.64 | 51.56 | 0.01 | (4987.77, 5128.44) |
| IA | 12323 | 13100 | 3404 | 0.26 | 12338.54 | 19.79 | 0 | (12322.76, 12376.14) |
| b. Spatial Model | | | | | | | | |
| AL | 293 | 350 | 75 | 0.21 | 355.51 | 34.51 | 0.1 | (293.47, 417.46) |
| AZ | 88 | 95 | 19 | 0.2 | 99.2 | 7.63 | 0.08 | (87.99, 114.07) |
| AR | 733 | 750 | 103 | 0.14 | 792.53 | 41.53 | 0.05 | (733.44, 872.04) |
| CA | 401 | 470 | 161 | 0.34 | 551.69 | 72.56 | 0.13 | (405.20, 677.03) |
| CO | 1418 | 1400 | 646 | 0.46 | 1480.8 | 41.02 | 0.03 | (1420.05, 1538.34) |
| CT | 22 | 26 | 5 | 0.18 | 26.35 | 2.12 | 0.08 | (22.47, 30.22) |
| DE | 166 | 175 | 50 | 0.29 | 188.37 | 16.05 | 0.09 | (166.23, 217.93) |
| FL | 78 | 100 | 48 | 0.48 | 110.13 | 19.61 | 0.18 | (78.32, 144.76) |
| GA | 420 | 460 | 69 | 0.15 | 462.48 | 26.63 | 0.06 | (420.50, 511.14) |
| ID | 342 | 400 | 107 | 0.27 | 406.32 | 40.33 | 0.1 | (343.01, 480.69) |
| IL | 10465 | 11200 | 3511 | 0.31 | 10519.4 | 32.92 | 0 | (10464.97, 10572.29) |
| IN | 4988 | 5400 | 772 | 0.14 | 5066.48 | 45.83 | 0.01 | (4992.59, 5139.01) |
| IA | 12323 | 13100 | 3404 | 0.26 | 12368.6 | 27.1 | 0 | (12322.83, 12411.41) |

NOTE: fp1 is FSA planted acres, p2 is survey indications, sp2 is survey variance and CVp2 is survey coefficient of variation. The constraint specifies the model estimates must be larger than the FSA value.



Estimated planted acres under the non-spatial model

Estimated planted acres under the spatial model

**Figure 2: Quintiles of posterior means of planted acres from the two models with constraint: The quintiles under the non-spatial (spatial) model are** 101 (92), 377 (356), 730 (668) **and** 3346 (3623), **and under the non-spatial (spatial) model, the minimum and maximum values are** 2.68 (2.32) **and** 12339 (12368).

larger than FSA values, which are unknown in June. We have provided two models for these data and we have demonstrated that the spatio-temporal model is a lot better than its non-spatio-temporal counterpart. While important indirect covariates, as used in this paper, are easily available, NASS has been pursuing more direct covariates such as temperature and precipitation, ethanol production capacity, and many others.

One would need to go down to lower level of disaggregation to accommodate variability. At the state level, there are actually a large number of records that go into the single number, thereby making variability relatively small because variance is generally inversely proportional to sample size. When survey indications are weighted up, there are no considerations of heterogeneity such as clustering (*e.g.*, counties) at intermediate levels, so that variability at the state level can be small. Young and Chen (2022) wrote, "Modeling at the state level is not always able to provide predictions of desired quality. Perhaps samples that provide valid estimates at lower geospatial scale should be considered; this will require major revisions in the current sample designs. Alternatively, if survey and non-survey data are linked at the farm level, then modeling could be conducted at that level."

It is now believed that modeling should be done at the level of Agricultural Statistics Districts (ASD); several ASDs might form a state. There are no ASD level survey indications and standard errors in June, so modeling is difficult to impossible at the ASD level; see Appendix A for a method to get ASD data from state data. Only state level indications and standard errors are available in June to NASS. We have been using the state level data to project backwards to the ASDs and the number counties within each ASD is used as the sample sizes (these are not presented) to get a rough idea of the indications and variances at the ASD level. A non-spatial model similar to the one discussed here is fit to the ASD level data, but now we need both an ASD level effect and a state level effect (so called sub-area or two-fold model). This will provide better state level model estimates. However, it is difficult to operationalize this model. At the ASD level, the NST model and the ST model are discussed in great detail in Nandram (2023), but again this second report is confidential. In addition, one may want to benchmark the states to the entire United States, but this is not attempted here. See Nandram, Ericulescu and Cruze (2019) for recent work on benchmarking.

A further problem of practical importance is the clustering of data at the state level, ASD level or county level. Many projects at NASS operates at county level such as cash rental rates and yield. The clustering does not have to be at geographical levels. For example, it does not have to be the case that the counties within a state have to form a cluster. Some counties in one state may be clustered with counties in another state. That is, there are unseen clusters among the sampling units (*e.g.*, counties), and these must be taken into consideration to avoid understating variability and biased estimates. Currently, this is on-going research activity in the Research and Development Division at NASS. Attempts are being made to accommodate this research activity for planted acres using the stick-breaking priors (Ishwaran and James, 2001); see Appendix B.

## Disclaimer, compliments, acknowledgements

1. The findings and conclusions in this paper are those of the author and should not be

construed to represent any official USDA or US Government determination or policy.

2. The author thanks Neomi Guidin, Grace Yoon, Lu Chen, Ashley Lockwood, Valbona Bejeleri, Lindsay Drunasky, Andrew Dau, Chris Hawthorn and Linda Young for ongoing discussions at the National Agricultural Statistics Service (NASS).

3. The research methodology was supported by USDA's NASS.

4. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

5. The paper was not written as part of Balgobin Nandram's normal duty at NASS, and a huge attempt was made to avoid confidential issues as much as possible.

6. Finally, I am grateful to the Organizing Committee, particularly Professor Vinod Gupta with his strong belief in me, for the invitation to present this Keynote Address. This was, indeed, a great honour for me.

## References

Berg, E., Im, J., Zhu, Z., Lewis-Beck, C., and Li, J. (2021). Integration of statistical and administrative agricultural data from Namibia. *Statistical Journal of the IAOS*, **37**, 557–578, DOI: 10.3233/SJI-200634.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Second Edition. California: Duxbury, ISBN: 0-534-24312-6.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011–2021, DOI: 10.1080/01621459.2019.1677241.

Chen, L., Nandram, B., and Cruze, N. (2022), Hierarchical Bayesian models with inequality constraints for US county estimates. *Journal of Official Statistics*, **38**, 709–732, DOI: 10.2478/jos-2020-0004.

Fay, R. E. and Herriot R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277, DOI: 10.2307/2286322.

Fuller, W. A. (1987),. *Measurement Error Models.* New York: Wiley, ISBN: 0-471-86187-1.

Goyal, S., Datta, G. S., and Mandal, A. (2020). A hierarchical Bayes unit-level small area estimation model for normal mixture populations. *Sankhya B*, S1-S27, DOI: 10.1007/s13571-019-00216-8.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–172.

Janicki, R., Raim, M. A., Scott, H. H., and Maples, J. J. (2022). Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations. *Annals of Applied Statistics*, **16**, 144–168, DOI: 10.1214/21-AOAS1494.

Kansas Farm Bureau (2020). *Counting crop acres NASS and WAOB vs. FSA*, 1–4.

Liu, J. S. (1994). The Collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966, DOI: 10.1080/01621459.1994.10476829.

Nandram, B. (2022). Temporal modeling of planted acres with spatial effects and covariates, *Technical Report*, Research Development Division, National Agricultural Statistics Service, USDA, 1–44.

Nandram, B. (2023). Temporal modeling of planted acres for Agricultural Statistics Districts with spatial effects and covariates, *Technical Report*, Research Development Division, National Agricultural Statistics Service, USDA, 1–77.

Nandram, B., Cruze, N. B., Erciulescu, A. L., and Chen, L. (2022). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. *RDD Research Report, Number RDD-22-02, National Agricultural Statistics Service, USDA*, 1–41.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of non-probability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 5–21, 10.5539/ijsp.v10n6p5.

Nandram, B. and Rao, J. N. K. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. *In JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 1568–1603.

Nandram, B. and Rao, J. N. K. (2023). Bayesian predictive inference when integrating a non-probability sample and a probability sample. *arXiv:2305.08997V1 [Stat.ME] 15 May 2023*, 1-35.

Nandram, B., Cruze, N. B., and Erciulescu, A. L. (2023). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. *Survey Methodology* (in press), 1–30.

Nandram, B., Erciulescu, A. L., and Cruze, N. B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, **45**, 365–390.

National Academies of Sciences, Engineering, and Medicine (2017). *Improving Crop Estimates by Integrating Multiple Data Sources.* The National Academies Press: Washington, DC, USA, 2017.

Office of the Chief Economist (2019). *Update of 2019 FSA acreage data and FAQs on USDA acreage*, United States Department of Agriculture (USDA). Robert Johansson (Chief Economist) and Ashley Hungerford (Economist) for questions, August 27, 2019, pg. 0–7.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 3–33, DOI: 10.1007/s13571-020-00227-w.

Ridgeway, J. (2016). Computation of Gaussian orthant probabilities in high dimension. *Statistics and Computing*, **26**, 899–916, DOI: 10.1007/s11222-015-9578-1.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs Sampler. *Journal of the American Statistical Association*, **87**, 861–868, DOI: 10.2307/2290225.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics*, **18**, 861–878, DOI: 10.1198/jcgs.2009.08153.

The National Agricultural Statistics Service (NASS) (2021 a). *Prospective Planting*, Agricultural Statistical Board, United States Department of Agriculture (USDA). Approved by Seth Meyer and Joseph L. Parson, March 31, 2021, ISSN: 1949-159X, pg. 1–36.

The National Agricultural Statistics Service (NASS) (2021 b). *Acreage*, Agricultural Statistical Board, United States Department of Agriculture (USDA). Approved by Seth Meyer and Joseph L. Parson, June 30, 2021, ISSN: 1949-1522, pg. 1–48.

Wang, J. C., Scott, H. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 84–106, DOI: 10.107/513253-011-0067-5.

Young, L. J. and Chen, L. (2022). Using small area estimation to produce official statistics. *Stats*, **5**, 881–897, DOI: 10.3390/stats5030051.

## APPENDIX A

### How to get ASD level data from state level data?

We consider a simple change of support (COS) analysis and we assume (normality is not required) that

$$\hat{\theta}_{ij} \overset{ind}{\sim} \text{Normal}(\theta_{ij}, \hat{\sigma}_{ij}^2), j = 1, \ldots, n_i, i = 1, \ldots, \ell,$$

where $n_i$ is the number of ASDs in the $i^{th}$ state (larger states have more ASDs). Let $m_{ij}$ denote the number of counties in the $j^{th}$ ASD. We do not know the $\hat{\theta}_{ij}$ and $\hat{\sigma}_{ij}^2$. However, note that $\sum_{j=1}^{n_i} \hat{\theta}_{ij} = \hat{\theta}_i$ and $\sum_{j=1}^{n_i} \hat{\sigma}_{ij}^2 = \hat{\sigma}_i^2$ (assuming independence).

Specifically, we assume that $\hat{\theta}_{ij} \propto m_{ij}$, and this gives

$$\hat{\theta}_{ij} = \left\{ \frac{m_{ij}}{\sum_{j=1}^{n_i} m_{ij}} \right\} \hat{\theta}_i, j = 1, \ldots, n_i, i = 1, \ldots, \ell.$$

We also assume that $\hat{\sigma}_{ij}^2 \propto {m_{ij}}^{-1}$, and this gives

$$\hat{\sigma}_{ij}^2 = \left\{ \frac{m_{ij}^{-1}}{\sum_{j=1}^{n_i} m_{ij}^{-1}} \right\} \hat{\sigma}_i^2, j = 1, \ldots, n_i, i = 1, \ldots, \ell.$$

Both of these imputation procedures are reasonable because bigger states (i.e., planted acres of corn) will have larger $\hat{\theta}_i$ and smaller $\hat{\sigma}_i^2$.

Historical data, FSA values and ASB estimates, are available at county level. However, FSA values for the current year in June are not available and a similar procedure can be performed on the state values. Covariates can be used at the state level or jittered to get ASD level covariates. NASS will need to put in a large effort to get the covariates at the ASD level.

# APPENDIX B

## Basic stick-breaking distribution

For planted acres, the stick-breaking distribution for state estimates, $\hat{\theta}_i, i = 1, \ldots, \ell$, is

$$f(\hat{\theta}_i - \theta_i \mid \theta_i) = \sum_{s=1}^{\ell_o} p_s \text{Normal}(z_s, \hat{\sigma}_i^2), \ell_o \leq \ell,$$

where, given the $\theta_i$, the $\hat{\theta}_i - \theta_i$ are independent and identically distributed, the $p_s$ are stick-breaking weights, the $z_s$ are a random sample from a baseline distribution, and $\ell_o$ (unknown) is the number of clusters; see Ishwaran and James (2001). Therefore, it is true that

$$f(\hat{\theta}_i \mid \theta_i) = \sum_{s=1}^{\ell_o} p_s \text{Normal}(\theta_i + z_s, \hat{\sigma}_i^2), \ell_o \leq \ell,$$

and, given the $\theta_i$, the $\hat{\theta}_i$ are now independent, not identically distributed.

Introducing latent variables, this can be rewritten as

$$f(\hat{\theta}_i, d_i) = \prod_{s=1}^{\ell_o} [p_s \text{Normal}(\theta_i + z_s, \hat{\sigma}_i^2)]^{I(d_i = s)}, \ell_o \leq \ell,$$

where $d_i$ maps the $i^{th}$ state into a cluster and $I(d_i = s)$ is the indicator function.

Here the stick-breaking weights are

$$p_1 = \nu_1, p_s = \nu_s \prod_{r=1}^{s-1} (1 - \nu_r), s = 2, \ldots, \ell_o - 1, \ldots, p_{\ell_o} = \prod_{s=1}^{\ell_o - 1} (1 - \nu_s),$$

and for the two-parameter Pitman-Yor process, we use the prior,

$$\nu_s \overset{ind}{\sim} \text{Beta}\{1 - \delta_1, \frac{1 - \delta_2}{\delta_2} + (s - 1)\delta_1\}, s = 1, \ldots, \ell_o, 0 < \delta_1, \delta_2 < 1.$$

As for the $z_s$, we take

$$z_s \overset{ind}{\sim} \text{Normal}\{0, \frac{\rho}{1 - \rho} \sigma^2\}, s = 1, \ldots, \ell_o, 0 < \rho < 1.$$

It is also possible to assume a stick-breaking prior on the $\theta_i$.

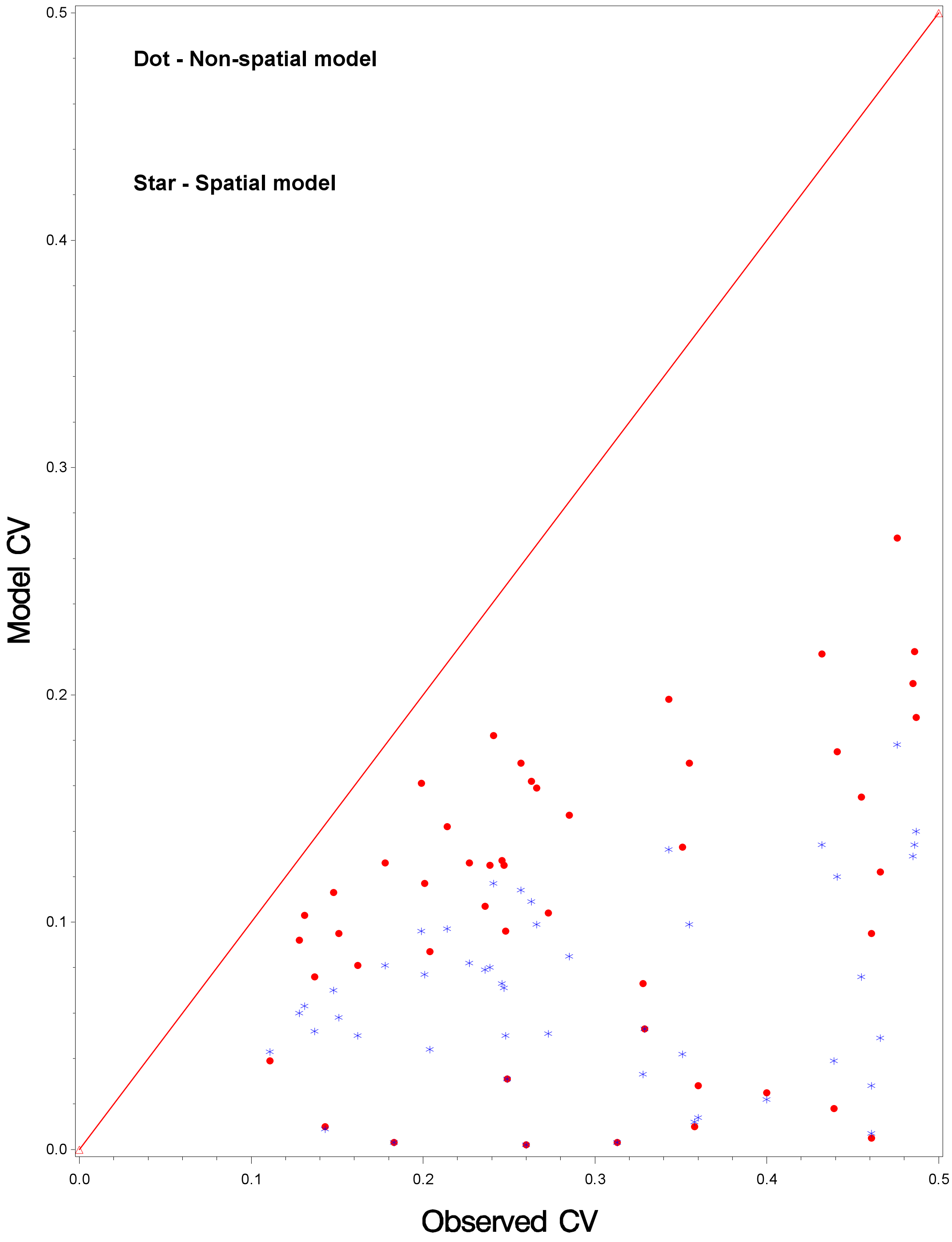


**Figure 3: Plots of the CVs of the two models versus the CVs of the observed data**

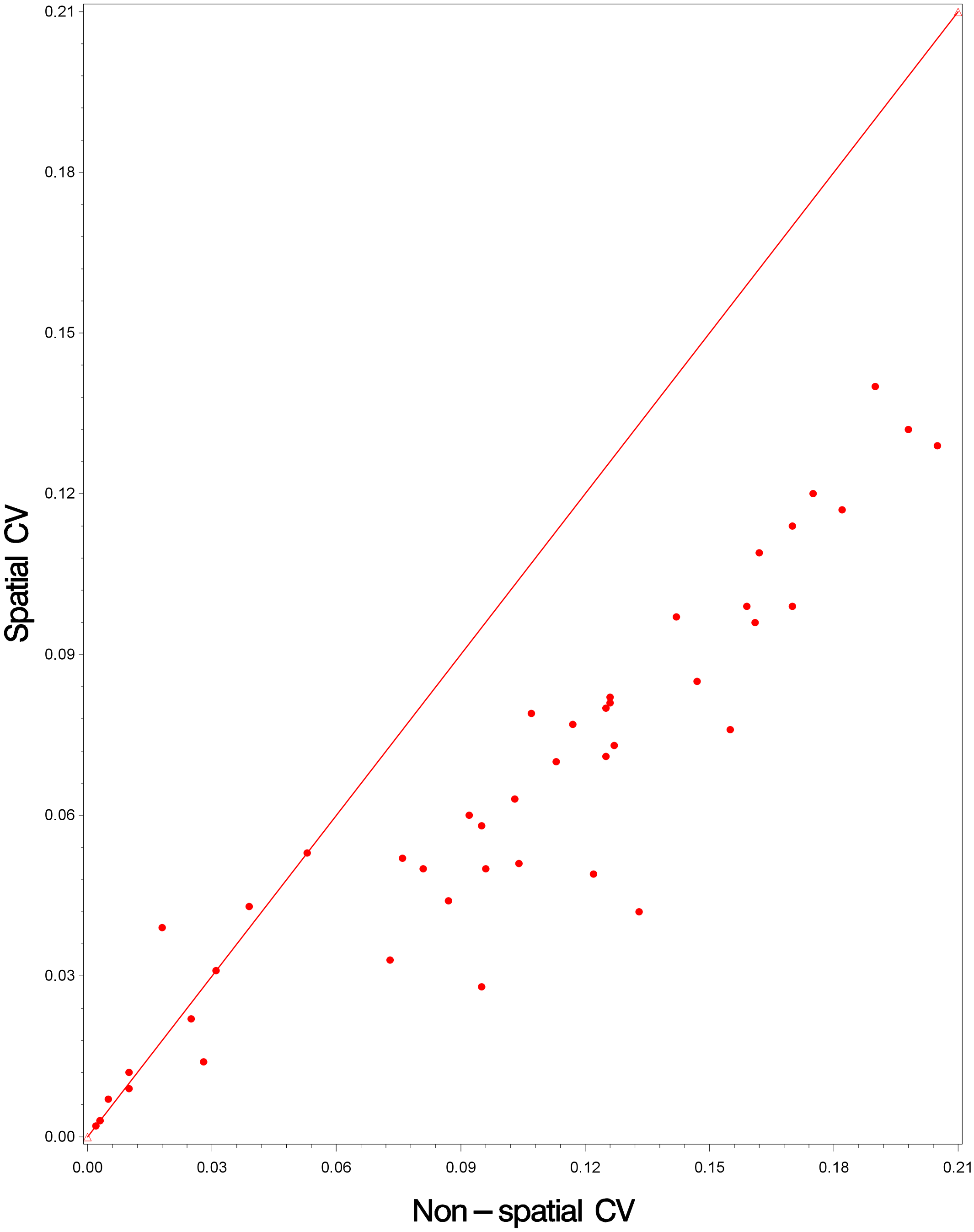


**Figure 4: Plots of the CVs of the spatial model versus those of the non-spatial model**

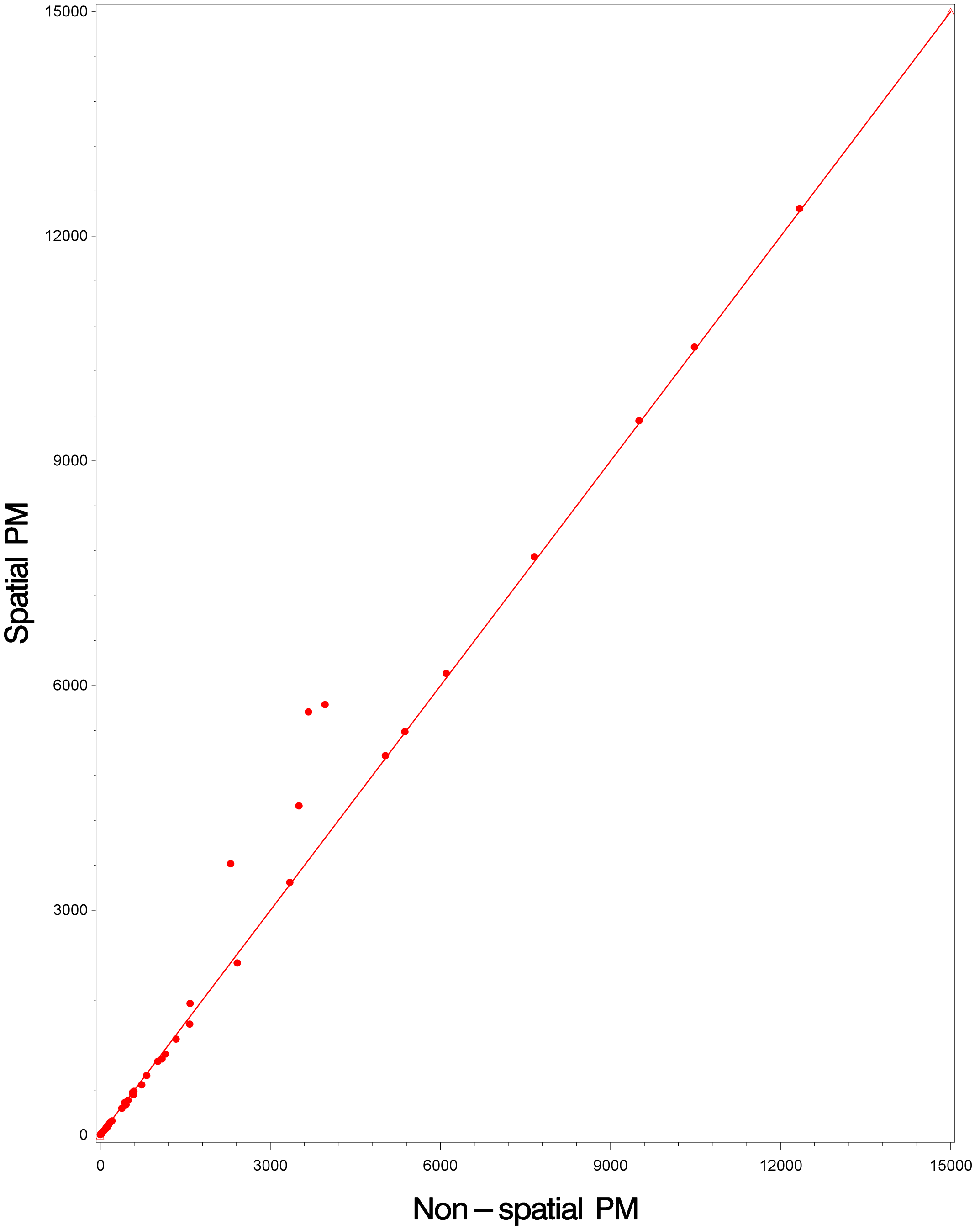


**Figure 5: Plots of the PMs of the spatial model versus those of the non-spatial model**

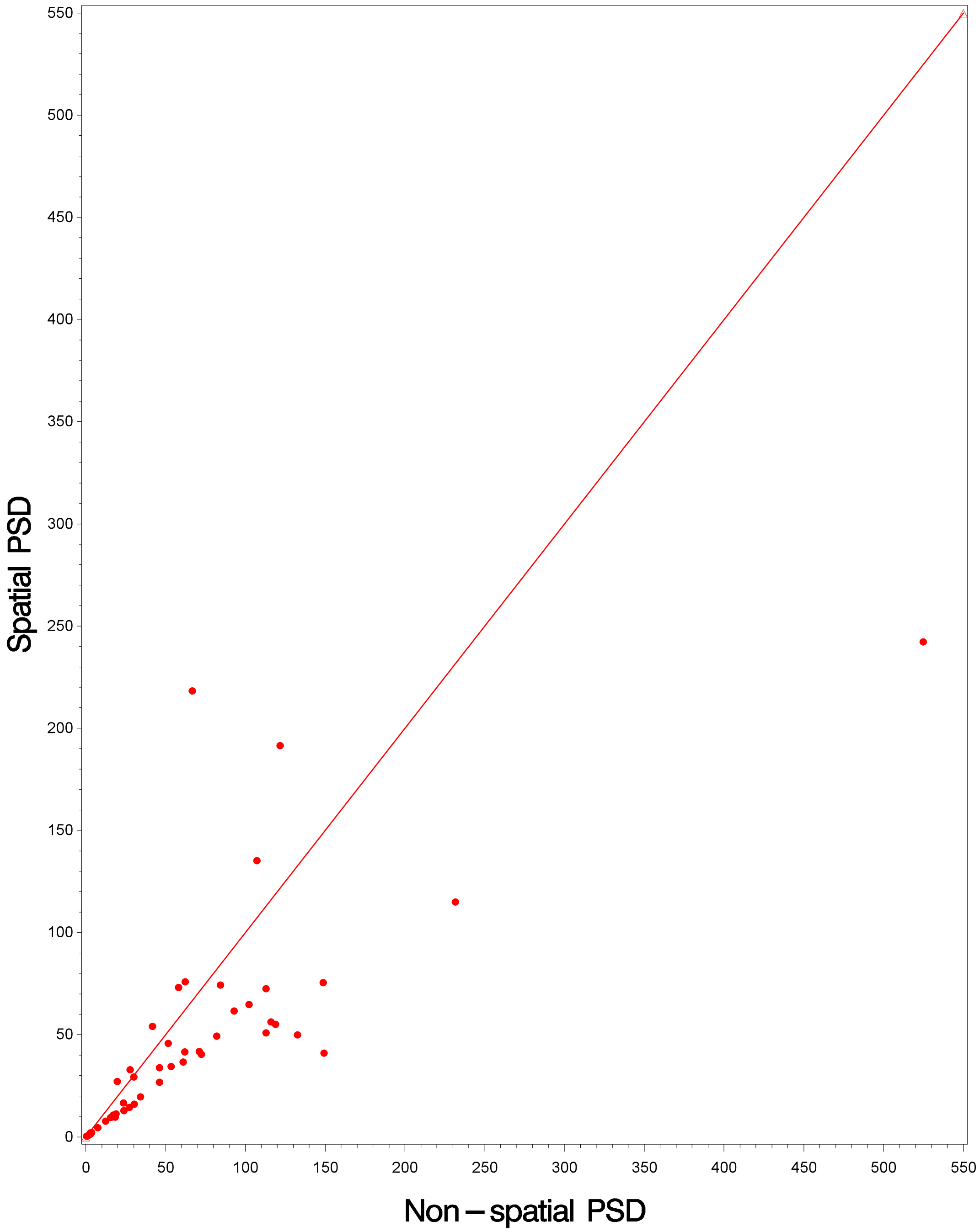


**Figure 6: Plots of the PSDs of the spatial model versus those of the non-spatial model**

# APPENDIX C

## A list of useful abbreviations

| Abbreviations | Meanings |
|---|---|
| USDA | United States Department of Agriculture |
| NASS | National Agricultural Statistics Service |
| FSA | Farm Service Agency |
| RDD | Research Development Division |
| ASB | Agricultural Statistics Board |
| APS | Agricultural Production Survey |
| JAS | June Area Survey |
| ASD | Agricultural Statistics District |

NOTE: NASS and FSA are two of the agencies of USDA, and RDD is a division of NASS. APS and JAS are the two surveys. All estimates are approved by the ASB before publication.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 79-92

# A New Development of Threshold Selection in Extreme Value Theory

**K. M. Sakthivel and V. Nandhini**
*Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu.*

Received: 15 June 2023; Revised: 21 July 2023; Accepted: 25 July 2023

## Abstract

Extreme value theory addresses the stochastic behavior of the extreme values in a process. There are two important methods used in modeling extreme value analysis and they are threshold selection and block maxima techniques. The threshold selection is important in many aspects of statistical inference of extreme or rare events because they use data more effectively than block maxima techniques. The inference derived from the threshold method mainly depends on the selection of the optimum threshold and it can be determined approximately using the parameter stability plot and mean residual life plot. Since the extreme value theory considers only extreme values in the given set of data. So there is an unresolved issue in determining the optimal threshold while using the peaks over threshold technique. Further exceedances above a high threshold have been shown to asymptotically follow the generalized Pareto distribution under the usual circumstances. In this paper, a new development in threshold selection technique is discussed in detail for modeling extreme values along with real-life applications.

*Key words*: Extremes; Tail behavior; Peaks over threshold; Block maxima; Return level.

**AMS Subject Classifications:** 62E20, 62M10.

## 1. Introduction

Extreme Value Theory (EVT) is a specialized field of statistics that provides methodologies and tools for the study and estimation of probabilities of events that have not been previously observed or rare events. Because these extreme events are sparse, extrapolation beyond the observed levels is required for estimation. EVT is designed explicitly for such extrapolation and utilizes asymptotic analyses as the foundation of extreme value models. This theory indicates that extreme value estimation is only related to the tail of the probabilistic distribution. The objective of extreme value analysis is to determine how likely it is that certain events will occur that are the least likely to have previously been observed. The techniques and models have been developed to describe the tails of the data and estimate the probabilities of extreme events.

Correponding Author: K.M. Sakthivel
Email: sakthithebest@buc.edu.in

In the literature on extreme value theory, several researchers have applied these special techniques in different real scenarios of the countries to come up with several estimates about extreme events. EVT has two commonly used approaches block maxima and peaks over threshold approaches. In block maxima (BM), the peak value of each block is considered an extreme value, and block sizes are usually taken on year. But in some cases, the block size may vary depending on the nature of the study areas. Such extreme values from block maxima also known as annual or cluster maxima, these values can be modeled using the generalized extreme value (GEV) distribution which came from the first theorem of EVT that is Fisher Tippett Gendenko theorem (1928). The peaks over a threshold (PoT) method is popular in the extreme value analysis because it prefixes the threshold for the whole observations, the values above the threshold are considered as the extreme values of the specific cases. The generalized Pareto (GP) distribution can be used to model the extreme value sequence from the PoT method. This GP distribution originated from the second theorem of EVT called the Pickands-Balkema-De Haan theorem (1975). However, some other methods can also be available, r-largest order statistic with GEV and point process (PP) with GP approaches.

So, according to Coles (2001), EVT can simulate the stochastic nature of processes involving events of unusually high or low intensity. Pickands (1975) proposed a method for making decisions about the upper tail of the distribution. It can be used to predict the likelihood that future extremely large observations. And GP distribution can be introduced to model extreme values. Smith (1989) proposed specific modifications based on the point-process view of high-level exceedances via a clustering approach with ozone data analysis. Davison and Smith (1990) talked about modeling the sizes and occurrences of exceedances in order to analyze data extremes. Katz *et al.* (2002) explained the evolution of extremes, which includes the development of a point process framework that incorporates block maxima and PoT techniques. Sanders (2005) shows the modeling of extreme events is becoming of increased importance to actuaries. Cooley (2011) investigated the definitions of return period and return level given by Olsen *et al* (1998) the $m$-year return level was the level for which the expected waiting time until the exceedances in $m$-years and Parey *et al.* (2007) was the $m$-year return level as the level for which the expected number of events in an $m$-year period is one can be considered under the nonstationary setting. Deidda (2010) introduced a multiple threshold method (MTM) to infer the parametes by using excess over the threshold applying againt the concepts of parameters threshold invariance, and also discussed the supremacy of the MTM fit against the single threshod fit. Scarrot and Mac-Donald (2012) developed the parameter stability plot, with an emphasis on estimating the shape and scale parameters in order to determine an appropriate threshold. De Zea *et al.* (2012) employed the PoT method to model the sample of excesses above a sufficiently high value of total cholesterol level of patients. Bader *et al.* (2018) developed the automated sequential threshold via ordered goodness of fit tests with adjustment for false discovery rate. Roux *et al.* (2020) studied the trends in 50 years' return levels of the ground snow loads using non-stationary extreme value models for the French Alps with its building standards. Hesarkazzazi *et al.* (2021) investigated the process of non-stationary annual maxima of river peak flow in northwest England and a regression model for the location parameter of the generalized logistic distribution (GLO) was also constructed. Tanprayoon *et al.* (2023) proposed a new Gompertz-generalized extreme value distribution for extreme value analysis and return-level estimation of the extreme rainfall.

In some sectors of science and technology, the extreme values of significant variables have special meanings and importance. The extreme value theory has recently been applied in terrestrial and solar climatology. The sunspot number series, which was recorded from 1818 to 2022, is used to study extreme values of solar activity. The observations of daily sunspot numbers have been collected from the database of Solar Influences Data Analysis Center (SIDC) - the solar physics research department of the Royal Observatory of Belgium. Sunspots are dark, planet-sized areas that appear on the surface of the sun. Sunspots are magnetic regions on the sun with magnetic field strengths thousands of times greater than the Earth's magnetic field. Sunspots appear in active regions, usually in opposite magnetic polarity pairs. Their number varies with the roughly 11-year solar cycle. Sunspot magnetic fields are extremely strong, keeping heat away from these regions of the sun's surface. The active region is a temporary region with a strong and complex magnetic field in the sun. They are often associated with sunspots and can be a source of eruptions like solar flares and coronal mass ejections (CMEs). Solar flares are a burst of energy caused by the tangling, crossing, or reorganization of magnetic field lines near sunspots. The variation in the number of sunspots and solar activity are closely related. Because solar activity can have an impact on Earth, scientists closely monitor it every day.

When sunspot counts are high, the sun is very active, and the peak in the sunspot count is referred to as a solar maximum, whereas a period when fewer or no sunspots appear is referred to as a solar minimum. Sunspots can cause geomagnetic storms in the Earth's magnetosphere. When sunspot numbers are at their peak during the solar maximum period, the sun emits more radiation than usual. A solar flare emits a large amount of radiation into the universe. Intense solar flares can interfere with radio waves, telecommunications, the electric power grid, and satellite navigation by releasing radiation that interferes with these systems. Therefore, due to the high number of sunspots in the sun's photosphere, there is a chance that solar flares and coronal mass ejections will appear. In this case, extreme value analysis is essential to find out the extreme occurrences of the sunspot number during the solar maximum period of this current solar cycle. The extreme values of previous events of sunspots decide the behavior of the future event of the study. Acero *et al.* (2017) used the block maxima method with the GEV distribution for modeling the maximum values of the sunspot numbers at yearly, monthly, and daily scales for each solar cycle and the PoT approach only for daily scales, which takes into account all sunspot numbers that exceed a predefined upper threshold and can be modeled using the GP distribution. The return levels were predicted for 10 (110 years), 50 (550 years), and 100 (1100 years) solar cycles. Elvidge *et al.* (2018) used EVT to investigate the likelihood of extreme solar flares with both GOES X-ray flux data and Kepler mission data.

In this paper, we are interested to develop a new threshold selection methodology that is superior to the existing PoT method. The sunspot numbers data set is used for this theory to estimate the return levels associated with the return periods, as well as to calculate the probability of exceedances. It is therefore essential to study and model these extremes to make accurate prognostications of return levels. As a result, new approaches for predicting extreme occurrences can be developed and they can be modeled with GP distribution in application to sunspot number series. This paper has been divided into five sections. Following this introduction, Section 2 presents research materials and methodologies, Section 3 performs preliminary data analysis, Section 4 describes the interpretation of the results, and Section 5 shows a summary and conclusion.

## 2. Methodology

In this section, we will discuss the procedures for both the traditional and proposed threshold selection methodologies for segregating extreme values from a series of observations. Those extreme values can be modeled with appropriate distribution to predict future events by the return period concept for this case.

### 2.1. Peaks over threshold

The PoT method were created by Pickands (1975) and it concentrates on observations that seem to go above a high threshold. The PoT with GP distribution can be used to avoid the problem of data waste, which is a common problem with the block maxima method. However, determining an appropriate threshold is an inherent problem specifically. If the threshold has been too low, the tail will satisfy the less convergence criterion, causing a large bias and an incorrect result. If the threshold is too high, however, very few values above the threshold will result in high variance and imprecise results. Thus, selecting an appropriate threshold necessitates balancing the bias and the variance.

The GP distribution has a continuous range of possible shapes, including special cases of the exponential and Pareto distributions. We can use either of these to model a specific set of exceedances. The two-parameter GP distribution with shape parameter $\xi$ and scale parameter $\sigma$ has the following representation.

The cumulative distribution function of the two-parameter GP distribution with a shape parameter $\xi$, the scale parameter $\sigma$ is given by

$$F(x|\sigma,\xi)=\begin{cases} 1-[1+\xi(\frac{x}{\sigma})]^{-\frac{1}{\xi}}; & \text{for} \quad \xi\neq 0\\ 1-\exp\{-[\frac{x}{\sigma}]\}; & \text{for} \quad \xi=0\end{cases} \tag{1}$$

where, $x>0$ when $\xi>0$ and $0\leq x\leq -\sigma/\xi$ when $\xi\leq 0$. and corresponding probability density function is

$$f(x|\sigma,\xi)=\begin{cases} \frac{1}{\sigma}[1+\xi(\frac{x}{\sigma})]^{-\frac{\xi-1}{\xi}}; & \text{for} \quad \xi\neq 0\\ \frac{1}{\sigma}\exp\{-[\frac{x}{\sigma}]\}; & \text{for} \quad \xi=0\end{cases} \tag{2}$$

If $\xi>0$, the above equation reduces to Pareto distribution, which is a heavy-tailed distribution. If $\xi=0$ it is reduced to the exponential distribution. If $\xi<0$ it is simply to obtain light-tailed distribution with finite endpoints such as short-tailed Pareto or uniform distribution. The mean and variance of a distribution is given by

$$E(x|\sigma,\xi)=\frac{\sigma}{1+\xi} \text{ and}$$

$$V(x|\sigma,\xi)=\frac{\sigma^2}{(1+\xi)^2(2\xi+1)} \text{ exists if } \xi>-1, \xi>-\tfrac{1}{2} \text{ respectively.}$$

### 2.2. Reduced threshold - a new approach

The newly developed reduced threshold (RT) method divides the entire set of observations into equal-sized non-overlapping periods and focuses on the extreme values in these periods. These extreme values are taken into account when determining the threshold point. When compared to the traditional BM and PoT method, the extreme values above this particular threshold point are considered special extreme values. There are '$m$' numbers of observations in each of the '$n$' periods. Therefore there is $m \times n$ number of total observations.

Let us consider the blocks $B_{ij}$ for $j = 1, 2, ..., k; i = 1, 2, ..., n$, where $i$ represent the position of each block consisting of $j$ is the number of independent and identically distributed observations. The maximum values of every block are considered extreme values which are represented as the following sequence,

$$Z_i = \{z_1, z_2, ..., z_k\}; \text{for every } i = 1, 2, ..., n$$

Let $Z_i$ be the sequence of iid random variables with CDF $F(z)$ and let $\xi_p$ denoted by $p^{th}$ quantile of $F$, so that $\xi_p = inf\{z|F(z) \geq p\}$. The $p^{th}$ quantile is defined as $F(\xi_p) = p$. Let $Q_p = Z_{(i)\lfloor np \rfloor:n}$ denote a sample $p^{th}$ quantile. Here $\lfloor np \rfloor$ denotes the greater integer $\leq np$. The weighted average of the distribution's median and quantiles $Q_p$ and $Q_{1-p}$ for $p \in (0, 1/2)$ is known as the ”Trimean estimator”, such that

$$\hat{\tau} = \frac{\alpha}{2}Q_p + (1-\alpha)Q_{1/2} + \frac{\alpha}{2}Q_{1-p} \tag{3}$$

The weights for the two quantiles are the same for $Q_p$ and $Q_{1-p}$, and the weight $\alpha \in [0, 1]$. The Tukey's Trimean estimator is obtained by taking $\alpha = \frac{1}{2}$ and $p = \frac{1}{4}$ in the above equation and it is a special case of the Trimean estimators. It can be defined as

$$\hat{\tau}_{TM} = \ \tfrac{1}{4}Q_{1/4} + \tfrac{1}{2}Q_{1/2} + \tfrac{1}{4}Q_{3/4} \tag{4}$$

The threshohd $u^*$ can be obtained by $\hat{\tau}_{TM}$ and first quartile of the extreme value sequence of iid's.

$$u^* = \frac{1}{2}[\hat{\tau}_{TM} + Q_{1/4}] \tag{5}$$

when the values exceed the threshold $u$ then as the special extreme values denote $Z_s^*$ for $s = 1, 2, ..., k$.

$$Z_s^* = \{z_1^*, z_2^*, ..., z_k^*\}, \text{for} Z_s^* \geq u, Z_s^* \neq Z_i. \tag{6}$$

These special extreme values from the RT method can be modeled with generalized Pareto distribution.

## 2.3. Tail dependence and declustering

In stationary sequences, extreme values can occur in clusters. The first step in making inferences is to identify clusters in the data, which is accomplished through the declustering process. Declustering could be effective at screening the dependent observation to a set of threshold exceedances. The empirical rule can be used to define the cluster of exceedances, and the maximum excess in each cluster can be determined. Runs and interval methods can be used in such cases to separate the clusters and estimate the extremal index. The extremal index $\theta$ is a measure of the degree of local dependence in the extremes of a stationary process it ranges from 0 to 1, that is $\theta \in [0,1]$, where to imply some dependence. When $\theta$ the value decreases there is evidence for greater dependence. In Runs declustering, a run length (the minimum gap between clusters) '$r$' can be fixed to choose the cluster, and the extremes are separated by fewer than '$r$'-non extremes belonging to the same cluster. The choice of '$r$' is critical because too small a value causes the problem of independence being unrealistic for nearby clusters, while too large a value causes the concatenation of clusters that could reasonably be considered independent, potentially resulting in the loss of valuable data.

### 2.4. Return periods and return levels

The return levels and return periods, which are crucial for the prediction of extreme events, can be discovered when the distribution is fitted. The return level predicts that the event will occur at least once over the following '$t$' years. For the GP model, the return level is given $x_q$ which defines the extreme level that exceeds at least once every '$q$' observations. The return period of the GP model is

$$P(X > x|X > u) = [1 + \xi(\tfrac{x-u}{\sigma})]^{-\frac{1}{\xi}}$$

Let $\xi_u = P(X > u) = r/n$, where $r$ is the number of upper order values exceeding the threshold $u$, and $n$ is the number of years of records then the return period can be simplified as follows

$$P(X > x) = \left[1 + \xi(\tfrac{x-u}{\sigma})\right]^{-\frac{1}{\xi}}$$

This implies that the data points exceed once in every '$m$' series of observations on average can be determined as

$$\xi_u \left[1 + \xi(\tfrac{x-u}{\sigma})\right]^{-\frac{1}{\xi}} = \tfrac{1}{m}$$

Finally, the $m$-year return level for GPD is given by

$$x_m = \begin{cases} u - \frac{\sigma}{\xi}[(m\xi_u)^{\xi} - 1]; & \xi \neq 0 \\ u - \sigma \log[m\xi_u]; & \xi = 0 \end{cases} \tag{7}$$

where $x_m$ is the return level associated with the return period $q = 1/m$.

The return level is the interesting final product of the extreme value analysis in the prediction of tail probabilities. Therefore, when '$m$' should be large enough, the return level $x_m$ exceeds the threshold '$u$'.

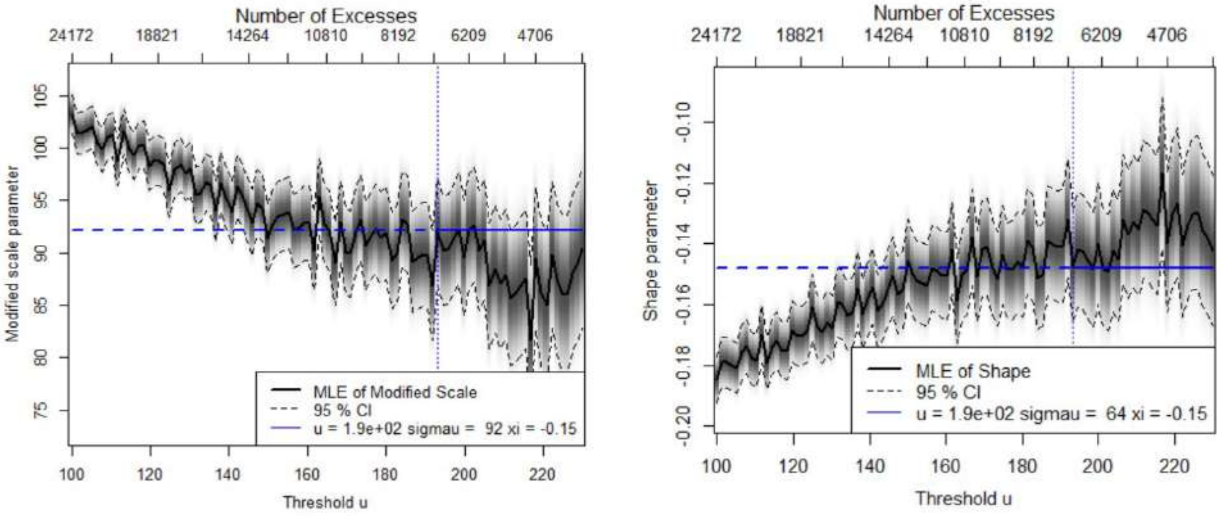
## 3. Application

### 3.1. Data source

The daily sunspot numbers dataset spans 205 years which is more than two centuries, beginning in January 1818 and ending in December 2022. The sunspot number daily of observations have been collected from the database of Solar Influences Data Analysis Center (SIDC) - the solar physics research department of the Royal Observatory of Belgium. It is publicly available on SIDC's Sunspot Index and Long-term Solar Observations (SILSO) website.

### 3.2. Data modeling and analysis

The excess over the threshold technique can be used to separate the extreme sunspot number from the non-extremes by selecting the appropriate threshold. The two approaches discussed above can be used to identify extreme sunspots, and the exceedances can be modeled with an appropriate distribution, which can then be used to estimate the return levels. The PoT approach takes into consideration only those sample sunspot number values that are significantly larger than a predetermined threshold $u$. The scale parameter of the distribution can be modified that is $\sigma^{*}= \sigma - \xi u$ against the threshold $u$ which has been shown in the modified scale parameter threshold stability plot in Figure 1. In this figure, the dark black line represents the estimated parameter value and the shaded area describes its confidence level for $u = 190$, the vertical line represents the threshold value, and the horizontal line shows the estimated parameter value, the threshold stability plot can be used to determine an appropriate threshold. The mean residual life (MRL) plot is another alternative way of choosing an appropriate value of the threshold which is shown in Figure 2. It plots an average value over a given threshold for a series of thresholds. A mean excess plot with a downward-sloping line indicates thin-tailed behavior. The MRL plot shows the mean number of excesses over the threshold $u$, in between a confidence interval (approx 95%). We look for approximate linearity (from the lowest possible threshold) whilst keeping in between the confidence bounds.



**(a) Modified scale parameter** **(b) Shape parameter**

**Figure 1: Threshold stability plot**

The RT is a new technique for determining an adaptive threshold $u$, particularly for dependence sequences. In this series of sunspot numbers, the RT approach uses trimean $\hat{\tau}_{TM}$

and $Q_{1/4}$ finds the threshold $u^*$ from the yearly maximum extreme value series, sometimes also known as annual maxima. Since maximum values of sunspot numbers are grouped into clusters, one should expect that there may be several consecutive days with the maximum exceeding the threshold. To avoid short-term dependencies in the time series, these expected clusters of exceedances would necessitate the use of a declustering procedure to identify approximately independent clusters of extreme observations within the sample.

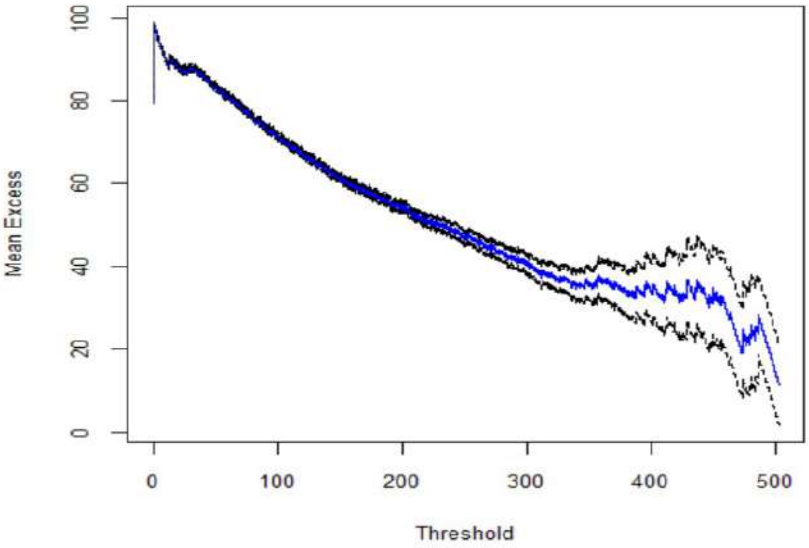


**Figure 2: Mean residual life plot**

The runs declustering process involves grouping exceedances into the same cluster if their distance from one another is less than the predetermined run length $r$. The extremal index for sunspot number from PoT threshold 190 is 0.01399, demonstrating the sequence's strong dependence. When 103 clusters are above the threshold and the appropriate run length is $r$=69, the maximum values of each cluster are regarded as extreme values. The weak dependence in the sequence is indicated by the associated extremal index, which is 0.8844 respectively. Figure 3 depicts the declustered sunspot numbers data using the runs method of declustering. The horizontal line in the figure represents the $u$=190 line over the years, and the values above the threshold are considered extreme values. These values are declustered to form 103 clusters, from which the higher order values for this study are taken. The new declustering series data can be modeled using the GP distribution. Maximum likelihood estimation is used to estimate the parameters. The value of the parameter estimates with its standard error for scale $\sigma = 74.0544(11.7214)$ and shape $\xi = 0.0215(0.1238)$. The variance-covariance matrix of the parameters for the peaks over threshold associated with GP is given as follows

$$CV = \begin{bmatrix} 137.3902 & -1.12069 \\ -1.12069 & 0.01534 \end{bmatrix}$$

The diagonals of the matrix are the variance for the fitted model. The 95% CI for the parameters scale has (51.0811, 97.0279) and shape has (-0.22132, 0.2642) respectively.

Figure 4 depicts the declustered sunspot numbers data using the runs method of declustering. The horizontal line in the figure represents the $u^*$=162 line over the years and

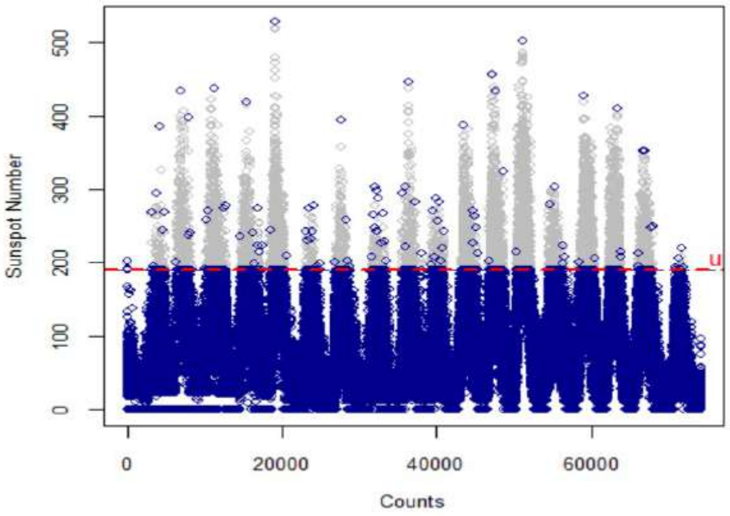


**Figure 3: Declustering series plot for PoT**

highlighted values above the threshold are considered extreme values of each cluster. The extremal index for RT is 0.01108 for the obtained threshold, demonstrating the severe dependence in sequence. In this case, the appropriate run length is $r$=55, with 126 clusters above the threshold, and the maximum values of each cluster are considered to be the extreme values. The associated extremal index is 0.9023 which describes the weak dependence in the sequence. The GP distribution can be used to model the new

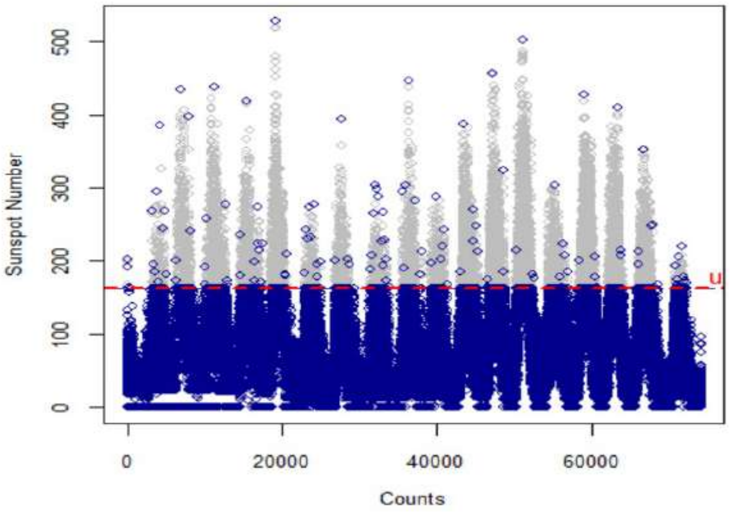


**Figure 4: Declustering series plot for RT**

declustering series data. The parameters are estimated using the maximum likelihood estimation. The value of the parameter estimates with its standard error for scale $\sigma = 62.70(9.3351)$ and shape $\xi = 0.1427(0.1202)$ respectively. The variance-covariance matrix of the parameters for the reduced threshold associated with GP is given by

$$CV = \begin{bmatrix} 90.0442 & -0.81475 \\ -0.81475 & 0.01367 \end{bmatrix}$$

The diagonals of the matrix are the variance of the parameters for the fitted model. The estimate $\xi > 0$ indicates that its domain of attraction is the Pareto (heavy-tailed) distribution. The 95% CI for the parameters scale has (44.2846, 80.8806) and the shape has (-0.09223, 0.38014) respectively.

## 4. Model diagnostics and return levels

In this section, we examine the results of two models: peaks over the threshold with GP distribution and reduced threshold with GP distribution. The model is chosen using the goodness of fit tests such as Anderson-Darling, Cramer-von Mises, and Kolmogorov-Smirnov, and the results are shown in Table 1. The models are ranked based on their performance, and return levels for various return periods are computed. Table 1 presents the goodness of fit test statistic as well as the $p$-value for the models under consideration in this study. The model with the least statistic is ranked with the lowest value in the range for each measure; among these, the reduced threshold-GP distribution shows a reasonable fit for this dataset.

**Table 1: Results of goodness of fit tests**

| **Models** | $A_n^2$ (**$p$-value**) | $W_n^2$ (**$p$-value**) | $D_n$ (**$p$-value**) |
|---|---|---|---|
| PoT-GP | 0.5664 (0.6800) | 0.0603 (0.8133) | 0.0587 (0.8695) |
| RT-GP | 0.4961 (0.7504) | 0.0392 (0.9381) | 0.0469 (0.9444) |

**Table 2: Estimated return levels with 95% confidence interval**

| **RP(yr)** | **PoT-GP** | **RT-GP** |
|---|---|---|
| 2023 | 139 (115, 163) | 133 (117, 148) |
| 2024 | 190 (176, 205) | 175 (164, 186) |
| 2025 | 221 (204, 237) | 202 (188, 216) |
| 2026 | 242 (223, 261) | 222 (205, 239) |

Table 2, above shows the estimated return level of the sunspot numbers for the maximum period of the current $25^{th}$ solar cycle, from 2023 to 2026. The solar maximum is expected to occur between 2024 and 2026. According to a NASA report, scientists anticipate a rise in solar activity leading up to the next maximum, which could occur in 2025. Usually, the maximum period is unknown because no one can predict it precisely. No one knows when the sun's polarities change precisely; it cannot happen at a precise time, but it does happen over an approximate period. We predicted the sunspot number for 2023 to 2026 as well, because the exact maximum period has not been exactly predicted by scientists, if the maximum period of the current cycle will extend to 2026, this prediction may be useful.

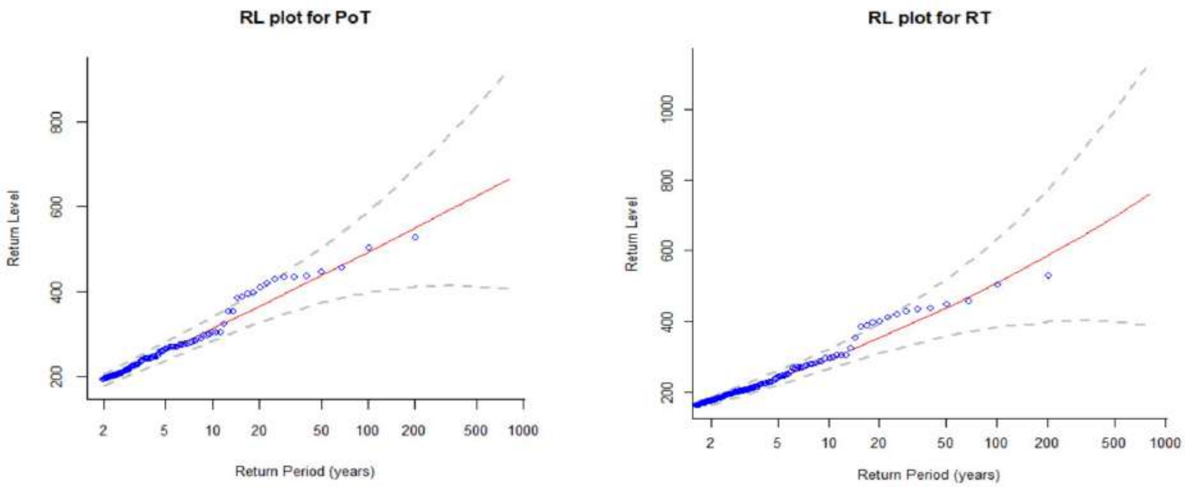


**Figure 5: Return Level Plot for PoT and RT**

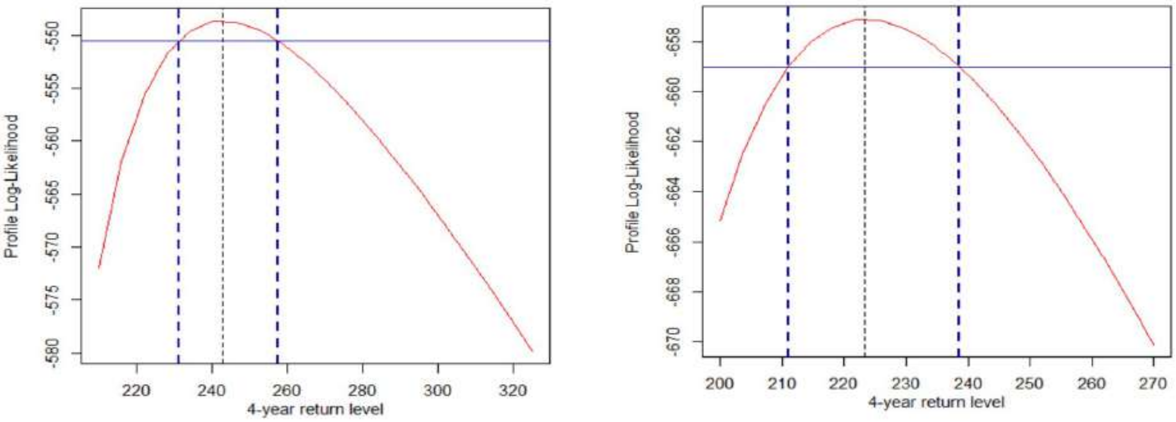


**Figure 6: Profile likelihood plot for PoT and RT**

The return level can be estimated by the delta method, for PoT is 242 with a 95% confidence interval is (223, 262), for RT is 223 with a 95% confidence interval is (206, 240) respectively. The return levels are graphically represented in Figure 5; it shows the estimated values associated with its confidence interval. When the time increases the estimated values of the sunspot numbers of the solar maximum period also increase with its confidence limits. The return levels can also be obtained using the profile likelihood method, we get the estimated value of the 4-year return level for the PoT method is 242 and its approximate 95% confidence interval is (223, 262), the estimated value of the 4-year return level for RT method is 223 and its approximate 95% confidence interval is (211, 238). Figure 6 displays the profile likelihood for the 4-year return level. Because the profile likelihood crosses both the blue vertical dashed and horizontal solid lines, the resulting intervals are believable. We select the RT method because it provides a better fit for the extreme values than the PoT method. The estimated return levels of sunspot numbers based on the RT method are thus taken into account.

## 5. Summary and conclusions

This study developed a new reduced threshold method for selecting a suitable threshold, and the generalized Pareto distribution can be used to model the extreme values. In the context of extreme value analysis, we applied the peaks over threshold and reduced threshold techniques to the sunspot number series from 1818 to 2022 years to establish the decision threshold. A generalized Pareto with shape and scale parameters can be used to model exceedances above the threshold. The behavior of the extreme value series is described by the estimated extreme value index. The shape parameter of the aforementioned two methods are positive in this situation, indicating that the distribution is heavily tailed according to the series. This study focuses on the maximum period of sunspots in the solar cycle because there is a possibility of solar flares and CMEs occurring during that period. The $m$-year return levels were estimated for the $25^{th}$ solar cycle's maximum periods, such as 2023-2026. This prediction could be helping to determine the next maximal event. We will discuss only the rare event rather than all occurrences because it will be more useful to observe the tail behavior with less probability. In this study, we explore how the peaks over threshold and the reduced threshold can be used to estimate the model's tail parameters. Among these, our suggested model has the narrowest return level confidence interval. The goodness of fit test can be used in conjunction with this study to evaluate the models and precision.

## References

Acero, F. J., Carrasco, V. M. S., Gallego, M. C., García, J. A., and Vaquero, J. M. (2017). Extreme value theory and the new sunspot number series. *The Astrophysical Journal*, **839**, 98.

Acero, F. J., Gallego, M. C., García, J. A., Usoskin, I. G., and Vaquero, J. M. (2018). Extreme value theory applied to the millennial sunspot number series. *The Astrophysical Journal*, **853**, 80.

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Annals of Applied Statistics*, **12**, 310–329.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). An introduction to statistical modeling of extreme values. *Springer.*

Cooley, D. (2012). Return periods and return levels under climate change. *In Extremes in a changing climate: Detection, analysis and uncertainty, Dordrecht: Springer Netherlands.*, **839**, 97–114.

Davison, A. C., and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, **52**, 393–425.

De Haan, L. (1976). Sample extremes: An elementary introduction. *Statistica Neerlandica*, **30**, 161–172.

Deidda, R. (2010). A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series. *Hydrology and Earth System Sciences*, **14**, 2559–2575.

De Zea Bermudez, P., and Mendes, Z. (2012). Extreme value theory in medical sciences: Modeling total high cholesterol levels. *Journal of Statistical Theory and Practice*, **6**, 468–491.

Elvidge, S. and Angling, M. J. (2018). Using extreme value theory for determining the probability of Carrington-like solar flares. *Space Weather*, **16**, 417–421.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *In Mathematical proceedings of the Cambridge philosophical society, Cambridge University Press.*, **24**, 180–190.

Gumbel, E. J. (1958). Statistics of Extremes. *Columbia University Press, New York.*

Hamdi, Y., Duluc, C. M., and Rebour, V. (2018). Temperature extremes: estimation of non-stationary return levels and associated uncertainties. *Atmosphere*, **9**, 129.

Hall, P. and Weissman, I. (1997). On the estimation of extreme tail probabilities. *Annals of Statistics*, 1311–1326.

Hesarkazzazi, S., Arabzadeh, R., Hajibabaei, M., Rauch, W., Kjeldsen, T. R., Prosdocimi, I., ... and Sitzenfrei, R. (2021). Stationary vs non-stationary modelling of flood frequency distribution across northwest England. *Hydrological Sciences Journal*, **66**, 729–744.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, 1163–1174.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29**, 339–349.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, **27**, 251–261.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, **25**, 1287–1304.

Leadbetter, M. R. and Rootzen, H. (1988). Extremal theory for stochastic processes. *The Annals of Probability*, **16**, 431–478.

Le Roux, E., Evin, G., Eckert, N., Blanchet, J., and Morin, S. (2020) Non-stationary extreme value analysis of ground snow loads in the French Alps: a comparison with building standards. *Natural Hazards and Earth System Sciences*, **20**, 2961–2977.

Olsen, J. R., Lambert, J. H., and Haimes, Y. Y. (1998). Risk of extreme events under nonstationary conditions. *Risk Analysis*, **18**, 497–510.

Parey, S., Hoang, T. T. H., and Dacunha-Castelle, D. (2010). Different ways to compute temperature return levels in the climate change context. *Environmetrics*, **21**, 698–718.

Parey, S., Malek, F., Laurent, C., and Dacunha-Castelle, D. (2007). Trends and climate evolution: Statistical approach for very high temperatures in France. *Climatic Change*, **81**, 331–352.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**, 119–131.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, **4**, 367–377.

Sanders, D. E. A. (2005). The modelling of extreme events. *British Actuarial Journal*, **11**, 519–557.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, **10**, 33–60.

Soukissian, T. H. and Arapi, P. M. (2011). The effect of declustering in the r-largest maxima model for the estimation of Hs-design values. *The Open Ocean Engineering Journal*, **4**, 34–43.

Shaikh, Y. H., Khan, A. R., Iqbal, M. I., Behere, S. H., and Bagare, S. P. (2008). Sunspots data analysis using time series. *Fractals*, **16**, 259–265.

Tawn, J. A. (1988). An extreme-value theory model for dependent observations. *Journal of Hydrology*, **101**, 227–250.

Tanprayoon, E., Tonggumnead, U., and Aryuyuen, S. (2023). A New Extension of Generalized Extreme Value Distribution: Extreme Value Analysis and Return Level Estimation of the Rainfall Data. *Trends in Sciences*, **20**, 4034–4034.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 93–106

# Statistical Inference for Fréchet Distribution Based on Dual Generalized Order Statistics

**Bhagwati Devi[1], Qazi J. Azhad[2], Ankita Sharma[3] and Ayush Tripathi[1]**
[1] *Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan, India*
[2] *Department of Mathematics, Shiv Nadar University, Greater Noida, Uttar Pradesh, India*
[3] *Government Maulana Azad Memorial P.G. College, Jammu, Jammu and Kashmir, India*

Received: 08 July 2023; Revised: 24 July 2023; Accepted: 28 July 2023

## Abstract

This article discusses the idea of an ordered random variable and its basic structure. Under the umbrella of dual generalized order statistics, the problem of Bayesian estimation of Fréchet distribution with parameters $\alpha$ and $\lambda$ is addressed. Both symmetric (squared error) and asymmetric (linear exponential and general entropy) loss functions are taken into account to enable flexibility in the outcomes. For the aim of estimation, two approximation methods (Lindley and Markov Chain Monte Carlo) have been employed and presented. Simulation tools have been used to elaborate the findings clearly.

*Key words:* Fréchet distribution; Dual generalized order statistics; Bayesian methods; Markov chain Monte Carlo.

**AMS Subject Classifications:** 62C10, 62F10, 62F15, 62G30

## 1. Introduction

While dealing with data analysis using statistical tools and techniques, extreme value theory is inevitable. Jenkinson (1955) described how Generalized Extreme Value (GEV) is the most preferred distribution in this regard. The cumulative density function is given by

$$F(y \mid \sigma, \mu, \xi) = \begin{cases} \exp\left(-[1+\xi(y-\mu)/\sigma]_+^{-1/y}\right\}, & \text{for } \xi \neq 0 \\ \exp(-\exp[-(y-\mu)/\sigma]], & \text{for } \xi = 0 \end{cases}$$

where $\sigma > 0, \mu, \xi \in \mathbb{R}$. The considered distribution in this manuscript is Fréchet which is a special cases of GEV distribution. Its name spawned from Maurice René Fréchet, a French mathematician, who developed this distribution in 1920 as a maximum value distribution. It is also known as the extreme value distribution of type II.

The probability density function (PDF), cumulative density function (CDF) and reliability function of the random variable y following Fréchet distribution are given as

$$f(\mathrm{y} \mid \lambda, \alpha) = \lambda\alpha y^{-(\alpha+1)} e^{-\lambda y^{-\alpha}}, \tag{1}$$

Corresponding Author: Bhagwati Devi
Email: bhagwatistats@gmail.com

$$F(\text{y} \mid \lambda, \alpha) = e^{-\lambda y^{-\alpha}}, \tag{2}$$

$$R(\text{t} \mid \lambda, \alpha) = 1 - e^{-\lambda t^{-\alpha}}. \tag{3}$$

where $y > 0$, $t > 0$, $\alpha > 0$ is shape parameter and $\lambda > 0$ is the scale parameter. Depending on the form of parameters, the PDF might be unimodal or declining, although the hazard function is always unimodal. This is the only CDF that can be established on non-negative real numbers and is also a limiting CDF for the maxima of random variables. For a range of engineering applications, this feature is crucial for simulating the issues associated with investigating the statistical behavior of material properties.

It was explained by Kotz and Nadarajah (2000) that how Fréchet distribution can be used in a variety of contexts, including accelerated life testing, natural disasters, horse racing, rainfall, grocery store lines, sea currents, wind speeds, track race records and so on. Harlow (2002) demonstrated that the Fréchet distribution is the best option for simulating the case where high values are crucial. The literature on Fréchet distribution is extensive. Maximum likelihood estimation has been performed by Calabria and Pulcini (1989), and the features of its estimator (MLE) have been studied. Maximum likelihood estimation was carried out by Ramos *et al.* (2017) in the presence of the cure fraction, and Loganathan and Uma (2017) compared the MLE, the LSE, the weighted LSE, and the method of moment estimation for the Fréchet distribution. In order statistics, the Fréchet distribution was investigated by Salman and AMER (2003), while generalised order statistics was researched by Maswadah (2003). Many scholars have also addressed the issue of Bayesian estimate for the Frechet distribution. For instance, Calabria and Pulcini (1994) and Kundu and Howlader (2010) have performed Bayesian estimation using Gamma or other informative or arbitrary priors. Fréchet distribution was examined using Jeffreys and reference priors in Abbas and Tang (2015).

After carefully searching the literature, we were unable to locate any articles addressing its application to order statistics or lower record data. Therefore, utilizing the setup of Dual Generalised Order Statistics (*dgos* ), we have addressed the Bayesian estimation of the Fr'echet distribution. The manuscript is arranged as follows: Mathematical formulation of *dgos* is thoroughly discussed in Section 2. Also, in this section, Bayesian framework for estimation using different loss functions is given. Bayes estimators are obtained using the Lindley approximation, a method for approximation that is detailed in Section 3. Bayes estimators are obtained in Section 4 using Markov chain Monte Carlo approach. Simulation analysis for *dgos* submodels such as order statistics and lower record values is provided in Section 5 along with conclusions regarding the obtained results.

## 2. Formulation of Bayesian framework

Let us take independent and identically distributed sequence containing $X_1, X_2, \ldots$ random variables having absolutely continuous distribution function $F(\cdot)$ and the probability density function $f(\cdot)$. Let $n \in \mathbb{N}$, $(n \geq 2)$, $k \geq 1$ and $m$ be the parameters such that $\gamma_r = k + (n-r)(m+1) > 0$, for all $r \in \{1, 2, \ldots, n-1\}$ and $Y(1, n, m, k), Y(2, n, m, k), \ldots, Y(n, n, m, k)$ be the $n$ *dgos*. Then the joint density function of $Y_1, Y_2, \ldots, Y_n$ is of the form

$$k \left( \prod_{j=1}^{n-1} \gamma_j \right) \left( \prod_{i=1}^{n-1} (F(y_i))^m f(y_i) \right) (F(y_n))^{k-1} f(y_n), \tag{4}$$

where $F^{-1}(1) > y_1 \geq y_2 \geq \cdots \geq y_n > F^{-1}(0)$, $Y_i = Y(i, n, m, k)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is the realization of $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$.

The *dgos* is a combination of many ordered random models, and we may create different models by accounting for different *dgos* model characteristics. For instance, when $m = 0$ and $k = 1$ are used, the *dgos* model reduces to reverse order statistics; when $m = -1$ is used, the *dgos* model reduces to the $k^{th}$ lower record values; and when $m = -1$ and $k = 1$ are used, the *dgos* model reduces to standard lower record values; *etc.* The following books and articles are suggested for readers who want to learn more about ordered statistics and record data: Ahsanullah (2004), Arnold *et al.* (2008), Devi *et al.* (2017), Arshad and Jamal (2019a,b), Sharma *et al.* (2019) Arshad and Baklizi (2019), Tripathi *et al.* (2019), Gupta and Jamal (2019), Anwar *et al.* (2020) and Azhad *et al.* (2021, 2022, 2023).

Now, let $Y_1, Y_2, \ldots, Y_n$ be the $n$ *dgos* drawn from Fréchet$(\alpha, \lambda)$, then by using equation (4), equation (1) and equation (2), the likelihood function is given as

$$L(\alpha, \lambda|\boldsymbol{y}) = k(\alpha\lambda)^n \left(\prod_{j=1}^{n-1} \gamma_j\right) \left(\prod_{i=1}^{n} y_i^{-(\alpha+1)} e^{-\lambda y_i^{-\alpha}}\right) \prod_{i=1}^{n-1} \left(e^{-\lambda y_i^{-\alpha}}\right)^m \left(e^{-\lambda y_n^{-\alpha}}\right)^{k-1}. \tag{5}$$

Assuming that informative priors are independent and have a two-parameter gamma distribution with the following set of hyperparameters, we now investigate informative priors for each parameter.

$$\left.\begin{aligned} \pi(\alpha) &= \frac{b_1^{a_1}}{\Gamma(a_1)} \alpha^{a_1-1} e^{-b_1\alpha}, \quad a_1,\ b_1,\ \alpha > 0, \\ \pi(\lambda) &= \frac{b_2^{a_2}}{\Gamma(a_2)} \lambda^{a_2-1} e^{-b_2\lambda}, \quad a_2,\ b_2,\ \lambda > 0. \end{aligned}\right\} \tag{6}$$

We take into account symmetric and asymmetric loss functions to demonstrate the adaptability of our findings and to provide a wide range of applicability for diverse real-life scenarios. The symmetric loss function is taken into consideration since it equally penalises underestimation and overestimation, which are typically highly helpful. The majority of the time, nevertheless, we observe that positive losses can sometimes be more severe than negative losses, and vice versa. Asymmetric loss functions are necessary in these circumstances. Here, we have taken into account one symmetric loss function, the squared error loss function (SELF), as well as two asymmetric loss functions, the linear exponential (LINEX) and general entropy (GE). For more details about these loss function, one may refer to Jaheen (2003), Dey (2009), Ali (2015), Zhang and Gui (2020), Nagamani *et al.* (2020). The SELF is defined as

$$L_1(\delta, \beta) = (\delta - \beta)^2, \quad \beta > 0. \tag{7}$$

The Bayes estimator under SELF is posterior mean $(\delta_{SEL})$. The LINEX loss function is defined as

$$L_2(\delta, \beta) = e^{c(\delta-\beta)} - c(\delta - \beta) - 1, \qquad c \neq 0 \tag{8}$$

with corresponding Bayes estimator as

$$\delta_{LINEX} = -\frac{1}{c} \ln\left(E(e^{-c\beta})\right).$$

The GE loss function is given as

$$L_3(\delta,\beta) \propto \left(\frac{\delta}{\beta}\right)^c - c\ln\left(\frac{\delta}{\beta}\right) - 1, \qquad c \neq 0 \tag{9}$$

with corresponding Bayes estimator as

$$\delta_{GE} = \left[E(\beta)^{-c}\right]^{-1/c}.$$

Now, the joint posterior density of $\alpha$ and $\lambda$ is obtained by using equation (5) and equation (6), and is given as

$$\pi(\alpha,\lambda|\boldsymbol{y}) \propto \alpha^{n+a_1-1}\lambda^{n+a_2-1}\left(\prod_{j=1}^{n-1}\gamma_j\right)\left(\prod_{i=1}^{n} y_i^{-(\alpha+1)}e^{-\lambda y_i^{-\alpha}}\right)\prod_{i=1}^{n-1}\left(e^{-\lambda y_i^{-\alpha}}\right)^m\left(e^{-\lambda y_n^{-\alpha}}\right)^{k-1} \tag{10}$$
$$e^{(-b_1\alpha-b_2\lambda)}; \alpha > 0, \lambda > 0.$$

Joint posterior density has a complex structure, making it difficult to construct exact Bayes estimators. Lindley approximation and the Markov chain Monte Carlo approach are two extensively used approximation techniques that are used to address this scenario.

## 3. Lindley approximation

Using the Taylor series expansion, Lindley (1980) estimated the ratio of the two integrals. The expectation of posterior densities can be calculated using this method to a reasonable extent. Typically, a Bayes estimator takes the following form for any loss function of the $\beta$ parameter:

$$E(z(\beta)|\boldsymbol{y}) = \frac{\int z(\beta)e^{\mathbb{L}(\beta)+\rho(\beta)}d\beta}{\int e^{\mathbb{L}(\beta)+\rho(\beta)}d\beta}, \tag{11}$$

where $\mathbb{L}$ denotes the logarithm of likelihood function, logarithm of the prior distribution of $\beta$ is denoted by $\rho$. In present case $\beta = (\alpha,\lambda)$, we can transform equation (11) to

$$E(z(\alpha,\lambda)|\boldsymbol{y}) = \frac{\int\int z(\alpha,\lambda)e^{\mathbb{L}(\alpha,\lambda)+\rho(\alpha,\lambda)}d\alpha d\lambda}{\int\int e^{\mathbb{L}(\alpha,\lambda)+\rho(\alpha,\lambda)}d\alpha d\lambda}, \tag{12}$$

The values of the quantities in above equation are $\mathbb{L}(\alpha,\lambda) = \ln L(\alpha,\lambda|\boldsymbol{y})$ and $\rho(\alpha,\lambda) = \ln\pi(\alpha) + \ln\pi(\lambda)$. Utilizing the method by, we get (see Lindley (1980))

$$E(z(\alpha,\lambda)|\boldsymbol{y}) \approx z(\alpha,\lambda) + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} z_{ij}\sigma_{ij} + \sum_{i=1}^{2}\rho_i Q_i \tag{13}$$
$$\frac{1}{2}\sum_{i=1}^{2}\mathbb{L}_{iii}\sigma_{ii}Q_i + \frac{1}{2}\left[\mathbb{L}_{112}(2\sigma_{12}Q_1 + \sigma_{11}Q_2) + \mathbb{L}_{122}(\sigma_{22}Q_1 + 2\sigma_{12}Q_2)\right],$$

where,

$$\left.\begin{aligned}
&z_1=\frac{\partial z(\alpha,\lambda)}{\partial\alpha},\ z_2=\frac{\partial z(\alpha,\lambda)}{\partial\lambda},\ z_{11}=\frac{\partial^2 z(\alpha,\lambda)}{\partial\alpha^2},\ z_{22}=\frac{\partial^2 z(\alpha,\lambda)}{\partial\lambda^2},\ z_{12}=\frac{\partial^2 z(\alpha,\lambda)}{\partial\alpha\partial\lambda}=z_{21},\\
&\mathbb{L}_{11}=\frac{\partial^2\ln L(\alpha,\lambda|\boldsymbol{y})}{\partial\alpha^2},\ \mathbb{L}_{22}=\frac{\partial^2\ln L(\alpha,\lambda|\boldsymbol{y})}{\partial\lambda^2},\ \mathbb{L}_{112}=\frac{\partial^3\ln L(\alpha,\lambda|\boldsymbol{y})}{\partial\alpha^2\partial\lambda},\ \mathbb{L}_{111}=\frac{\partial^3\ln L(\alpha,\lambda|\boldsymbol{y})}{\partial\alpha^3},\\
&\mathbb{L}_{222}=\frac{\partial^3\ln L(\alpha,\lambda|\boldsymbol{y})}{\partial\lambda^3},\ \rho_1=\frac{\partial\rho(\alpha,\lambda)}{\partial\alpha},\ \rho_2=\frac{\partial\rho(\alpha,\lambda)}{\partial\lambda},\ Q_r=\textstyle\sum_{j=1}^{2} z_j\sigma_{rj}
\end{aligned}\right\}\tag{14}$$

and $\sigma_{rj}$ denotes $(r,j)^{th}$ element of the inverse of matrix $[-\mathbb{L}_{ij}]$. For obtaining Bayes estimator, we have to calculate all the unknown values in equation (13) by using the MLES of $\alpha$ and $\lambda$.

We have deduced the unknown quantities in equation (14) as per our problem. These are :

$$\left.\begin{aligned}
\mathbb{L}_{11}=&\ -\frac{n}{\alpha^2}-(k-1)\lambda(\ln y_n)^2y_n^{-\alpha}-m\lambda\textstyle\sum_{i=0}^{n-1}(\ln y_i)^2y_i^{-\alpha}-\sum_{i=0}^{n}\lambda(\ln y_i)^2y_i^{-\alpha}\\
\mathbb{L}_{12}=&\ -(k-1)(\ln y_n)^2y_n^{-\alpha}-m\textstyle\sum_{i=0}^{n-1}(\ln y_i)^2y_i^{-\alpha}-\sum_{i=0}^{n}(\ln y_i)^2y_i^{-\alpha}\\
\mathbb{L}_{22}=&\ -\frac{n}{\lambda^2},\ \mathbb{L}_{222}=\frac{2n}{\lambda^3},\ \mathbb{L}_{122}=0,\ \rho_1=\frac{a_1-1}{\alpha}-b_1,\ \rho_2=\frac{a_2-1}{\lambda}-b_2\\
\mathbb{L}_{111}=&\ \frac{2n}{\alpha^3}+(k-1)\lambda(\ln y_n)^3y_n^{-\alpha}-m\lambda\textstyle\sum_{i=0}^{n-1}-(\ln y_i)^3y_i^{-\alpha}-\sum_{i=0}^{n}-\lambda(\ln y_i)^3y_i^{-\alpha}
\end{aligned}\right\}\tag{15}$$

According to the defined loss functions, we have derived the quantities required. It is evident that except $z(\alpha,\lambda)$ and its derivatives, all the other quantities are same.

We know that the posterior mean is the Bayes estimator in SELF. So, Bayes estimator of $\alpha$, is obtained using

$$z(\alpha,\lambda)=\alpha,\ z_1=1,\ z_2=0=z_{11}=z_{12}=z_{22}=z_{21}.$$

Similarly, the quantities

$$z(\alpha,\lambda)=\lambda,\ z_2=1,\ z_1=0=z_{11}=z_{12}=z_{22}=z_{21}.$$

are used for Bayes estimator of $\lambda$,
and,

$$\begin{aligned}
&z(\alpha,\lambda)=1-e^{-\alpha(t^{-\lambda}-1)},\ z_1=-e^{-t^{-\alpha}\lambda}t^{-\alpha}\lambda\ln t,\ z_2=e^{-t^{-\alpha}\lambda}t^{-\alpha},\\
&z_{11}=e^{-t^{-\alpha}\lambda}t^{-2\alpha}\left(t^{\alpha}-\lambda\right)\lambda(\ln t)^2,\ z_{22}=-e^{-t^{-\alpha}\lambda}t^{-2\alpha},\\
&z_{12}=-e^{-t^{-\alpha}\lambda}t^{-2\alpha}\left(t^{\alpha}-\lambda\right)\ln t=z_{21}.
\end{aligned}$$

are used for Bayes estimator of $R(t)$.

Under LINEX loss function, following quantities are used for the Bayes estimator of $\alpha$ and $\lambda$, respectively

$$z(\alpha,\lambda)=e^{-c\alpha},\ z_1=-ce^{-c\alpha},\ z_{11}=c^2e^{-c\alpha},\ z_2=0=z_{12}=z_{21}=z_{22}.$$

$$z(\alpha,\lambda)=e^{-c\lambda},\ z_2=-ce^{-c\lambda},\ z_{22}=c^2e^{-c\lambda},\ z_1=0=z_{12}=z_{21}=z_{11}.$$

and for $R(t)$, the used quantities are:

$$\begin{aligned}
&z(\alpha,\lambda)=e^{-c\left(1-e^{-\lambda t^{-\alpha}}\right)},\ z_1=ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\lambda\ln t\\
&z_{11}=-ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\lambda(\ln t)^2+ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\lambda\ln t\\
&\qquad\times\left(t^{-\alpha}\lambda\ln t+ce^{-t^{-\alpha}\lambda}t^{-\alpha}\lambda\ln t\right)\\
&z_2=-ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\\
&z_{22}=-ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\left(-t^{-\alpha}-ce^{-t^{-\alpha}\lambda}t^{-\alpha}\right)\\
&z_{12}=ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\ln t+ce^{-c\left(1-e^{-t^{-\alpha}\lambda}\right)-t^{-\alpha}\lambda}t^{-\alpha}\left(-t^{-\alpha}-ce^{-t^{-\alpha}\lambda}t^{-\alpha}\right)\lambda\ln t=z_{21}.
\end{aligned}$$

Similarly, in case of GE loss function, Bayes estimator of $\alpha$ can be obtained by the following quantities

$$z(\alpha,\lambda)=\alpha^{-c},\ z_1=-c\alpha^{-c-1},\ z_{11}=c(c+1)\alpha^{-c-2},\ z_2=0=z_{12}=z_{21}=z_{22}.$$

For Bayes estimator of $\lambda$, we have,

$$z(\alpha,\lambda)=\lambda^{-c},\ z_2=-c\lambda^{-c-1},\ z_{22}=c(c+1)\lambda^{-c-2},\ z_1=0=z_{12}=z_{21}=z_{11}.$$

and for Bayes estimator of $R(t)$, following quantities are used

$$\begin{aligned}
&z(\alpha,\lambda)=\left(1-e^{-\lambda t^{-\alpha}}\right)^{-c},\ z_1=ce^{-t^{-\alpha}\lambda}\left(1-e^{-t^{-\alpha}\lambda}\right)^{-1-c}t^{-\alpha}\lambda\ln t\\
&z_2=-ce^{-t^{-\alpha}\lambda}\left(1-e^{-t^{-\alpha}\lambda}\right)^{-1-c}t^{-\alpha}\\
&z_{11}=\frac{c\left(1-e^{-t^{-\alpha}\lambda}\right)^{-c}t^{-2\alpha}\lambda\left(-\left(e^{t^{-\alpha}\lambda}-1\right)t^{\alpha}+\left(c+e^{t^{-\alpha}\lambda}\right)\lambda\right)(\ln t)^2}{\left(e^{t^{-\alpha}\lambda}-1\right)^2}\\
&z_{22}=\frac{c\left(1-e^{-t^{-\alpha}\lambda}\right)^{-c}\left(c+e^{t^{-\alpha}\lambda}\right)t^{-2\alpha}}{\left(e^{t^{-\alpha}\lambda}-1\right)^2}\\
&z_{12}=\frac{c\left(1-e^{-t^{-\alpha}\lambda}\right)^{-c}t^{-2\alpha}\left(-\left(e^{t^{-\alpha}\lambda}-1\right)t^{\alpha}+\left(c+e^{t^{-\alpha}\lambda}\right)\lambda\right)\ln t}{\left(e^{t^{-\alpha}\lambda}-1\right)^2}=z_{21}.
\end{aligned}$$

## 4. Markov chain Monte Carlo

From equation (10), we see that posterior density is complex in nature and exact Bayes estimates of parameters are not easy to compute. To tackle this situation, one of the most popular tools known as Markov chain Monte Carlo (MCMC) is applied here. MCMC is is a powerful computational method used for generating samples from complex probability distributions and obtaining approximate Bayes estimates of the unknown parameters. This tools has significant popularity in various scientific fields, including statistics, machine learning, physics, and computational biology. To derive the approximate Bayes estimator of $\alpha$, $\lambda$ and $R(t)$, we use the MCMC technique in this part. With the use of posterior densities, the MCMC method is utilised to generate a random samples of unknown quantities. The Bayes estimator for the loss functions is then obtained using the generated samples. For this we first derived the conditional posterior densities of $\alpha$ and $\lambda$, from equation (10) as,

$$\left.\begin{array}{ll}\pi(\alpha|\lambda,\boldsymbol{y}) \propto & \alpha^{n+a_1-1}\left(\prod_{i=1}^{n} y_i^{-(\alpha+1)} e^{-\lambda y_i^{-\alpha}}\right)\prod_{i=1}^{n-1}\left(e^{-\lambda y_i^{-\alpha}}\right)^{m}\left(e^{-\lambda y_n^{-\alpha}}\right)^{k-1} e^{-b_1\alpha} \\ \pi(\lambda|\alpha,\boldsymbol{y}) \propto & \lambda^{n+a_2-1}\left(\prod_{i=1}^{n} e^{-\lambda y_i^{-\alpha}}\right)\prod_{i=1}^{n-1}\left(e^{-\lambda y_i^{-\alpha}}\right)^{m}\left(e^{-\lambda y_n^{-\alpha}}\right)^{k-1} e^{-b_2\lambda}\end{array}\right\}. \qquad (16)$$

From equation (16), we observe that the marginal posterior densities of $\alpha$ and $\lambda$ do not have known form of any probability distribution. So, we adopt the technique of Metropolis Hasting (MH) algorithm with normal distribution (see Gelman *et al.* (2013)) as the proposal density to generate samples. The algorithm and steps are followed from Arshad *et al.* (2021).

## 5. Simulation study

This section comprises of studying the behavior of the derived estimators on the simulated model. Various configurations of the parameters, sample sizes and priors have been tested and reported in this section. Since *dgos* is an umbrella term containing many models having different configurations for random variables of ordered nature, we have confined ourselves to study the lower record data and order statistics. To assess the credibility of Bayes estimators, risk function is taken to be the measure. The first thing is to generate the random samples from the *dgos* setup. For this purpose the algorithm discussed by Azhad *et al.* (2021) is considered here. Using the generated samples, for 1000 replications, all the estimators are obtained along with the risks in their estimation. For assessing the different possibilities, we have considered two set of priors i.e., Prior I : $(a_i, b_i) = (2, 2), i = 1, 2$ and Prior II : $(a_i, b_i) = (0.05, 0.05), i = 1, 2$, and different configurations of shape and scale parameters. The calculation is performed using R software (R Core Team (2022)). In addition to this, the convergence behaviour of generated Markov chain is tested with the aid of Gelman Rubin (GR) diagnostic (See Gelman *et al.* (2013)). With GR diagnostic we find that as we increase the number of iterations, the value of shrink reduction factor is getting close to 1. Hence, we conclude that convergence is achieved. The risks of various estimators are reported in Table [1-4] (see Appendix). From these tables, the following observations are made.

(i) The Table [1] (see Appendix) reports risks of Bayes estimates obtained using Lindley Approximation method for lower record values. From the table, it is observed that risks based on asymmetric loss functions (LINEX and GELF) are much smaller than symmetric loss function.

(ii) The Table [2] (see Appendix) reports risks of Bayes estimates obtained using MCMC method for lower record values. From the table, it is observed that risks based on asymmetric loss functions (LINEX and GELF) are much smaller than symmetric loss function. It is also observed that mostly risks of estimators based on MCMC method smaller than risk of estimators based on Lindley method.

(iv) The Table [3 - 4] (see Appendix) report risks of Bayes estimates obtained using Lindley and MCMC method for order statistics, respectively. Similar observations are seen for risks of all estimators for order statistics as these were for lower record values.

(v) From all the Tables, it is observed that the risks of all estimators are decreasing as we increase the sample size irrespective of ordered random models. Also, on average, Prior I seems to have showm lesser risk that Prior II.

(vi) From these observations it is evident that Bayes estimators based on asymmetric loss functions (LINEX and GELF) are performing better based on their risks. So, In practical scenarios where the underlying assumptions considered in this study are satisfied, it is recommended to use asymmetric loss functions as it provides more flexibility to the model. Also, estimators based on MCMC method are performing better than Lindley estimators.

## 6. Discussion and conclusions

In the present manuscript Fréchet distribution is considered and Bayesian perspective on estimation is explored under the *dgos* configuration. The considered distribution has many applications like it is used in hydrology to describe severe occurrences like annual maximum one-day rainfall and river discharges, used to depict a falling pattern in time series data of oil or gas production rate over time for a well, employed to simulate the idiosyncratic element of people's preferences for various goods, places , or businesses *etc.* The reliability function and Bayes estimators of unknown quantities are thoroughly addressed. For Bayesian methods, it makes sense to take distinct loss functions into account. In addition, a discussion of the findings for order statistics under *dgos's* setup and lower record values under *dgos's* setup is given. After careful examination of the simulation results, we come to the conclusion that, MCMC is a better choice than Lindley approximation for estimation of parameters $\alpha$, $\lambda$, and $R(t)$ in both the cases of lower record values and order statistics for the considered distribution.

For future studies scaled squared error loss function, precautionary loss function, K-loss function, regression loss function, *etc.*, may be used. This research may possibly be expanded by assuming additional estimating techniques and applying them on censored data.

## Acknowledgment

We sincerely appreciate the Editors' advice and support. I am also appreciative that the conference organisers gave us the chance to present our work.

## References

Abbas, K. and Tang, Y. (2015). Analysis of Frechet distribution using reference priors. *Communications in Statistics-Theory and Methods*, **44**, 2945–2956.

Ahsanullah, M. (2004). *Record Values–Theory and Applications.* University Press of America.

Ali, S. (2015). On the Bayesian estimation of the weighted Lindley distribution. *Journal of Statistical Computation and Simulation*, **85**, 855–880.

Anwar, Z., Gupta, N., Khan, M. A. R., and Jamal, Q. A. (2020). Recurrence relations for marginal and joint moment generating functions of topp-leone generated exponential distribution based on record values and its characterization. *Journal of Modern Applied Statistical Methods*, **18**, 25.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A First Course in Order Statistics*. SIAM.

Arshad, M. and Baklizi, A. (2019). Estimation of common location parameter of two exponential populations based on records. *Communications in Statistics-Theory and Methods*, **48**, 1545–1552.

Arshad, M., J. Azhad, Q., Gupta, N., and Pathak, A. K. (2021). Bayesian inference of Unit Gompertz distribution based on dual generalized order statistics. *Communications in Statistics-Simulation and Computation*, **00**, 1–19.

Arshad, M. and Jamal, Q. A. (2019a). Estimation of common scale parameter of several heterogeneous Pareto populations based on records. *Iranian Journal of Science and Technology, Transactions A: Science*, **43**, 2315–2323.

Arshad, M. and Jamal, Q. A. (2019b). Statistical inference for Topp–Leone generated family of distributions based on records. *Journal of Statistical Theory and Applications*, **18**, 65–78.

Azhad, Q. J., Arshad, M., Devi, B., Khandelwal, N., and Ali, I. (2023). Record-based transmuted kumaraswamy generalized family of distributions: Properties and application *G Families of Probability Distributions*. CRC Press.

Azhad, Q. J., Arshad, M., and Khandelwal, N. (2022). Statistical inference of reliability in multicomponent stress strength model for Pareto distribution based on upper record values. *International Journal of Modelling and Simulation*, **42**, 319–334.

Azhad, Q. J., Arshad, M., and Misra, A. K. (2021). Estimation of common location parameter of several heterogeneous exponential populations based on generalized order statistics. *Journal of Applied Statistics*, **48**, 1798–1815.

Calabria, R. and Pulcini, G. (1989). Confidence limits for reliability and tolerance limits in the inverse Weibull distribution. *Reliability Engineering & System Safety*, **24**, 77–85.

Calabria, R. and Pulcini, G. (1994). Bayes 2-sample prediction for the inverse Weibull distribution. *Communications in Statistics-Theory and Methods*, **23**, 1811–1824.

Devi, B., Kumar, P., and Kour, K. (2017). Entropy of Lomax probability distribution and its order statistic. *International Journal of Statistics and System*, **12**, 175–181.

Dey, S. (2009). Comparison of Bayes estimators of the parameter and reliability function for Rayleigh distribution under different loss functions. *Malaysian Journal of Mathematical Sciences*, **3**, 247–264.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gupta, N. and Jamal, Q. A. (2019). Inference for Weibull generalized exponential distribution based on generalized order statistics. *Journal of Applied Mathematics and Computing*, **61**, 573–592.

Harlow, D. G. (2002). Applications of the Fr'echet distribution function. *International Journal of Materials and Product Technology*, **17**, 482–495.

Jaheen, Z. F. (2003). A Bayesian analysis of record statistics from the Gompertz model. *Applied Mathematics and Computation*, **145**, 307–320.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications.* world scientific.

Kundu, D. and Howlader, H. (2010). Bayesian inference and prediction of the inverse Weibull distribution for Type-II censored data. *Computational Statistics & Data Analysis*, **54**, 1547–1558.

Lindley, D. V. (1980). Approximate Bayesian methods. *Trabajos de estadística y de investigación operativa*, **31**, 223–245.

Loganathan, A. and Uma, A. (2017). Comparison of estimation methods for inverse Weibull parameters. *Global and Stochastic Analysis*, **4**, 83–93.

Maswadah, M. (2003). Conditional confidence interval estimation for the inverse Weibull distribution based on censored generalized order statistics. *Journal of Statistical Computation and Simulation*, **73**, 887–898.

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramos, P. L., Nascimento, D., and Louzada, F. (2017). The Long Term Fr\'echet distribution: Estimation, Properties and its Application. *Biometrics & Biostatistics International Journal*, **6**, 1–6.

Salman, M. and AMER, S. S. M. (2003). Order statistics from inverse Weibull distribution and characterizations. *Metron*, **61**, 389–401.

Sharma, A., Kumar, P., and Devi, B. (2019). Entropy estimation of inverse Rayleigh probability distribution and its order statistics. *International Journal of Electronics Engineering*, **11**, 508–513.

Tripathi, A., Singh, U., and Singh, S. K. (2019). Inferences for the DUS-exponential distribution based on upper record values. *Annals of Data Science*, **8**, 387–403.

Zhang, F. and Gui, W. (2020). Parameter and reliability inferences of inverted exponentiated half-logistic distribution under the progressive first-failure censoring. *Mathematics*, **8**, 1–29.

## APPENDIX

**Table 1: Risk of Lindley Bayes estimates based on lower record values for $(c, t) = (0.5, 0.5)$**

| $(\alpha, \lambda)$ | $n$ | SELF | | | Linex | | | General Entropy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\widehat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\widehat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\widehat{R(t)}_{risk}$ |
| | | $(a_1, a_2, b_1, b_2) = (2, 2, 2, 2)$ | | | | | | | | |
| (1,1) | 5 | 0.1493 | 0.2405 | 0.2423 | 0.0200 | 0.0340 | 0.0304 | 0.0124 | 0.0362 | 0.0467 |
| | 10 | 0.0896 | 0.1096 | 0.2209 | 0.0079 | 0.0125 | 0.0263 | 0.0033 | 0.0200 | 0.0361 |
| | 15 | 0.0306 | 0.0690 | 0.2208 | 0.0017 | 0.0047 | 0.0260 | 0.0011 | 0.0103 | 0.0330 |
| (1.5,1) | 5 | 0.3810 | 0.3609 | 0.2753 | 0.0611 | 0.0358 | 0.0330 | 0.0408 | 0.0476 | 0.0562 |
| | 10 | 0.2700 | 0.1910 | 0.2628 | 0.0181 | 0.0331 | 0.0330 | 0.0107 | 0.0275 | 0.0445 |
| | 15 | 0.1532 | 0.1246 | 0.2595 | 0.0059 | 0.0146 | 0.0302 | 0.0041 | 0.0194 | 0.0393 |
| (1,1.5) | 5 | 0.1600 | 0.3802 | 0.3791 | 0.0211 | 0.0348 | 0.0452 | 0.0109 | 0.0633 | 0.0628 |
| | 10 | 0.0912 | 0.3383 | 0.3664 | 0.0083 | 0.0252 | 0.0423 | 0.0064 | 0.0409 | 0.0537 |
| | 15 | 0.0398 | 0.3117 | 0.3620 | 0.0040 | 0.0237 | 0.0416 | 0.0041 | 0.0315 | 0.0498 |
| (1.5,1.5) | 5 | 0.2758 | 0.3346 | 0.3793 | 0.0349 | 0.0271 | 0.0451 | 0.0231 | 0.0640 | 0.0640 |
| | 10 | 0.1552 | 0.2322 | 0.3720 | 0.0083 | 0.0166 | 0.0432 | 0.0056 | 0.0316 | 0.0559 |
| | 15 | 0.0812 | 0.2025 | 0.3599 | 0.0041 | 0.0119 | 0.0411 | 0.0035 | 0.0213 | 0.0504 |
| | | $(a_1, a_2, b_1, b_2) = (0.05, 0.05, 0.05, 0.05)$ | | | | | | | | |
| (1,1) | 5 | 0.1853 | 0.3233 | 0.2036 | 0.0239 | 0.0337 | 0.0254 | 0.0221 | 0.0501 | 0.0453 |
| | 10 | 0.1394 | 0.2609 | 0.2030 | 0.0167 | 0.0284 | 0.0248 | 0.0154 | 0.0415 | 0.0381 |
| | 15 | 0.0849 | 0.2423 | 0.1996 | 0.0107 | 0.0272 | 0.0240 | 0.0100 | 0.0347 | 0.0342 |
| (1.5,1) | 5 | 0.1254 | 0.3322 | 0.2742 | 0.0155 | 0.0347 | 0.0335 | 0.0188 | 0.0492 | 0.0560 |
| | 10 | 0.0864 | 0.2873 | 0.2716 | 0.0107 | 0.0305 | 0.0310 | 0.0119 | 0.0410 | 0.0455 |
| | 15 | 0.0648 | 0.2483 | 0.2568 | 0.0082 | 0.0279 | 0.0322 | 0.0086 | 0.0366 | 0.0440 |
| (1,1.5) | 5 | 0.1919 | 0.5153 | 0.2899 | 0.0233 | 0.0537 | 0.0353 | 0.0211 | 0.0878 | 0.0575 |
| | 10 | 0.1363 | 0.4065 | 0.2890 | 0.0175 | 0.0434 | 0.0345 | 0.0162 | 0.0700 | 0.0496 |
| | 15 | 0.0987 | 0.3688 | 0.2732 | 0.0128 | 0.0400 | 0.0323 | 0.0120 | 0.0620 | 0.0437 |
| (1.5,1.5) | 5 | 0.1092 | 0.4841 | 0.3193 | 0.0138 | 0.0507 | 0.0387 | 0.0169 | 0.0851 | 0.0620 |
| | 10 | 0.0834 | 0.4338 | 0.3143 | 0.0106 | 0.0458 | 0.0374 | 0.0115 | 0.0734 | 0.0531 |
| | 15 | 0.0736 | 0.3466 | 0.3122 | 0.0096 | 0.0382 | 0.0367 | 0.0100 | 0.0601 | 0.0488 |

**Table 2: Risk of MCMC Bayes estimates based on lower record values for $(c, t) = (0.5, 0.5)$**

| $(\alpha, \lambda)$ | $n$ | SELF | | | Linex | | | General Entropy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ |
| $(a_1, a_2, b_1, b_2) = (2, 2, 2, 2)$ | | | | | | | | | | |
| (1,1) | 5 | 0.0626 | 0.0314 | 0.1837 | 0.0079 | 0.0040 | 0.0214 | 0.0057 | 0.0033 | 0.0380 |
| | 10 | 0.0601 | 0.0302 | 0.1832 | 0.0076 | 0.0038 | 0.0214 | 0.0056 | 0.0032 | 0.0379 |
| | 15 | 0.0578 | 0.0287 | 0.1816 | 0.0073 | 0.0036 | 0.0212 | 0.0054 | 0.0030 | 0.0375 |
| (1.5,1) | 5 | 0.0646 | 0.2124 | 0.1874 | 0.0082 | 0.0247 | 0.0218 | 0.0059 | 0.0179 | 0.0390 |
| | 10 | 0.0589 | 0.2060 | 0.1833 | 0.0075 | 0.0239 | 0.0214 | 0.0055 | 0.0159 | 0.0379 |
| | 15 | 0.0557 | 0.1928 | 0.1801 | 0.0070 | 0.0225 | 0.0210 | 0.0052 | 0.0174 | 0.0371 |
| (1,1.5) | 5 | 0.0875 | 0.0329 | 0.1848 | 0.0106 | 0.0041 | 0.0215 | 0.0064 | 0.0035 | 0.0380 |
| | 10 | 0.0839 | 0.0319 | 0.1839 | 0.0102 | 0.0040 | 0.0214 | 0.0061 | 0.0033 | 0.0383 |
| | 15 | 0.0837 | 0.0289 | 0.1836 | 0.0101 | 0.0036 | 0.0214 | 0.0061 | 0.0030 | 0.0380 |
| (1.5,1.5) | 5 | 0.0873 | 0.2180 | 0.1844 | 0.0106 | 0.0253 | 0.0215 | 0.0064 | 0.0184 | 0.0382 |
| | 10 | 0.0840 | 0.2139 | 0.1808 | 0.0101 | 0.0248 | 0.0211 | 0.0061 | 0.0181 | 0.0373 |
| | 15 | 0.0835 | 0.2022 | 0.1816 | 0.0102 | 0.0235 | 0.0212 | 0.0061 | 0.0168 | 0.0375 |
| $(a_1, a_2, b_1, b_2) = (0.05, 0.05, 0.05, 0.05)$ | | | | | | | | | | |
| (1,1) | 5 | 0.0627 | 0.0319 | 0.1851 | 0.0079 | 0.0040 | 0.0216 | 0.0058 | 0.0035 | 0.0384 |
| | 10 | 0.0591 | 0.0308 | 0.1834 | 0.0075 | 0.0039 | 0.0211 | 0.0055 | 0.0032 | 0.0372 |
| | 15 | 0.0560 | 0.0306 | 0.1804 | 0.0071 | 0.0039 | 0.0214 | 0.0052 | 0.0032 | 0.0379 |
| (1.5,1) | 5 | 0.0649 | 0.2036 | 0.1839 | 0.0082 | 0.0240 | 0.0215 | 0.0060 | 0.0173 | 0.0381 |
| | 10 | 0.0605 | 0.2031 | 0.1821 | 0.0077 | 0.0237 | 0.0213 | 0.0056 | 0.0169 | 0.0376 |
| | 15 | 0.0585 | 0.2068 | 0.1804 | 0.0074 | 0.0237 | 0.0211 | 0.0054 | 0.0168 | 0.0371 |
| (1,1.5) | 5 | 0.0863 | 0.0328 | 0.1835 | 0.0105 | 0.0041 | 0.0214 | 0.0063 | 0.0034 | 0.0380 |
| | 10 | 0.0845 | 0.0299 | 0.1833 | 0.0102 | 0.0038 | 0.0214 | 0.0062 | 0.0032 | 0.0379 |
| | 15 | 0.0832 | 0.0296 | 0.1833 | 0.0101 | 0.0037 | 0.0214 | 0.0061 | 0.0031 | 0.0379 |
| (1.5,1.5) | 5 | 0.0880 | 0.2113 | 0.1848 | 0.0106 | 0.0245 | 0.0215 | 0.0064 | 0.0176 | 0.0383 |
| | 10 | 0.0841 | 0.2087 | 0.1820 | 0.0102 | 0.0243 | 0.0212 | 0.0061 | 0.0174 | 0.0376 |
| | 15 | 0.0833 | 0.2053 | 0.1819 | 0.0101 | 0.0239 | 0.0212 | 0.0061 | 0.0172 | 0.0375 |

**Table 3: Risk of Lindley Bayes estimates based on order statistics for $(c, t) = (0.5, 0.5)$**

| $(\alpha, \lambda)$ | $n$ | SELF | | | Linex | | | General Entropy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ |
| $(a_1, a_2, b_1, b_2) = (2, 2, 2, 2)$ | | | | | | | | | | |
| (1,1) | 5 | 0.0711 | 0.0632 | 0.6544 | 0.0086 | 0.0049 | 0.0718 | 0.0102 | 0.0074 | 0.0744 |
| | 10 | 0.0643 | 0.0599 | 0.6168 | 0.0078 | 0.0079 | 0.0680 | 0.0078 | 0.0087 | 0.0719 |
| | 15 | 0.0481 | 0.0369 | 0.4837 | 0.0059 | 0.0073 | 0.0540 | 0.0057 | 0.0074 | 0.0631 |
| (1.5,1) | 5 | 0.1965 | 0.0571 | 0.7713 | 0.0154 | 0.0045 | 0.0837 | 0.0178 | 0.0060 | 0.0864 |
| | 10 | 0.0608 | 0.0562 | 0.7173 | 0.0074 | 0.0069 | 0.0782 | 0.0083 | 0.0077 | 0.0818 |
| | 15 | 0.0503 | 0.0380 | 0.5412 | 0.0063 | 0.0069 | 0.0597 | 0.0067 | 0.0071 | 0.0684 |
| (1,1.5) | 5 | 0.0882 | 0.3452 | 0.8049 | 0.0112 | 0.0412 | 0.0871 | 0.0153 | 0.0420 | 0.0898 |
| | 10 | 0.0640 | 0.0974 | 0.7617 | 0.0079 | 0.0106 | 0.0828 | 0.0074 | 0.0131 | 0.0869 |
| | 15 | 0.0471 | 0.0711 | 0.6257 | 0.0058 | 0.0092 | 0.0690 | 0.0056 | 0.0103 | 0.0799 |
| (1.5,1.5) | 5 | 0.2417 | 0.2860 | 0.8531 | 0.0249 | 0.0417 | 0.0920 | 0.0286 | 0.0403 | 0.0947 |
| | 10 | 0.0799 | 0.0996 | 0.7967 | 0.0096 | 0.0095 | 0.0863 | 0.0094 | 0.0119 | 0.0903 |
| | 15 | 0.0500 | 0.0698 | 0.6197 | 0.0064 | 0.0089 | 0.0682 | 0.0067 | 0.0102 | 0.0780 |
| $(a_1, a_2, b_1, b_2) = (0.05, 0.05, 0.05, 0.05)$ | | | | | | | | | | |
| (1,1) | 5 | 0.1562 | 0.1643 | 0.6613 | 0.0203 | 0.0209 | 0.0725 | 0.0194 | 0.0239 | 0.0750 |
| | 10 | 0.1004 | 0.1094 | 0.6263 | 0.0130 | 0.0131 | 0.0690 | 0.0121 | 0.0136 | 0.0731 |
| | 15 | 0.0653 | 0.0870 | 0.5379 | 0.0082 | 0.0108 | 0.0599 | 0.0077 | 0.0113 | 0.0694 |
| (1.5,1) | 5 | 0.1124 | 0.1681 | 0.7867 | 0.0135 | 0.0196 | 0.0853 | 0.0154 | 0.0224 | 0.0880 |
| | 10 | 0.0713 | 0.1163 | 0.7492 | 0.0091 | 0.0142 | 0.0815 | 0.0097 | 0.0145 | 0.0860 |
| | 15 | 0.0583 | 0.0836 | 0.6535 | 0.0071 | 0.0102 | 0.0720 | 0.0073 | 0.0100 | 0.0821 |
| (1,1.5) | 5 | 0.1566 | 0.2145 | 0.8007 | 0.0193 | 0.0250 | 0.0867 | 0.0177 | 0.0307 | 0.0895 |
| | 10 | 0.0920 | 0.1135 | 0.7594 | 0.0111 | 0.0144 | 0.0826 | 0.0101 | 0.0164 | 0.0872 |
| | 15 | 0.0589 | 0.0814 | 0.6712 | 0.0076 | 0.0103 | 0.0738 | 0.0073 | 0.0111 | 0.0840 |
| (1.5,1.5) | 5 | 0.0961 | 0.1915 | 0.8692 | 0.0122 | 0.0225 | 0.0936 | 0.0147 | 0.0289 | 0.0964 |
| | 10 | 0.0702 | 0.1213 | 0.8295 | 0.0091 | 0.0146 | 0.0897 | 0.0096 | 0.0164 | 0.0944 |
| | 15 | 0.0612 | 0.0841 | 0.7239 | 0.0076 | 0.0104 | 0.0792 | 0.0077 | 0.0112 | 0.0899 |

**Table 4: Risk of MCMC Bayes estimates based on order statistics for $(c, t) = (0.5, 0.5)$**

| $(\alpha, \lambda)$ | $n$ | SELF | | | Linex | | | General Entropy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ | $\hat{\alpha}_{risk}$ | $\hat{\lambda}_{risk}$ | $\hat{R(t)}_{risk}$ |
| $(a_1, a_2, b_1, b_2) = (2, 2, 2, 2)$ | | | | | | | | | | |
| (1,1) | 5 | 0.0428 | 0.2054 | 0.2548 | 0.0052 | 0.0249 | 0.0294 | 0.0077 | 0.0670 | 0.0603 |
| | 10 | 0.0401 | 0.1966 | 0.2512 | 0.0049 | 0.0238 | 0.0290 | 0.0072 | 0.0632 | 0.0591 |
| | 15 | 0.0379 | 0.1961 | 0.2473 | 0.0047 | 0.0237 | 0.0285 | 0.0068 | 0.0629 | 0.0579 |
| (1.5,1) | 5 | 0.0457 | 0.8683 | 0.2593 | 0.0056 | 0.0952 | 0.0299 | 0.0083 | 0.1353 | 0.0618 |
| | 10 | 0.0442 | 0.8594 | 0.2554 | 0.0054 | 0.0942 | 0.0294 | 0.0081 | 0.1318 | 0.0608 |
| | 15 | 0.0422 | 0.8490 | 0.2549 | 0.0052 | 0.0932 | 0.0294 | 0.0077 | 0.1288 | 0.0605 |
| (1,1.5) | 5 | 0.4668 | 0.2074 | 0.2550 | 0.0525 | 0.0251 | 0.0294 | 0.0452 | 0.0671 | 0.0602 |
| | 10 | 0.4553 | 0.2028 | 0.2500 | 0.0513 | 0.0245 | 0.0289 | 0.0439 | 0.0656 | 0.0588 |
| | 15 | 0.4471 | 0.1914 | 0.2465 | 0.0504 | 0.0232 | 0.0285 | 0.0427 | 0.0602 | 0.0575 |
| (1.5,1.5) | 5 | 0.4587 | 0.8904 | 0.2509 | 0.0516 | 0.0975 | 0.0290 | 0.0441 | 0.1410 | 0.0573 |
| | 10 | 0.4562 | 0.8883 | 0.2500 | 0.0514 | 0.0973 | 0.0288 | 0.0439 | 0.1392 | 0.0587 |
| | 15 | 0.4417 | 0.8818 | 0.2454 | 0.0498 | 0.0966 | 0.0283 | 0.0422 | 0.1352 | 0.0589 |
| $(a_1, a_2, b_1, b_2) = (0.05, 0.05, 0.05, 0.05)$ | | | | | | | | | | |
| (1,1) | 5 | 0.0402 | 0.2067 | 0.2564 | 0.0054 | 0.0250 | 0.0289 | 0.0080 | 0.0673 | 0.0591 |
| | 10 | 0.0403 | 0.1973 | 0.2509 | 0.0049 | 0.0239 | 0.0288 | 0.0073 | 0.0635 | 0.0585 |
| | 15 | 0.0438 | 0.1884 | 0.2493 | 0.0049 | 0.0229 | 0.0295 | 0.0072 | 0.0596 | 0.0609 |
| (1.5,1) | 5 | 0.0432 | 0.8842 | 0.2544 | 0.0053 | 0.0968 | 0.0293 | 0.0078 | 0.1371 | 0.0603 |
| | 10 | 0.0405 | 0.8697 | 0.2527 | 0.0050 | 0.0954 | 0.0291 | 0.0073 | 0.1357 | 0.0596 |
| | 15 | 0.0378 | 0.8654 | 0.2477 | 0.0046 | 0.0949 | 0.0286 | 0.0067 | 0.1345 | 0.0579 |
| (1,1.5) | 5 | 0.4582 | 0.2015 | 0.2515 | 0.0516 | 0.0244 | 0.0290 | 0.0443 | 0.0651 | 0.0591 |
| | 10 | 0.4582 | 0.2014 | 0.2510 | 0.0516 | 0.0244 | 0.0290 | 0.0442 | 0.0641 | 0.0590 |
| | 15 | 0.4562 | 0.1980 | 0.2505 | 0.0514 | 0.0240 | 0.0289 | 0.0442 | 0.0634 | 0.0592 |
| (1.5,1.5) | 5 | 0.4581 | 0.8825 | 0.2517 | 0.0516 | 0.0967 | 0.0290 | 0.0442 | 0.1380 | 0.0593 |
| | 10 | 0.4557 | 0.8772 | 0.2507 | 0.0511 | 0.0961 | 0.0288 | 0.0438 | 0.1368 | 0.0589 |
| | 15 | 0.4535 | 0.8746 | 0.2496 | 0.0513 | 0.0958 | 0.0289 | 0.0436 | 0.1353 | 0.0585 |

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 107–118

# Long Memory in Volatility: Application of Fractionally Integrated GARCH Model

**Debopam Rakshit and Ranjit Kumar Paul**
*ICAR- Indian Agricultural Statistics Research Institute, New Delhi*

Received: 10 July 2023; Revised: 24 July 2023; Accepted: 28 July 2023

## Abstract

Volatility is an important characteristic of time series. If the volatility of a series at any time epoch is affected by its distant counterpart, then it is known as long memory in volatility. The (FIGARCH) model is useful for addressing the long memory in volatility. In this paper, for empirical illustration, the daily modal spot price of mustard from four markets of Rajasthan namely Khedli (Laxmangarh), Atru, Nimbahera and Anoopgarh, are used. The GARCH, EGARCH, APARCH, GJR-GARCH, and FIGARCH models are fitted to the log return series of the selected datasets. It is seen that the FIGARCH model is the best-fitted model for all the time series and it confirmed the presence of long memory in volatility.

*Key words:* GARCH; Long memory; Nonlinear models; Time series; Volatility.

## 1. Introduction

Time series analysis is used to identify patterns and trends in the dataset, and it helps make predictions about future values. Time series modelling is a crucial aspect for understanding the price behaviour and movement of any economic goods including the prices of agricultural commodities. The major breakthrough in time series modelling was first pioneered by Box and Jenkins (1970) through the introduction of the autoregressive integrated moving average (ARIMA) model. The ARIMA model is based on the assumptions of linearity and stationarity of the dataset and the homoscedasticity of the error variance. Lots of applications of the ARIMA model can be found in the literature (Paul *et al.*, 2014, 2020; Agarwal *et al.*, 2021). Linear models take advantage of their analytical and implementable easiness over the others. But it is irrational to assume a priori about the linear process for time series. Volatility is the nonlinear aspect of time series. It is the degree of unexpected variation of its realizations over a certain period. Engle (1982) introduced the autoregressive conditional heteroscedastic (ARCH) model for capturing the volatility of any time series. Later, its generalization, i.e. generalized ARCH (GARCH) model was proposed by Bollerslev (1986) and Taylor (1986) independent of each other. Applications of the GARCH model can be found in Paul *et al.* (2009, 2015), etc. The GARCH model is symmetric. It does not

Corresponding Author: Ranjit Kumar Paul
Email: ranjit.paul@icar.gov.in

account for the sign of shocks and only takes into consideration the amount of shocks' effects on volatility. Hence, it cannot capture the asymmetric behaviour of price volatility, i.e., reactions to the volatility may differ depending on whether the positive and negative shocks are of the same magnitude. The exponential GARCH (EGARCH) model (Nelson, 1991), Asymmetric Power ARCH (APARCH) model (Ding *et al.*, 1993), and GJR-GARCH model (Glosten *et al.*, 1993) are better alternatives to the GARCH model for addressing asymmetric volatility. Again, the realizations of a time series may have long term dependency. In the presence of long term dependency, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are significant for a long lag. This is known as hyperbolic decay. The long memory process can be present in both linear and nonlinear dynamics of a time series. If long memory is present in the linear model then the autoregressive fractionally integrated moving average (ARFIMA) model (Granger and Joyeux, 1980) is useful. Fractional integration is a generalization of ordinary integration, where the integral is taken to a fractional power. Some applications of the ARFIMA model can be found in Paul (2014) and Rakshit *et al.* (2022). Similarly, the fractionally integrated GARCH (FIGARCH) model (Baillie *et al.*, 1996) is useful for capturing the long memory in volatility. Paul *et al.* (2016) applied the FIGARCH model for modelling long memory in the volatility of the spot price of gram in Delhi, India. In the presence of long memory both in the mean and variance structure, Mitra *et al.* (2018) applied the ARFIMA-FIGARCH models for modelling the potato price of the Agra and Amritsar markets, India.

Agriculture is the backbone of the Indian economy. Around 60% of the Indian population depends on agriculture for their livelihood. As per the Second Advance Estimates of National Income, 2022-23 released by the Ministry of Statistics and Programme Implementation (MoSPI), the share of Gross value added (GVA) of agriculture and allied sectors in the total economy is 18.3% at current prices. The volatility study of the price series of agricultural commodities is an important aspect to social science researchers (Paul and Garai, 2021; Rakshit *et al.*, 2021, 2023; Garai *et al.*, 2023). Mustard is an important oilseed crop in India. It is grown in the rabi (winter) season and is a major source of edible oil for the country. The oilcake from mustard seeds is used as a feed for livestock. In addition to its edible oil, mustard has a number of other uses. The leaves of the mustard plant can be eaten as a vegetable, and the flowers can be used to make mustard seed paste, which is used as a condiment. Mustard seeds also have medicinal properties and have been used traditionally to treat a variety of ailments, including arthritis, rheumatism, and respiratory problems. Mustard cultivation provides livelihood opportunities for a large number of farmers, especially in the states of Rajasthan, Uttar Pradesh, Madhya Pradesh, and Punjab, where it is extensively grown. Earlier, it is seen that the modelling and forecasting of rapeseed and mustard prices helps in improving decision making in Rajasthan (Bhardwaj *et al.*, 2015). In the present study, the modal daily spot price series of mustard for Khedli (Laxmangarh), Atru, Nimbahera and Anoopgarh markets of Rajasthan are used. The GARCH, EGARCH, APARCH, GJR-GARCH and FIGARCH models are applied to the selected time series. Section 2 includes a description of the used models. The empirical illustration is given in Section 3 followed by concluding remarks in Section 4.

## 2. Materials and methods

### 2.1. The ARCH and GARCH models

ARIMA is a linear model that cannot address the nonlinear dynamics of a time series. Homoscedasticity in the error variance is a basic assumption of this model. By relaxing the linear and homoscedasticity assumptions, the ARCH model is introduced by taking into account substantial autocorrelations present in the squared residual series to capture the nonlinear dynamics of a time series. A process $\{\varepsilon_t\}$ is said to follow the ARCH $(q)$ model if the conditional distribution of $\{\varepsilon_t\}$ given the available information $(\psi_{t-1})$ up to $t-1$ time epoch is represented as:

$$\varepsilon_t|\psi_{t-1} \sim \; N\left(0, h_t\right) \; and \; \varepsilon_t = \sqrt{h_t}\nu_t \tag{1}$$

where $\nu_t$ is identically and independently distributed (IID) innovation with zero mean and unit variance. The conditional variance $h_t$ of ARCH $(q)$ model is calculated as

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2, \; \alpha_0 > 0, \; \alpha_i \geq 0 \; \forall \; i \; and \; \sum_{i=1}^{q} \alpha_i < 1 \tag{2}$$

The GARCH model is a more parsimonious version of the ARCH model where the number of parameters to be estimated is less. Here, the conditional variance is treated as a linear function of its own lags. The GARCH $(p, q)$ model has the following form of conditional variance

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \tag{3}$$

provided $\alpha_0 > 0, \alpha_i \geq 0 \; \forall \; i \, \beta_j \geq 0 \; \forall \; j$

$\alpha_i$ and $\beta_j$ parameters indicate how previous shocks and volatility have influenced current volatility, respectively. The GARCH $(p, q)$ model is said to be weakly stationary if and only if

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1 \tag{4}$$

The GARCH model only considers the dependencies of volatility on the magnitude of the shocks, and it does not consider the sign of the shocks that influence the degree of volatility. The EGARCH, APARCH, and GJR-GARCH models can be useful to overcome this gap.

### 2.2. EGARCH model

The EGARCH model is introduced by defining the conditional variance in terms of the logarithm function. The main advantage of this model over the GARCH model, aside from addressing the asymmetric volatility, is that no restriction is imposed on the parameters as the positivity of the conditional variance is always achieved. The conditional variance for the EGARCH model is defined as

$$\ln h_t = \; \alpha_0 + \sum_{j=1}^{p} \beta_j \ln h_{t-j} + \sum_{i=1}^{q} \left( \alpha_i \left| \frac{\varepsilon_{t-i}}{\sqrt{h_{t-i}}} \right| + \; \gamma_i \frac{\varepsilon_{t-i}}{\sqrt{h_{t-i}}} \right) \tag{5}$$

where, $\gamma_i$ is the asymmetric factor which explains the asymmetric effect due to external shocks. For EGARCH (1,1) model conditional variance $h_t$ is reduced to

$$\ln h_t = \alpha_0 + \beta_1 \ln h_{t-1} + \left( \alpha_1 \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| + \gamma \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right) \tag{6}$$

## 2.3. APARCH model

The APARCH model considers some asymmetric power to the conditional variance $h_t$. The conditional variance of the APARCH model is defined as

$$h_t^{\frac{\delta}{2}} = \alpha_0 + \sum_{j=1}^{p} \beta_j h_{t-j}^{\frac{\delta}{2}} + \sum_{i=1}^{q} \alpha_i \left( |\varepsilon_{t-i}| - \gamma \varepsilon_{t-i} \right)^{\delta} \tag{7}$$

where $\gamma(-1 < \gamma < 1)$ is the parameter for asymmetry and $\delta(> 0)$ is the power term parameter. The APARCH model is a general framework of models. Different orders of GARCH models can be fitted within the APARCH model by defining specific values to the parameters. For $\delta = 2$ and $\gamma = 0$, the APARCH model is the same as the GARCH model. The conditional variance $h_t$ for APARCH (1,1) model is reduced to

$$h_t^{\frac{\delta}{2}} = \alpha_0 + \beta_1 h_{t-1}^{\frac{\delta}{2}} + \alpha_1 \left( |\varepsilon_{t-1}| - \gamma \varepsilon_{t-1} \right)^{\delta} \tag{8}$$

## 2.4. GJR-GARCH model

The GJR-GARCH model considers the impact of $\varepsilon_{t-1}^2$ on the conditional variance based on the sign of $\varepsilon_{t-1}$. An indicator variable is introduced to capture the sign dependency. The conditional variance of the GJR-GARCH model is defined as

$$h_t = \alpha_0 + \sum_{j=1}^{p} \beta_j h_{t-j} + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \gamma \varepsilon_{t-1}^2 I_{t-1} \tag{9}$$

where $\gamma(-1 < \gamma < 1)$ is the asymmetric parameter and $I_{t-1}$ is the indicator variable, such that

$$\begin{aligned} I_{t-1} = 1 \quad & if\ \varepsilon_{t-1} < 0 \\ 0 \quad & if\ \varepsilon_{t-1} \geq 0 \end{aligned}$$

For GJR-GARCH (1,1) model conditional variance $h_t$ is reduced to

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} + \gamma \varepsilon_{t-1}^2 I_{t-1} \tag{10}$$

## 2.5. FIGARCH model

The FIGARCH model is useful when the volatility is symmetric i.e. positive and negative shocks of the same magnitude exhibit the same response to volatility and the volatility exhibits long term persistence. The FIGARCH model is derived by introducing a fractional differencing parameter in the GARCH model after some algebraic operations.

Tayefi and Ramanathan (2012) provided a thorough review of the FIGARCH model. The FIGARCH $(p, d, q)$ model can be expressed as

$$[1-\alpha(L)-\beta(L)](1-L)^d \varepsilon_t^2=\alpha_0+[1-\beta(L)]z_t \tag{11}$$

where, $\alpha(L)$ and $\beta(L)$ are polynomials in lag operator and $(1-L)^d$ is the fractional difference operator. Here, $d$ is a fraction and $0 < d < 1$.

## 3. Empirical illustration

### 3.1. Data description

For empirical illustration purposes, the daily modal spot prices (Rs./q) of mustard for four markets in Rajasthan namely Khedli (Laxmangarh), Atru, Nimbahera and Anoopgarh are collected from the Ministry of Agriculture and Farmers' Welfare, Government of India for the study period of 1st January 2010 to 31st May 2023 (total number of observation is 4899). Since the square of return is regarded as the realization of volatility, the analysis is done with the log return series of the selected time series data. For a time series $\{y_t\}$ the log return series $\{r_t\}$ is calculated as

$$r_t=\ln\frac{y_t}{y_{t-1}} \tag{12}$$

The latest 250 realizations of the log return series of each of the selected markets are used as the model validation set, while the remaining previous portion is used as the model building set.

### 3.2. Descriptive statistics

The descriptive statistics of the selected price series are given in Table 1. The Khedli market has the highest mean price, while the Nimbahera market has the lowest mean price. The Atru market has the highest median price and the Nimbahera market is the lowest one. Regarding the minimum price, the Khedli market minimum price is significantly lower than the others. The Khedli market has the highest maximum price and the Nimbahera has the lowest maximum price. All the selected price series are positively skewed and leptokurtic. Figure 1 shows the time plots of the selected price series. The time plots of all the price series show a similar pattern of price variation.

### 3.3. Test for stationarity

The stationarity of the time series is a prior assumption for the GARCH modelling. Using the Augmented Dickey-Fuller (ADF) test (Dickey and Fuller, 1979), Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski *et al.*, 1992), and the Phillips-Perron (PP) test (Phillips and Perron, 1988), the stationarity of the log return series and the squared log return series are tested (Table 2). For ADF and PP tests, the null hypothesis is that the unit root is present in the time series. For the KPSS test, the null hypothesis is that the unit root is not present in the time series. All three tests terminate the possibility of the

**Table 1: Descriptive statistics of selected price series**

| Statistics | Khedli | Atru | Nimbahera | Anoopgarh |
|---|---|---|---|---|
| Mean (Rs./q) | 4040.02 | 3883.61 | 3748.34 | 3863.27 |
| Median (Rs./q) | 3523.35 | 3589.03 | 3475.82 | 3588.00 |
| Minimum (Rs./q) | 1055.00 | 2026.00 | 2000.00 | 2108.00 |
| Maximum (Rs./q) | 8300.00 | 8091.00 | 7715.00 | 8031.00 |
| S.D. (Rs./q) | 1215.30 | 1232.00 | 1175.12 | 1234.29 |
| CV (%) | 30.08 | 31.72 | 31.35 | 31.95 |
| Skewness | 1.45 | 1.15 | 1.00 | 1.11 |
| Kurtosis | 1.32 | 0.84 | 0.39 | 0.75 |

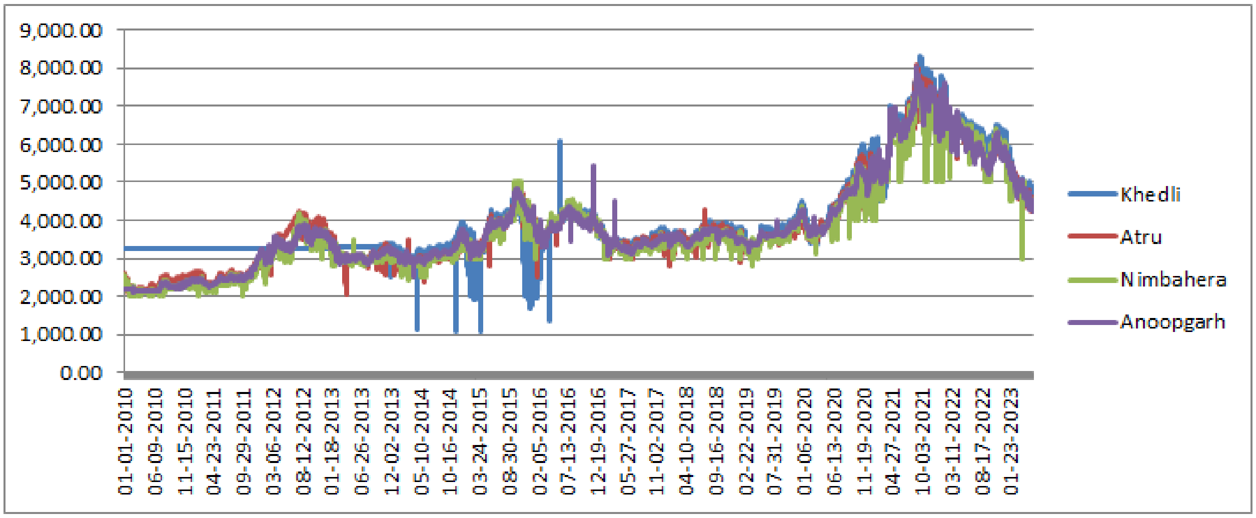


**Figure 1: Time plots of the daily price series**

presence of a unit root in the log return series and the squared log return series ($p$-values are given in parenthesis).

**Table 2: Test for stationarity**

| Market | Khedli | Atru | Nimbahera | Anoopgarh |
|---|---|---|---|---|
| Series | Log return | Log return | Log return | Log return |
| ADF | -23.44 | -18.5 | -18.89 | -17.97 |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| KPSS | 0.01 | 0.04 | 0.03 | 0.10 |
| | (0.10) | (0.10) | (0.10) | (0.10) |
| PP | -5674.1 | -6079.5 | -5850.8 | -5724.6 |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Series | Squared log return | Squared log return | Squared log return | Squared log return |
| ADF | -15.15 | -14.81 | -14.98 | -15.85 |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| KPSS | 0.39 | 0.25 | 1.22 | 0.41 |
| | (0.08) | (0.10) | (0.10) | (0.07) |
| PP | -2356.8 | -2527 | -2644 | -2781.9 |
| | (0.01) | (0.01) | (0.01) | (0.01) |

### 3.4. Test for long memory

The GPH test (Geweke and Porter-Hudak, 1983) is used to check the presence of long memory in the log return series and the squared log return series (Table 3). It is seen that the fractional differencing parameters for the log return series are not significant. But, they are significant for their corresponding squared log return series. It implies that the long memory is present in the squared log return series but not in the log return series.

**Table 3: GPH test**

| Market | Log return | | | Squared log return | | |
|---|---|---|---|---|---|---|
| | *d* | s.e. | Z | *d* | s.e. | Z |
| Khedli | -0.088 | 0.081 | -1.085 | 0.211 | 0.073 | 2.884 |
| Atru | 0.157 | 0.088 | 1.776 | 0.128 | 0.065 | 1.976 |
| Nimbahera | -0.137 | 0.097 | -1.403 | 0.24 | 0.106 | 2.259 |
| Anoopgarh | -0.095 | 0.096 | -0.987 | 0.224 | 0.086 | 2.596 |

### 3.5. ACF and PACF plots

The ACF and PACF plots help to examine the statistical relationships between the realizations of a time series through visualization. Figure 2 depicts the ACF and PACF plots of the selected log return series and the ACF plots of the squared log return series. The ACF and PACF plots of the log return series are decaying at exponential rates. It implies the absence of long memory in the mean model. But, hyperbolic decay is visible in the ACF plots of the squared log return series. It implies the presence of long memory in volatility. The GPH test's results also support the same conclusions.

### 3.6. Fitting of models

In the first step, the AR (1) model is fitted as the mean model in all the log return series. After that, the residuals are obtained and tested for the presence of conditional heteroscedasticity using the ARCH-LM test. The null hypothesis for this test is the absence of the ARCH effect in the residual series. It is seen that the ARCH-LM test is significant for all residual series and the presence of ARCH effect in all the residual series is confirmed. After that, the GARCH, EGARCH, APARCH, GJR-GARCH, and FIGARCH models are fitted to the residual series. The parameters are estimated using the maximum likelihood estimation procedure. The best fitted model for all the time series is chosen based on the degree of fitting in terms of three popularly used error functions namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) in the model building set. These error functions are calculated as

$$RMSE = \left[\frac{1}{k}\sum_{t=1}^{k}(y_t - \hat{y}_t)^2\right]^{\frac{1}{2}} \tag{13}$$

$$MAE = \frac{1}{k}\sum_{t=1}^{k}|y_t - \hat{y}_t| \tag{14}$$

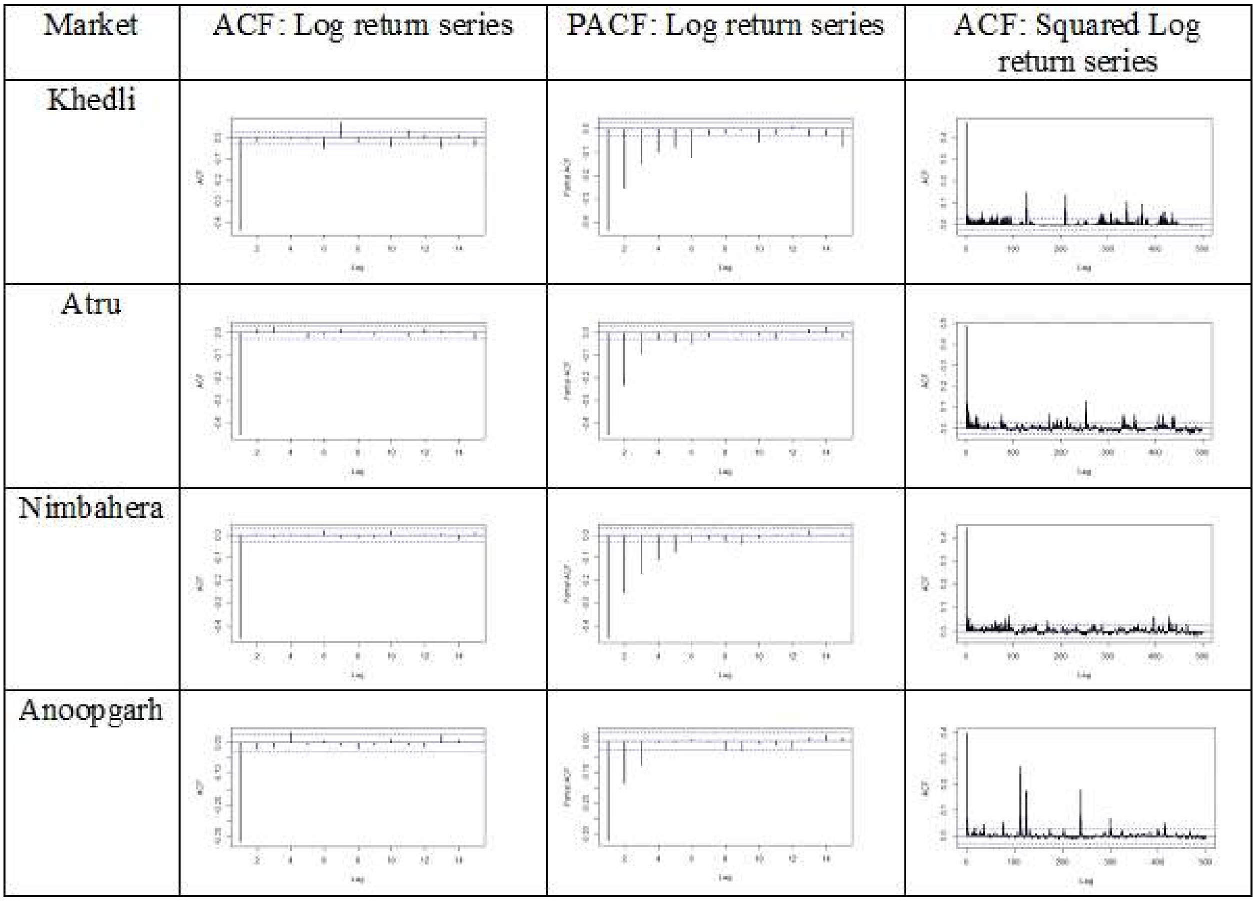


**Figure 2: ACF and PACF plots**

$$MAPE = \frac{1}{k}\sum_{t=1}^{k}\frac{|y_t - \hat{y}_t|}{y_t} \times 100 \tag{15}$$

where $k$ denotes the number of realizations used, $y_t$ is the observed value and $\hat{y}_t$ is the corresponding predicted value.

The estimated parameters of the best-fitted models are given in Table 4. It is seen that the FIGARCH model is the best-fitted model for all the markets. For all the series the parameters $\alpha_1$, $\beta_1$ and $d$ are highly significant. This implies that the current volatility significantly depends on previous volatility as well as previous shock. The presence of long memory is also significant for all cases. Fitting performances of the used models in the model building set in terms of RMSE, MAE and MAPE are given in Table 5. It is seen that for all the markets the best fitted model is the AR (1)-FIGARCH (1, $d$, 1) model. The ACF and PACF plots of the residual series, after fitting the AR (1)-FIGARCH (1, $d$, 1) model, for all the markets, do not exhibit any systematic trend and almost all the correlations lie within the 95% confidence interval.

In the model validation set, the rolling window forecast for 50 days, 100 days, 150 days, 200 days and 250 days are obtained and they are given in Table 6. It can be seen that for Khedli and Anoopgarh markets the forecasting performance is improving by increasing

**Table 4: Estimate of parameters of the best-fitted models**

| Market | Khedli | Atru | Nimbahera | Anoopgarh |
|---|---|---|---|---|
| parameter | AR(1) - FIGARCH<br>(1, $d$, 1) | AR(1) - FIGARCH<br>(1, $d$, 1) | AR(1) - FIGARCH<br>(1, $d$, 1) | AR(1) - FIGARCH<br>(1, $d$, 1) |
| | | Mean Model | | |
| Constant | 0.000<br>(0.000)*** | -0.001<br>(0.000)*** | -0.000<br>(0.000) | 0.000<br>(0.000) |
| AR(1) | -0.550<br>(0.000)*** | -0.327<br>(0.016)*** | -0.410<br>(0.016)*** | -0.230<br>(0.019)*** |
| | | Variance Model | | |
| Constant | 0.000<br>(0.000) | 0.000<br>(0.000)*** | 0.000<br>(0.000) | 0.000<br>(0.000)*** |
| $\alpha_1$ | 0.152<br>(0.000)*** | 0.961<br>(0.004)*** | 0.820<br>(0.194)*** | 0.311<br>(0.026)*** |
| $\beta_1$ | 0.649<br>(0.000)*** | 0.942<br>(0.003)*** | 0.797<br>(0.219)*** | 0.824<br>(0.012)*** |
| $d$ | 0.727<br>(0.000)*** | 0.577<br>(0.013)*** | 0.463<br>(0.050)*** | 0.631<br>(0.033)*** |

**Table 5: Fitting performance of the selected models in the model building set**

| Market | Model | RMSE | MAE | MAPE (%) |
|---|---|---|---|---|
| Khedli | AR(1)-GARCH (1,1) | 103.165 | 35.684 | 0.930 |
| | AR(1)-EGARCH(1,1) | 107.785 | 36.921 | 0.962 |
| | AR(1)-APARCH (1,1) | 99.267 | 36.163 | 0.702 |
| | AR(1)-GJRGARCH (1,1) | 98.542 | 36.408 | 0.688 |
| | AR(1)-FIGARCH (1, $d$, 1) | 67.986 | 23.223 | 0.575 |
| Atru | AR(1)-GARCH (1,1) | 39.918 | 20.478 | 0.520 |
| | AR(1)-EGARCH(1,1) | 39.871 | 18.858 | 0.478 |
| | AR(1)-APARCH (1,1) | 38.233 | 19.554 | 0.496 |
| | AR(1)-GJRGARCH (1,1) | 39.344 | 20.167 | 0.512 |
| | AR(1)-FIGARCH (1, $d$, 1) | 32.608 | 16.317 | 0.410 |
| Nimbahera | AR(1)-GARCH (1,1) | 56.147 | 28.414 | 0.755 |
| | AR(1)-EGARCH(1,1) | 50.373 | 25.513 | 0.678 |
| | AR(1)-APARCH (1,1) | 56.484 | 28.583 | 0.760 |
| | AR(1)-GJRGARCH (1,1) | 59.615 | 30.157 | 0.802 |
| | AR(1)-FIGARCH (1, $d$, 1) | 48.295 | 24.481 | 0.651 |
| Anoopgarh | AR(1)-GARCH (1,1) | 19.260 | 9.822 | 0.240 |
| | AR(1)-EGARCH(1,1) | 16.133 | 8.229 | 0.218 |
| | AR(1)-APARCH (1,1) | 16.999 | 8.670 | 0.212 |
| | AR(1)-GJRGARCH (1,1) | 18.385 | 9.376 | 0.229 |
| | AR(1)-FIGARCH (1, $d$, 1) | 14.963 | 7.652 | 0.202 |

the forecast horizon. For the Atru market, the numerical values of these three error functions first increase and then decrease. For the Nimbahera market, they decrease, then increase and again then decrease. All these are because of the presence of long memory in volatility. Long memory in volatility plays a crucial role in increasing the forecast efficiency while increasing the forecast horizon at different levels.

**Table 6: Rolling window forecasting performance of best-fitted models in the model validation set**

| Market | Model | Horizon | RMSE | MAE | MAPE (%) |
|---|---|---|---|---|---|
| Khedli | AR (1) - FIGARCH (1, $d$ , 1) | 50 | 138.469 | 89.017 | 1.595 |
| | | 100 | 131.694 | 87.371 | 1.529 |
| | | 150 | 129.799 | 86.693 | 1.488 |
| | | 200 | 124.836 | 85.922 | 1.472 |
| | | 250 | 117.413 | 81.497 | 1.427 |
| Atru | AR (1) - FIGARCH (1, $d$ , 1) | 50 | 75.871 | 61.047 | 1.054 |
| | | 100 | 84.717 | 66.929 | 1.145 |
| | | 150 | 89.824 | 68.346 | 1.207 |
| | | 200 | 88.569 | 65.951 | 1.206 |
| | | 250 | 87.919 | 65.031 | 1.202 |
| Nimbahera | AR (1) - FIGARCH (1, $d$ , 1) | 50 | 202.693 | 116.820 | 2.039 |
| | | 100 | 166.421 | 102.061 | 1.748 |
| | | 150 | 185.467 | 113.154 | 2.072 |
| | | 200 | 241.414 | 114.319 | 2.249 |
| | | 250 | 218.968 | 101.744 | 2.024 |
| Anoopgarh | AR (1) - FIGARCH (1, $d$ , 1) | 50 | 141.472 | 106.313 | 1.874 |
| | | 100 | 131.277 | 98.890 | 1.719 |
| | | 150 | 126.541 | 96.082 | 1.710 |
| | | 200 | 122.185 | 94.336 | 1.709 |
| | | 250 | 118.971 | 91.240 | 1.707 |

## 4. Conclusions

In this article, the mustard price volatility of four markets from the state of Rajasthan is studied. The presence of long memory in volatility for these series is confirmed using the GPH test. The GARCH, EGARCH, APARCH, GJR-GARCH and FIGARCH models are fitted to the log return series of the selected markets and it is seen that the FIGARCH model is the best fitted model for all the markets. The presence of long memory in volatility helps increase the model forecasting efficiency on a larger horizon. A better understanding of price volatility in the presence of long memory in volatility can help improve decision scenarios.

## References

Agarwal, M., Tripathi, P. K., and Pareek, S. (2021). Forecasting infant mortality rate of india using arima model: a comparison of bayesian and classical approaches. *Stat Appl*, **19**, 101–114.

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

Bhardwaj, S., Paul, R. K., and Singh, K. (2015). Price forecast an instrument for improvement in agricultural production and marketing in rajasthan-a case study of rape & mustard seeds. *Indian Journal of Agricultural Marketing*, **29**, 155–163.

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

Box, G. E. P. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control.* Holden- Day: San Francisco, CA, USA.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Ding, Z., Granger, C. W., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83–106.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, **50**, 987–1007.

Garai, S., Paul, R. K., Rakshit, D., Yeasin, M., Emam, W., Tashkandy, Y., and Chesneau, C. (2023). Wavelets in combination with stochastic and machine learning models to predict agricultural prices. *Mathematics*, **11**, 2896.

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, **4**, 221–238.

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

Granger, C. W. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, **1**, 15–29.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, **54**, 159–178.

Mitra, D., Paul, R. K., Paul, A. K., et al. (2018). Statistical modelling for forecasting volatility in potato prices using arfima-figarch model. *Indian Journal of Agricultural Sciences*, **88**, 268–272.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, **59**, 347–370.

Paul, R., Prajneshu, and Ghosh, H. (2009). Garch nonlinear time series analysis for modelling and forecasting of india's volatile spices export data. *Journal of the Indian Society of Agricultural Statistics*, **63**, 123–131.

Paul, R. K. (2014). Forecasting wholesale price of pigeon pea using long memory time-series models. *Agricultural Economics Research Review*, **27**, 167–176.

Paul, R. K., Alam, W., and Paul, A. K. (2014). Prospects of livestock and dairy production in india under time series framework. *Indian Journal of Animal Sciences*, **84**, 130–134.

Paul, R. K., Bhardwaj, S. P., Singh, D. R., Kumar, A., Arya, P., and Singh, K. N. (2015). Price volatility in food commodities in india-an empirical investigation. *International Journal of Agricultural and Statistical Sciences*, **11**, 395–401.

Paul, R. K. and Garai, S. (2021). Performance comparison of wavelets-based machine learning technique for forecasting agricultural commodity prices. *Soft Computing*, **25**, 12857–12873.

Paul, R. K., Gurung, B., Paul, A. K., and Samanta, S. (2016). Long memory in conditional variance. *Journal of the Indian Society of Agricultural Statistics*, **70**, 243–254.

Paul, R. K., Paul, A., and Bhar, L. (2020). Wavelet-based combination approach for modeling sub-divisional rainfall in india. *Theoretical and Applied Climatology*, **139**, 949–963.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression.

*Biometrika*, **75**, 335–346.

Rakshit, D., Paul, R. K., and Panwar, S. (2021). Asymmetric price volatility of onion in india. *Indian Journal of Agricultural Economics*, **76**, 245–260.

Rakshit, D., Paul, R. K., Yeasin, M., Emam, W., Tashkandy, Y., and Chesneau, C. (2023). Modeling asymmetric volatility: A news impact curve approach. *Mathematics*, **11**, 2793.

Rakshit, D., Roy, A., Atta, K., Adhikary, S., and Vishwanath (2022). Modeling temporal variation of particulate matter concentration at three different locations of delhi. *International Journal of Environment and Climate Change*, **12**, 1831–1839.

Tayefi, M. and Ramanathan, T. (2012). An overview of figarch and related time series models. *Austrian Journal of Statistics*, **41**, 175–196.

Taylor, S. J. (1986). *Modelling Financial Time Series*. John Wiley Sons, Ltd., Chichester, Great Britain.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 119-129

# Assessment of Child Mortality and its Socioeconomic and Demographic Determinants- Evidences from the Latest National Family Health Survey of India

**Anuradha Rajkonwar Chetiya**[1] **and Vishal Deo**[2]
[1]*Department of Statistics, Ramjas College, University of Delhi, Delhi, India*
[2]*National Institute of Medical Statistics, ICMR, New Delhi, India*

Received: 10 July 2023; Revised: 27 July 2023; Accepted: 29 July 2023

## Abstract

Even though the childhood mortality rates have been on a steady decline in India, they are still unacceptably high across many parts of the country. As per WHO estimates, India experienced 490 thousand new born deaths in 2020, the highest in the world. To achieve the Sustainable Development Goal (SDG) goal of reducing under-five mortality to at least as low as 25 per 1,000 live births in every country by 2030, India needs to identify factors acting as barriers in the implementation of health policies and programmes to improve accessibility, utilization and outreach of quality public healthcare systems to all.

The objective of this study is to identify and assess the impact of demographic, socio-economic and health resource factors associated with infant mortality in India. A comparative assessment of the current status of child mortality in India has been presented. Further, risk of infant deaths associated with these factors has been evaluated using a binary logistic regression. Individual child level data from the fifth National Family Health Survey (NFHS 5) has been used for the analysis.

Out of the four factors included in the model, education level of mother has come out to be the most significant determinant of infant mortality. Results show that the odds of infant mortality increase consistently as the education level of mother decreases. Those born to a mother with no education are at more than two times risk of dying within 1 year as compared to those born to a mother with higher education. An interesting finding, contrary to the historical trend, is that the risk of infant mortality in male child is significantly higher (by around 25%), as compared to that of a female child.

*Key words:* Maternal and Child Health; Healthcare; Under-five mortality; Infant mortality; Excess girl child mortality; Female literacy.

Corresponding Author: Anuradha Rajkonwar Chetiya
Email: anuradha_rc@hotmail.com

## 1. Introduction

There is a potential association between the causes of infant mortality and factors that are likely to influence health status of the whole population, see Crevoiserat and Kim (2013). Three significant measures of child mortality are neonatal mortality – probability of dying in the first month of life, infant mortality – probability of dying before reaching the first birthday, and the under-five mortality – probability of dying before the fifth birthday. India as a member state of the United Nations had adopted the now named Sustainable Development Goals (SDGs) goals, previously known as the Millennium Development Goals, since 2000. These goals, 17 in all, aim at reducing economic and social inequities among nations - 'they recognize that ending poverty and other deprivations must go hand-in-hand with strategies that improve health and education, reduce inequality, and spur economic growth – all while tackling climate change and working to preserve our oceans and forests (https://sdgs.un.org/goals).The Sustainable Development Goal 3 (SDG 3) - Good Health and Well Being, is aimed at ensuring healthy lives and promoting well-being for all at all ages. In particular, the target of goal 3.2 is to end preventable deaths of newborns to at least as low as 12 per 1000 live births in every country and reduce under-five mortality to at least as low as 25 per 1,000 live births in every country by 2030. In India, recent data from the NFHS 5 survey report of 2019-21 have estimated neonatal mortality at 25, infant mortality rate at 35 and the under-five mortality at 42 per 1000 live births. In terms of actual number of child deaths, this is very high considering India is now expected to surpass China as the most populous country in the world with a population exceeding one billion. UNICEF has reported an estimate of 490 thousand newborn deaths in India in 2020 (Figure 1). This is nearly twice that of Nigeria which has the second highest estimated number of new born deaths at 271 thousand.

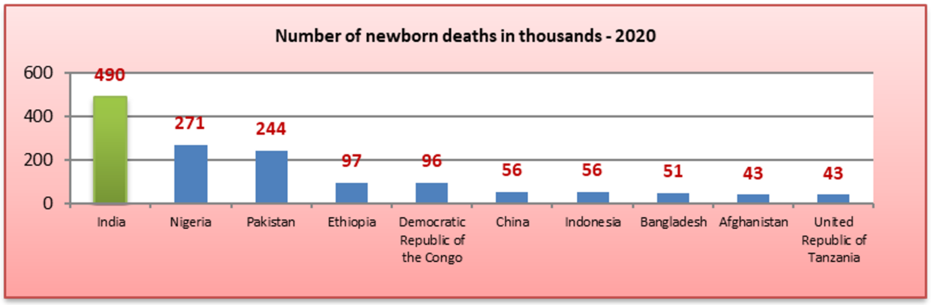


**Figure 1: Number of new born deaths (in thousands) in 2020 [Data Source: https://www.unicef.org]**

These numbers for India remain worrisome, even after considerable reductions in the last four decades. Evolving evidences on demographic and socioeconomic determinants of child mortality from latest national surveys will be crucial to inform health policies and overcome programmatic gaps. This paper examines the current status of child mortality in India using survey data from various sources, primarily the NFHS 5. It also analyzes the impact of some determinants currently associated with infant mortality in India.

Section 2 contains a comparative analysis of the current state of child mortality in India with respect to the other G20 nations. In section 3, we have presented a synopsis of the major socio-economic and demographic determinants of child mortality in India based on exploratory findings from the NFHS 5 and SRS datasets. Furthermore, based on the

NFHS 5 data, impact of these determinants on infant mortality in India has been assessed using multiple logistic regression in section 4. Section 5 provides a comprehensive discussion on the findings of the sections 2, 3, and 4.

## 2. Child mortality in India – An assessment

The Infant Mortality Rate (IMR) of India was recorded as 125 per 1000 live births in 1978 and, almost forty five years later, it now stands at 28 per 1000 live births in 2020 (SRS bulletins, 1997-2021). The World Bank estimate of the IMR of India for the same year is 27 per 1000 live births. The overall IMR and the IMRs for both male and female child have been on a decline since 1998 as evident from the data (Figure 2). Despite the substantial drop over the years, the current IMR for India (2020) does not compare well with the other G20 countries. G20 is a consortium of 19 countries and the European Union that together represent around 85% of the global GDP, over 75% of the global trade, and about two-thirds of the world population (https://www.g20.org). G20 initially focused on broad macroeconomic issues but it has since expanded its agenda to include sustainable development and health in its ambit. According to estimates of child mortality of the G20 nations in 2020 (https://data.worldbank.org), neonatal mortality rate of India is highest among these nations at 20.3 (Figure 3). 15 of these nations are below the 5 mark, with Indonesia having the next highest at 11.7. IMR of India is also the highest (Figure 4) at 27, with Japan having the lowest at 1.8. Again, 15 of these G20 nations have an IMR less than 10. The under-five mortality rate of India stands at 32.6 with Japan being the lowest again at 2.5 (Figure 5).

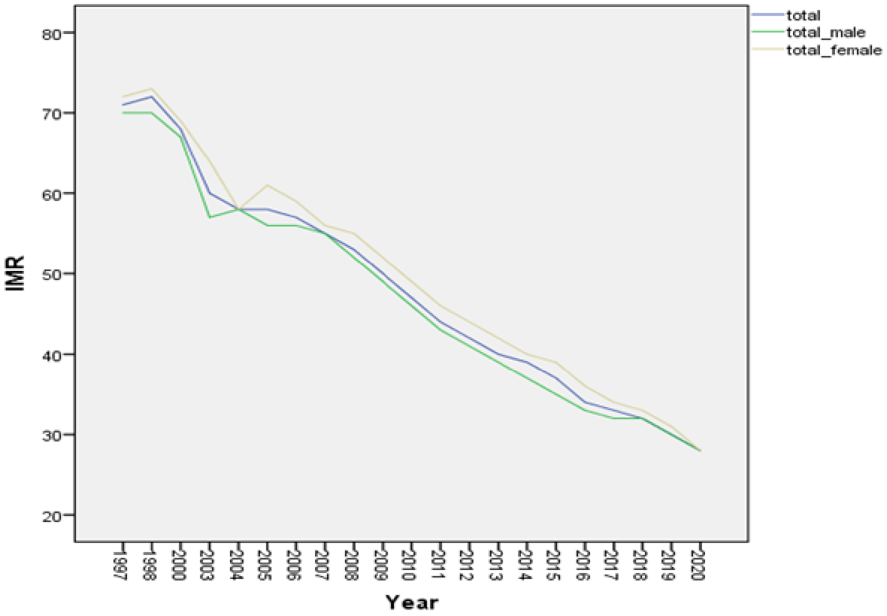


**Figure 2: IMR trend in India [Data Source: SRS bulletins, 1997–2021]**

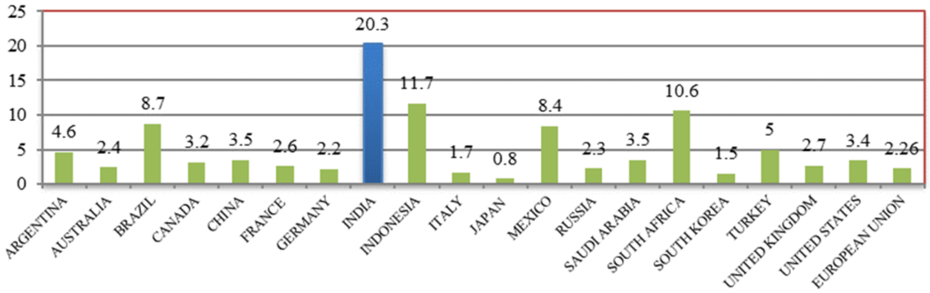


**Figure 3: G20 nations: neonatal mortality rates (2020) [Data Source: https://data.worldbank.org]**

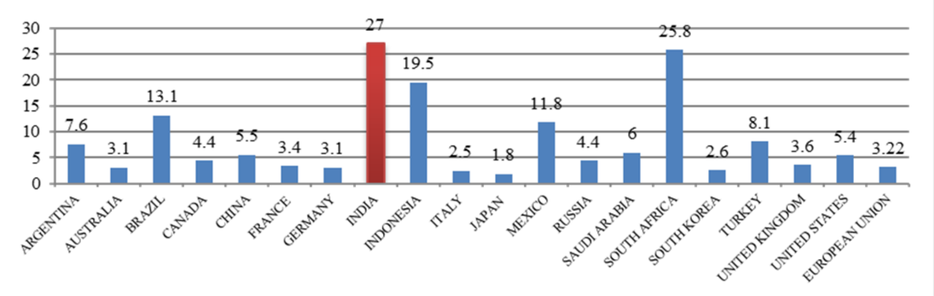


**Figure 4: G20 nations: infant mortality rates (2020) [Data Source: https://data.worldbank.org)]**

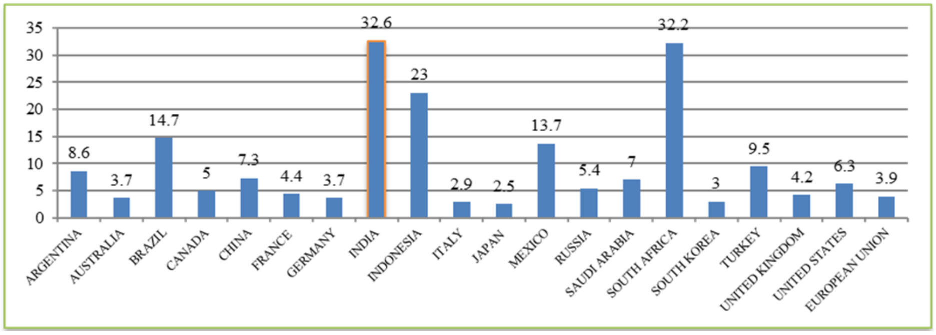


**Figure 5: G20 nations: under-five mortality rates (2020) [Data Source: https://data.worldbank.org]**

## 3. Socioeconomic and demographic risk factors of child mortality

According to World Bank estimates (https://data.worldbank.org), 65% population of India lives in rural areas. Statistics also indicate that the IMR has been uneven across the rural urban divide (Figure 6). The National Rural Health Mission was launched in 2005. One of the targets was to reduce IMR in rural areas to 30 by 2012. The Janani Suraksha Yojana (JSY) is one such program under the National Rural Health Mission. This program was introduced in 2005 with the objective of reducing maternal and neonatal mortality by promoting institutional delivery among poor pregnant women. The JSY is currently being implemented through the Accredited Social Health Activists (ASHA) and Anganwadi Workers (AWW). Under this scheme, a comprehensive package of free and cashless services is offered to all pregnant women, and sick infants up to the age of one year, in government health institutions. The Janani Shishu Suraksha Karyakram (JSSK), thereby is aimed at reducing financial barriers to care and improving access to health services by eliminating out of pocket expenditure in all government facilities.

Twenty years of IMR data of rural and urban population from 2000 to 2020, from the SRS bulletins was examined and analysed. It was observed that the IMR has gone down steadily in both rural and urban areas from 2000 to 2020. The gap between the two still remains high with IMR in urban areas at 19 and in rural areas is higher at 31 per thousand in 2020 (Figure 6). A t -test confirms that this difference between IMR values in rural and urban areas in 2020 is statistically significant.

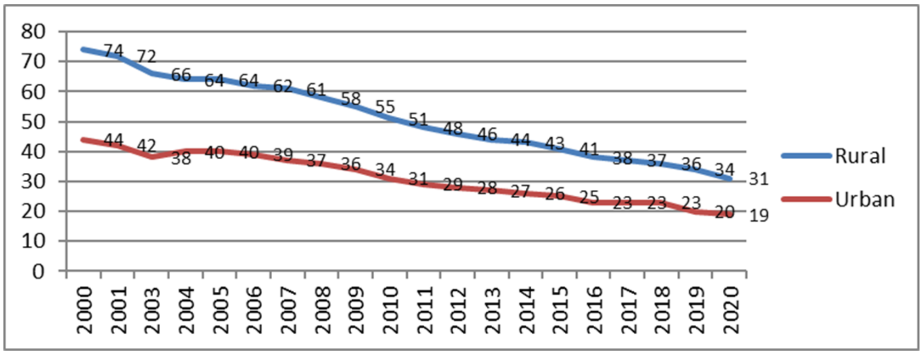


**Figure 6: IMR in rural and urban areas in India, 2000-2020 [Data Source: SRS Bulletin, 2000- 2020]**

Several studies have established that there is an inverse relationship between female literacy and infant mortality rate (Rao *et al.* (1996), Gokhale *et al.* (2002), Gakidou *et al.* (2010), Singh *et al.* (2011), Balaj *et al.* (2021), Okui (2023)). A meta analysis of surveys from 92 countries by Balaj *et al.* (2021) observed a reduction in under-5 mortality of $31 \cdot 0\%$ for children born to mothers with 12 years of education (*i.e.*, completed secondary education). A basic minimum level of education empowers females and helps creates awareness about health practices. Maternal health is an immediate and important factor in determining child mortality. Factors such as low birth weight, nutritional deficiency in infants are all tied to maternal health and can affect the child's survival. Socioeconomic and demographic factors identified in various studies include nutritional status of mother, age of the mother, gaps between two deliveries, access to healthcare services that ensure safe delivery along with ante natal and post natal care (Thakkar *et al.* (2023), Patel and Olickal (2021), Bora (2020) Singh *et al.* (2011)). From the observed data of educational status and percentage of institutional deliveries obtained from the NFHS 5 survey, it can be seen in Figure7 that higher the education level of the mother, the more likely she is to go for a safer delivery in a health facility with trained medical staff. Among women who had completed 12 years of schooling and above, 97 percent had opted for institutional deliveries as compared to women with no schooling among which the percentage was 75.

From Table 1, it can be observed that the most common reason for not delivering in a health facility for both rural as well as urban areas was that the woman did not think it was necessary. In rural areas 19.5 percent women said that the husband or family did not allow them to have the delivery in a health facility, 17.4 percent of women said that a health facility was too far or there was no transportation, and 15.1 percent said it costs too much. 27.6 percent women in rural and 30.5 percent women in urban areas did not feet that it was necessary.

**Table 1: Rural vs. urban response percentage on reasons for not delivering in a health facility**

| Sl. no. | Reason for not delivering in health facility | Urban | Rural |
|---|---|---|---|
| 1 | Costs too much | 15.2 | 15.1 |
| 2 | Facility not open | 9.1 | 9.8 |
| 3 | Too far/ no transportation | 12.4 | 17.4 |
| 4 | Don't trust/poor quality service | 6.8 | 4.7 |
| 5 | No female provider at facility | 4.3 | 3.9 |
| 6 | Husband /family did not allow | 18.1 | 19.5 |
| 7 | Not necessary | 30.5 | 27.6 |
| 8 | Not customary | 3.6 | 3.5 |
| 9 | Other | 19.1 | 16.4 |

Data Source: NFHS 5, 2019-21

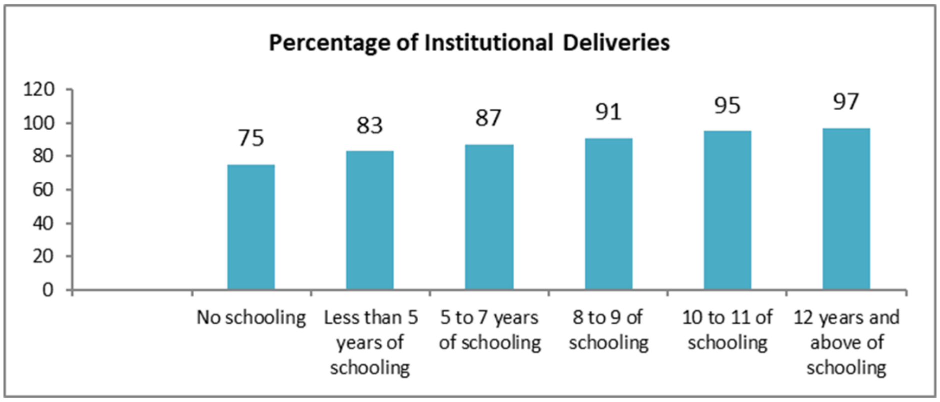


**Figure 7: Percentage of institutional deliveries for different levels of schooling [Data Source: NFHS 5, 2019- 2021]**

In NFHS 5 data, wealth index is a composite measure of a household's cumulative living standard and relative economic status. The wealth index is calculated using data on a household's ownership of certain selected assets which include consumer items such as a television and car; dwelling characteristics such as flooring material, type of drinking water source, toilet facilities and other characteristics that are related to wealth status (NFHS 5 India Report). On the basis of household scores, the population is divided into five equal categories (quintiles) each consisting of 20% of the population. Table 2 presents the wealth index from the NFHS 5 survey report (2019-21). The rural urban divide in economic condition is further evident from the data in this table. Nearly 76% of the wealthiest population falling in the highest two quintiles reside in urban areas as opposed to 24% in rural areas, and more than 50% of the rural population falls in the lowest two quintiles.

Table 3 indicates that, in India, the brunt of high child deaths is borne by the marginalized and socioeconomically disadvantaged sections of the population. For example, the Infant Mortality Rate in the poorest 20 percent of the population is more than 3 times higher than that in the richest 20 percent of the population. This means that an

**Table 2: Distribution (in%) of wealth index by residency**

| Residency | Lowest | Second | Middle | Fourth | Highest | Total |
|---|---|---|---|---|---|---|
| India | 20 | 20 | 20 | 20 | 20 | 100 |
| Rural | 3.2 | 7.2 | 15.5 | 28.6 | 45.5 | 100 |
| Urban | 27.8 | 26 | 22.1 | 16 | 8.1 | 100 |

Data Source: NFHS 5 Report, 2019-21

infant born in a relatively poor family is more than three times likely to die in infancy than an infant born in a better off family. Similar conclusions follow for neonatal and under-five mortality. A study by Goel *et al.* (2015), investigated the link between maternal health and wealth index. They concluded that the number of antenatal care (ANC) visits increased as the wealth index increased. Another study by Thakkar *et al.* (2023) came up with similar results - women with less formal education, from poorer households and belonging to rural areas had higher odds of inadequate visits. Among the reasons given for not delivering in a health facility – ‘it costs too much’ and ‘was too far / no transportation’, together constituted 27.6% respondents in urban and 32.5% respondents in urban areas respectively.

**Table 3: Child mortality rates by wealth quintiles**

| Wealth Quintile | Neonatal Mortality | Infant Mortality | Under- Five Mortality |
|---|---|---|---|
| Lowest | 39.2 | 53.1 | 63.4 |
| Second | 25.4 | 34.8 | 43.6 |
| Middle | 22.3 | 34.5 | 40.2 |
| Fourth | 19.4 | 29.2 | 33.7 |
| Highest | 10.9 | 16.2 | 19.4 |

Data Source: NFHS 5 Report, 2019-21

## 4. Risk assessment of determinants of infant mortality using NFHS 5 data

### 4.1. Socioeconomic and demographic determinants

Based on the review presented in the previous sections, four major factors are selected for generating evidence on associated risk of infant mortality using the NFHS 5 data. These are - type of place of residence, education level of the mother, wealth index of the household, and sex of the child. Type of place of residence is classified into two categories- rural and urban. Four levels of education were considered - no education, primary education, secondary education and higher education. All five quintiles of the wealth index from NFHS 5 survey are considered – poorest, poorer, middle, richer and richest are considered.

### 4.2. Methodology and results

Individual level data of children from NFHS 5 (file name: IAKR7EDT), downloaded from the website of DHS [www.dhsprogram.com], was used for the analyses. Respondents (mothers) are between 15 to 49 years of age. Since the age at death is in rounded-off months, deaths till 11 months of age have been categorized as infant mortality. Children who were alive at the time of interview and were 12 months or older are considered as those who did not experience infant mortality. Summary of the data by different factors are provided in

Table 4. A binary logistic regression with response variable as infant mortality status of children has been fitted with the four factors (type of place of residence, education level of the mother, wealth index of the household, and sex of the child), while adjusting for the age of the respondent (mother) at the birth of her first child. Results of the fitted logistic regression are presented in Table 5. All analyses have been performed using R software.

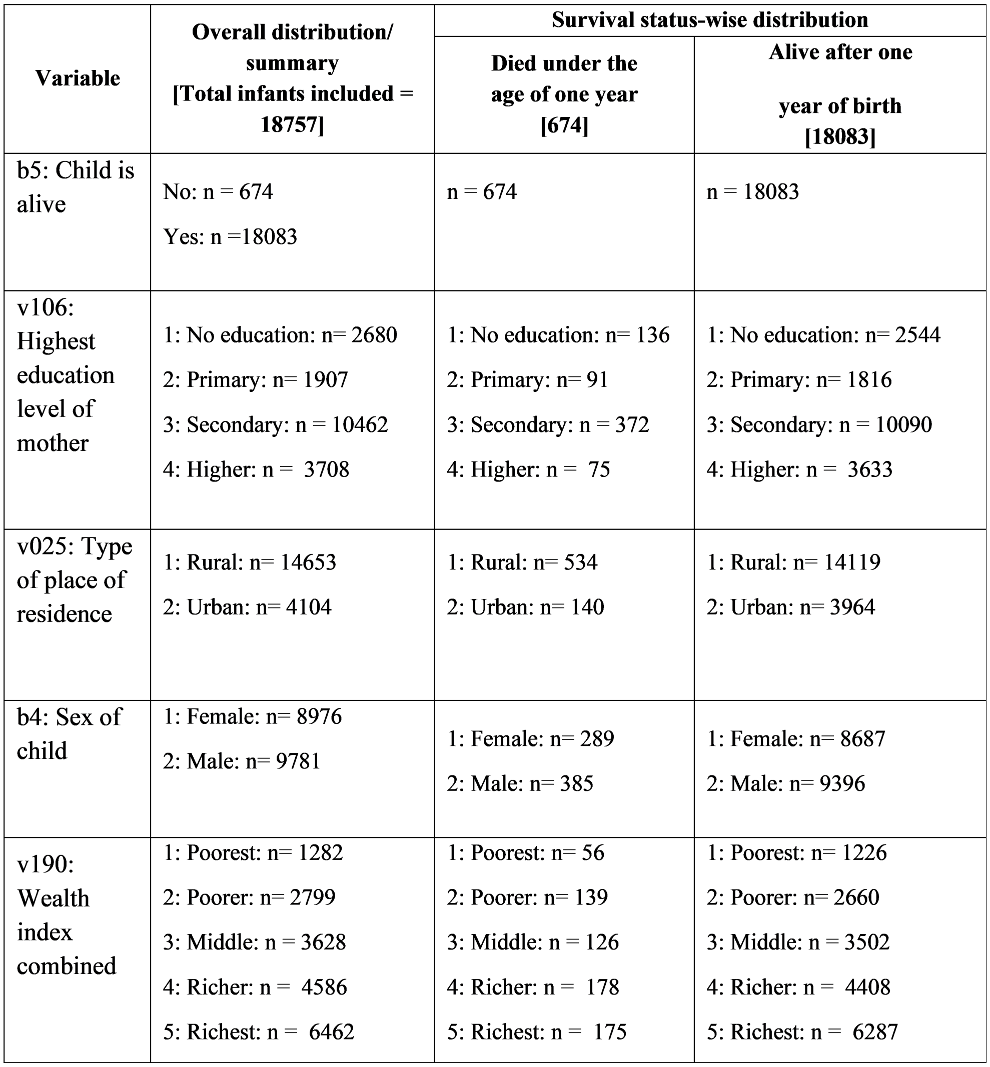
Out of the four factors included in the model, education level of mother has come out to be the most significant determinant of infant mortality. Results show that the odds of infant mortality increase consistently as the education level of mother decreases. Those born to a mother with no education are at more than two times risk of dying within 1 year as compared to those born to a mother with higher education (college and above). Similarly, odds of infant mortality among children born to mothers with primary and secondary education are around 2 times and 1.6 times of the odds for those born to mothers with higher education. Although the estimated odds of infant mortality associated with lower wealth index quintiles, as compared to that for the Richest group, are all higher, only two of the odds ratios are statistically significant. That is, it does indicate higher odds of infant mortality in relatively poorer households, but the odds do not increase consistently through the subsequent lower levels of wealth quintiles. An interesting result is that there is no significant difference in the odds of infant mortality among children born in households residing in rural areas as compared to those residing in urban households. The risk of infant mortality in male child is significantly higher (by around 25%), as compared to that of a female child.

## 5. Discussion

Child mortality rates in India are highly variable across the rural urban divide. Despite two decades of implementation of policies and programs to improve child mortality with particular focus in rural areas, IMR in rural areas of India continue to be significantly and consistently higher than in urban areas. The insignificant odds ratio of infant mortality in urban areas as compared to rural areas, may be indicative of the narrowing gap, but in terms of the IMR, NFHS 5 India report specifies that the under-five and infant mortality rates are still considerably higher in rural areas than in urban areas. Similar conclusions were given by Kumar *et al.* (2022) based on data from the earlier NFHS 4 survey. Their study found an existing rural-urban gap in under-five mortality and the authors suggested that the social and health policies need to reach rural children from poor families and uneducated mothers.

Based on the results of the present study we can conclude that female literacy remains one of the risk factors associated with child mortality in India. Improvements in the literacy rate of women will have a positive impact in reducing child mortality. The fact that 27.6 percent women in rural and 30.5 percent women in urban areas felt that it was not necessary to go for institutional deliveries, indicates a lack of awareness about safe deliveries, importance of antenatal and postnatal care, proper nutrition of the mother. It reflects a casual approach towards the birthing process. With reproductive and child health services being improved through Health and Wellness centres and primary health care centres, the decision to avail Antenatal care (ANC) and Postnatal care (PNC) services may depend more on awareness than on economic status. Consequently, the education level of mother can be expected to be a more significant determinant of infant mortality than household wealth index quintiles. Ensuring education of women up to a minimum level would play a vital role here to empower women to make better choices regarding health services. Policies and programs need to be

**Table 4: Data summary and distribution with respect to factors**

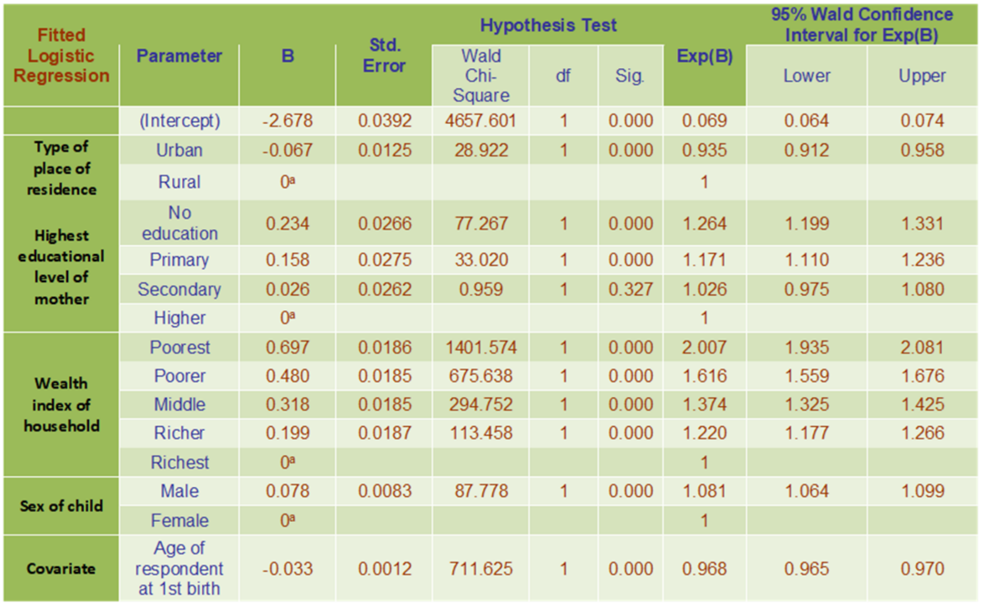


| Variable | Overall distribution/ summary [Total infants included = 18757] | Survival status-wise distribution | |
|---|---|---|---|
| | | Died under the age of one year [674] | Alive after one year of birth [18083] |
| b5: Child is alive | No: n = 674<br>Yes: n =18083 | n = 674 | n = 18083 |
| v106: Highest education level of mother | 1: No education: n= 2680<br>2: Primary: n= 1907<br>3: Secondary: n = 10462<br>4: Higher: n = 3708 | 1: No education: n= 136<br>2: Primary: n= 91<br>3: Secondary: n = 372<br>4: Higher: n = 75 | 1: No education: n= 2544<br>2: Primary: n= 1816<br>3: Secondary: n = 10090<br>4: Higher: n = 3633 |
| v025: Type of place of residence | 1: Rural: n= 14653<br>2: Urban: n= 4104 | 1: Rural: n= 534<br>2: Urban: n= 140 | 1: Rural: n= 14119<br>2: Urban: n= 3964 |
| b4: Sex of child | 1: Female: n= 8976<br>2: Male: n= 9781 | 1: Female: n= 289<br>2: Male: n= 385 | 1: Female: n= 8687<br>2: Male: n= 9396 |
| v190: Wealth index combined | 1: Poorest: n= 1282<br>2: Poorer: n= 2799<br>3: Middle: n = 3628<br>4: Richer: n = 4586<br>5: Richest: n = 6462 | 1: Poorest: n= 56<br>2: Poorer: n= 139<br>3: Middle: n = 126<br>4: Richer: n = 178<br>5: Richest: n = 175 | 1: Poorest: n= 1226<br>2: Poorer: n= 2660<br>3: Middle: n = 3502<br>4: Richer: n = 4408<br>5: Richest: n = 6287 |

designed to ensure that women complete at least a basic minimum level of 12 to 15 years of schooling for the effects of education to reflect truly on the individuals.

The results for male and female IMR indicate that a male child is at a significantly higher risk of mortality than a female child during the first year of life. Historically, IMR for females in India have been higher than that of males (Figure 2). However, as per the finding from our study, the risk of infant mortality is higher for male children than female children. This is corroborated by the higher male IMR as per the NFHS 5 India report which may be indicative of a recent change in trend in male and female IMR in India. Some earlier studies, like, Graunt (1977), Naeye *et al.* (1971), Waldron (1983), have attributed childhood mortality differences in sex to genetic and biological factors arguing that male children are more susceptible to diseases as compared to their female counterparts, and hence have lower survival rates. However, some later studies, like, Garenne (2003), and Pongou (2013), have argued that while prenatal environment and child biology are important contributing factors to sex differences in infant mortality, the effect of biology is much less important than the literature suggests. In the absence of any conclusive evidence on this research question, there is a need for further investigation through well designed India-specific studies to identify and understand the possible changes in the infant mortality trends in India.

**Table 5: Results of the logistic regression**



| Fitted Logistic Regression | Parameter | B | Std. Error | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Wald Chi-Square | df | Sig. | | Lower | Upper |
| | (Intercept) | -2.678 | 0.0392 | 4657.601 | 1 | 0.000 | 0.069 | 0.064 | 0.074 |
| Type of place of residence | Urban | -0.067 | 0.0125 | 28.922 | 1 | 0.000 | 0.935 | 0.912 | 0.958 |
| | Rural | $0^a$ | | | | | 1 | | |
| Highest educational level of mother | No education | 0.234 | 0.0266 | 77.267 | 1 | 0.000 | 1.264 | 1.199 | 1.331 |
| | Primary | 0.158 | 0.0275 | 33.020 | 1 | 0.000 | 1.171 | 1.110 | 1.236 |
| | Secondary | 0.026 | 0.0262 | 0.959 | 1 | 0.327 | 1.026 | 0.975 | 1.080 |
| | Higher | $0^a$ | | | | | 1 | | |
| Wealth index of household | Poorest | 0.697 | 0.0186 | 1401.574 | 1 | 0.000 | 2.007 | 1.935 | 2.081 |
| | Poorer | 0.480 | 0.0185 | 675.638 | 1 | 0.000 | 1.616 | 1.559 | 1.676 |
| | Middle | 0.318 | 0.0185 | 294.752 | 1 | 0.000 | 1.374 | 1.325 | 1.425 |
| | Richer | 0.199 | 0.0187 | 113.458 | 1 | 0.000 | 1.220 | 1.177 | 1.266 |
| | Richest | $0^a$ | | | | | 1 | | |
| Sex of child | Male | 0.078 | 0.0083 | 87.778 | 1 | 0.000 | 1.081 | 1.064 | 1.099 |
| | Female | $0^a$ | | | | | 1 | | |
| Covariate | Age of respondent at 1st birth | -0.033 | 0.0012 | 711.625 | 1 | 0.000 | 0.968 | 0.965 | 0.970 |

To conclude, socioeconomic and demographic factors continue to contribute to the disparities in the risk of infant mortality. These factors have the potential to create barriers for effective implementation of health programmes. With thousands of children in India still not being able to make it beyond the initial crucial years of their lives, there is an urgent need to identify and address such barriers in implementation and utilization of public healthcare programmes. Such steps will be imperative for India to achieve the Sustainable Development Goal (SDG) goal of reducing under-five mortality to at least as low as 25 per 1,000 live births by 2030.

A limitation of this analysis is that regional variations have not been considered and analysis has been performed at the national level only. Also, factors, like ANC visits, PNC, *etc.* have not been included as there was a lot of missing data.

## Acknowledgements

We are indeed grateful to the Editors and the reviewers for their guidance and counsel towards improving this paper. I am also thankful to the organizers of the Conference for giving me the opportunity to present our work.

## References

Balaj, M., York, H. W., Sripada, K., Besnier, E., Vonen, H. D., Aravkin, A., Friedman, J., Griswold, M., Jensen, M. R., Mohammad, T., Mullany, E. C., Solhaug, S., Sorensen, R., Stonkute, D., Tallaksen, A., Whisnant, J., Zheng, P., Gakidou, E., and Eikemo,

T. A. (2021). Parental education and inequalities in child mortality: a global systematic review and meta-analysis. *The Lancet*, **398**, 608–620.

Bora, J. K. (2020). Factors explaining regional variation in under-five mortality in india: An evidence from nfhs-4. *Health & Place*, **64**, 102363.

Crevoiserat, J. and Kim, J. (2013). Infant mortality kansas. Technical report, Kansas Department of Health and Environment.

Gakidou, E., Cowling, K., Lozano, R., and Murray, C. J. (2010). Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *The Lancet*, **376**, 959–974.

Garenne, M. (2003). Sex differences in health indicators among children in african dhs surveys. *Journal of Biosocial Science*, **35**, 601–614.

Goel, M., Roy, P., Rasania, S., Roy, S., Kumar, Y., and Kumar, A. (2015). Wealth index and maternal health care: Revisiting nfhs-3. *Indian Journal of Public Health*, **59**, 217.

Gokhale, M. K., Rao, S. S., and Garole, V. R. (2002). Infant mortality in india: use of maternal and child health services in relation to literacy status. *Journal of health, population, and nutrition*, **20**, 138–47.

Graunt, J. (1977). *Mathematical Demography*, chapter Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality, pages 11–20. Springer Berlin, Heidelberg.

Kumar, C., Piyasa, and Saikia, N. (2022). An update on explaining the rural-urban gap in under-five mortality in india. *BMC Public Health*, **22**, 2093.

Naeye, R. L., Burt, L. S., Wright, D. L., Blanc, W. A., and Tatter, D. (1971). Neonatal mortality, the male disadvantage. *Pediatrics*, **48**, 902–6.

Okui, T. (2023). Association between infant mortality and parental educational level: An analysis of data from vital statistics and census in japan. *PLOS ONE*, **18**, e0286530.

Patel, N. and Olickal, J. J. (2021). Maternal and child factors of under-five mortality in india. findings from nfhs-4. *Clinical Epidemiology and Global Health*, **12**, 100866.

Pongou, R. (2013). Why is infant mortality higher in boys than in girls? a new hypothesis based on preconception environment and evidence from a large sample of twins. *Demography*, **50**, 421–444.

Rao, R. S. P., Chakladar, B. K., Nair, N. S., Kutty, P. R., Acharya, D., Bhat, V., Chandrasekhar, S., Rodrigues, V. C., Kumar, P., Nagaraj, K., Prasad, K. N., and Krishnan, L. (1996). Influence of parental literacy and socio-economic status on infant mortality. *The Indian Journal of Pediatrics*, **63**, 795–800.

Singh, A., Pathak, P. K., Chauhan, R. K., and Pan, W. (2011). Infant and child mortality in india in the last two decades: A geospatial analysis. *PLoS ONE*, **6**, e26856.

Thakkar, N., Alam, P., and Saxena, D. (2023). Factors associated with underutilization of antenatal care in india: Results from 2019–2021 national family health survey. *PLOS ONE*, **18**, e0285454.

Waldron, I. (1983). Sex differences in human mortality: The role of genetic factors. *Social Science & Medicine*, **17**, 321–333.

Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 131–137

# Statistical and Artificial Neural Network Approaches for the Classification of Rice Genotypes based on Morphological Characters

**Shavi Gupta[1], Manish Sharma[1], S. E. H. Rizvi[1], R. K. Salgotra[2] and S. K. Gupta[3]**
[1]*Division of Statistics and Computer Science, SKUAST- Jammu*
[2]*School of Bio Technology, SKUAST- Jammu*
[3]*Division of Silviculture and Agroforestry, SKUAST- Jammu*

Received: 12 July 2023; Revised: 28 July 2023; Accepted: 30 July 2023

## Abstract

Classification is that arena of science which deals with grouping of objects on the basis of information available about those objects. It plays a significant role for planning purposes in agriculture system. The aim of this study was to classify the rice genotypes using statistical methods like discriminant analysis and artificial neural network (ANN), such as multilayer perceptron neural network for different classes of yield. These methods are fitted to primary data recorded for 100 genotypes of rice for five morphological variables and the data has been collected from the trial laid in SKUAST, Jammu. The class variable grain yield was categorized into 3 classes and was considered as dependent variable and all morphological characters as independent variables. The ability measures of classification such as Accuracy Rate and Kappa Statistics were used for testing samples. Number of days for full maturity was found to be important attributing character followed by number of effective tillers per plant for classification. Artificial Neural Network model (85 %) performed better than Discriminant Analysis (75 %) for classification of genotypes for different classes of yield of rice genotypes.

*Key words*: Classification; Rice; Discriminant analysis; Multilayer perceptron neural network; Accuracy rate; Kappa statistics.

## 1. Introduction

Agriculture sector plays a very crucial role in the economy of developing countries and it is the main source of income, employment and food for their population. With the aim of producing more and better crops, the agricultural sector has gone through many new technologies. According to UN Report 2017, the world population is expected to have an increase of 9.8 billion in 2050 and 11.2 billion in 2100. So, there should be need to increase world food production by 50% to feed the estimated world production.

Rice is an important staple food in Jammu and Kashmir as well as all over the world and its production plays an important role in the life of all farmers. Agriculture and Food security policymakers all over the world should give their attention in promoting the research work and projects for studying the processing, food manufacturing, improvement in nutritive

Corresponding Author: Manish Sharma
Email: manshstat@gmail.com

values and potential health benefits of rice by considering its different varieties to promote their utilization as food in respective places.

Classification is playing a very important role in the field of research in agriculture sector. It is a data mining technique used for prediction of class of objects and is an example of supervised learning as suggested by Kumar *et. al*. 2012. Classification predicts categorical label either discrete or ordered. Classification problems can be done using either statistical methods or machine learning methods or both. For classification through statistical methods, discriminant analysis and through machine learning methods, artificial neural network can be used. The classification of genotypes for different classes of yield, can helps to create genetic variability among the genotypes with respect to a particular character.

The primary goal of the research work is to provide a best approach to classify the rice genotypes for different classes of yield on the basis of different characters.

## 2. Material and method

The primary data collected on yield and attributing characters of rice genotypes such as average plant height ($X_1$), number of effective tillers per plant ($X_2$), number of days for 50 percent maturity ($X_3$), number of days for full maturity ($X_4$) and 1000 grains weight ($X_5$) of 100 rice genotypes from the trail laid in SKUAST Jammu. The yield of rice genotypes is considered as dependent variable which has been classified into three categories as given below

Low : Yield less than 150 grams
Medium : Yield between 150 & 300 grams
High : Yield greater than 300 grams

and all other physical characters are considered as independent variables. The data set is divided randomly into training data consists of 80 percent of data and test data consists remaining 20%. Discriminant Analysis is a multivariate technique introduced by Fisher (1936) to differentiate between groups. The maximum number of discriminant functions that can be computed is equal to minimum of *K*-1 and *t*, where *K* is the number of groups and *t* is the number of variables. Suppose the first discriminant function is

$$D_1 = A_{11}X_1 + A_{12}X_2 + \cdots + A_{1t}X_t$$

where the $A_{1j}$ is the weight of the *j*th variable for the first discriminant function. The weights of the discriminant function are such that the ratio is

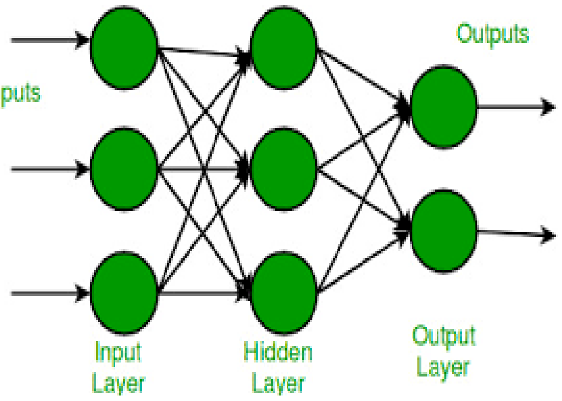
$$\lambda_1 = \frac{Between\ groups\ SS\ of\ D_1}{Within\ groups\ SS\ of\ D_1}.$$

Suppose the second discriminant function is given by $D_2 = A_{21}X_1 + A_{22}X_2 + \cdots + A_{2t}X_t$.

The weights of above discriminant function are estimated such that the ratio is

$$\lambda_2 = \frac{Between\ groups\ SS\ of\ D_2}{Within\ groups\ SS\ of\ D_2}.$$

The above $\lambda_i$'s are maximized subject to the condition that $D_i$ and $D_{i-1}$are uncorrelated. The procedure is repeated until all possible discriminant functions are identified. Once the identification of discriminant functions is done, the next step is to determine a rule for classifying the future observations.



The other technique used is multilayer perceptron (MLP), the most known and most frequently used type of neural network for classification problems. In this neural network there are multiple layers of neurons that are present between input and output. MLPs are also known as feedforward neural networks which means that data flow in one direction from the input to the output layer. Layers that are present in between the input and output layers are referred as hidden layers. The hidden layer performs useful intermediary computations before directing the input to the output layer. The input layer neurons are linked to the hidden layer neurons through some weights known as input-hidden layer weights. Similarly, the hidden layer neurons are linked to the output layer neurons by hidden-output layer weights.

In order to check the classification ability of statistical models and artificial neural network model we use measures like Accuracy rate and Kappa statistics given as

$$\text{Accuracy Rate} = \frac{\text{correctly classified data}}{\text{total data}} \times 100$$

$$\text{Kappa Statistics} = \frac{N\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{ir}x_{ic}}{N^2 - \sum_{i=1}^{k} x_{ir}x_{ic}}$$

where $x_{ii}$ is the count of diagonal elements of the confusion matrix; $x_{ir}$ and $x_{ic}$ are the total of rows and columns of confusion matrix respectively and *N* is the total number of observations.

## 3. Results and discussions

The Discriminant analysis and Multilayer Perceptron Neural Network (MLPNN) used for classification of research data and the results of these methods have been discussed.

**Table 1: Classification table of discriminant analysis for yield of rice**

| Sample | Observed (Number of genotypes) | Predicted (Number of genotypes) | | | |
|---|---|---|---|---|---|
| | | Low | Medium | High | % Correct |
| Training (80 %) | Low (9) | 8 | 0 | 1 | 88.9 |
| | Medium (50) | 10 | 24 | 16 | 48.0 |
| | High (21) | 1 | 8 | 12 | 57.1 |
| | Overall % | | | | 55.0 |
| Testing | Low (5) | 5 | 0 | 0 | 100.0 |

| | | | | | |
|---|---|---|---|---|---|
| (20%) | Medium (13) | 3 | 8 | 2 | 61.5 |
| | High (2) | 0 | 0 | 2 | 100.0 |
| | Overall % | | | | 75.0 |

The classification of rice genotypes using discriminant analysis is given by Table 1 which represents that in training dataset of discriminant analysis, 8 out of the 9 Low yield genotypes with 88.9 percent of accuracy, 24 out of the 50 Medium yield genotypes with 48.0 percent of accuracy, 12 out of 21 High yield genotypes with 57.1 percent of accuracy are correctly classified and overall, 55.0 percent of the training cases are classified correctly. In testing dataset of discriminant analysis, 5 out of 5 low yield genotypes are classified correctly with 100 percent accuracy, 8 out of 13 medium yield genotypes are correctly with 61.5 percent of accuracy, 2 out of 2 high yield genotypes are correctly with 100 percent of accuracy and overall, 75 percent of the testing cases are classified correctly.

**Table 2: Tests of equality of group means**

| Variable | Wilks' Lambda | *F* | DF1 | DF2 | *p*-value |
|---|---|---|---|---|---|
| $X_1$ | 0.865 | 1.326 | 2 | 17 | $0.292^{ns}$ |
| $X_2$ | 0.893 | 1.014 | 2 | 17 | $0.384^{ns}$ |
| $X_3$ | 0.576 | 6.250 | 2 | 17 | 0.009** |
| $X_4$ | 0.557 | 6.755 | 2 | 17 | 0.007** |
| $X_5$ | 0.971 | 0.256 | 2 | 17 | $0.777^{ns}$ |

ns: non-significant **: significant at 1% level of significance

The Table 2 represents that the Wilks' Lambda statistics for variables average plant height, number of effective tillers per plant, number of days for 50 percent maturity, number of days for full maturity and 1000 grains weight which was 0.865, 0.893, 0.576, 0.557 and 0.971 respectively. As per values of Wilks' Lambda, the smaller the value of Wilks' Lambda, the more important the independent variable. Therefore, it indicates that the important independent variable is number of days for full maturity followed by number of days for 50 percent maturity, average plant height, number of effective tillers and 1000 grains weight for yield classes of rice genotypes. Also, it is concluded that the variables such as number of days for 50 percent maturity and number of days for full maturity are highly significant and these regressors are the main contributors for differences in means of three classes for yield of rice.

The architecture of Multilayer Perceptron Neural Network (MLPNN) in Figure 1 depicts that there are 5 input nodes and 4 hidden nodes for yield of rice, the lines with light colour represents weights greater than zero and the dark colour lines display weights less than zero.

Table 3 depicts that in training dataset of MLPNN, 7 out of the 9 Low yield genotypes are correctly classified with 77.8 percent of accuracy, 39 out of 50 Medium yield genotypes are classified correctly with 78.0 percent of accuracy, 3 out of 21 High yield genotypes are correctly classified with 14.3 percent of accuracy and overall, 61.3 percent of the training cases are classified correctly. In testing dataset of MLPNN, 4 out of the 5 Low yield genotypes are correctly classified with 80 percent of accuracy, 13 out of 13 Medium yield genotypes are classified correctly with 100 percent of accuracy, 0 out of 2 High yield genotypes are correctly classified with 0 percent of accuracy and overall, 85 percent of the testing samples are classified correctly.

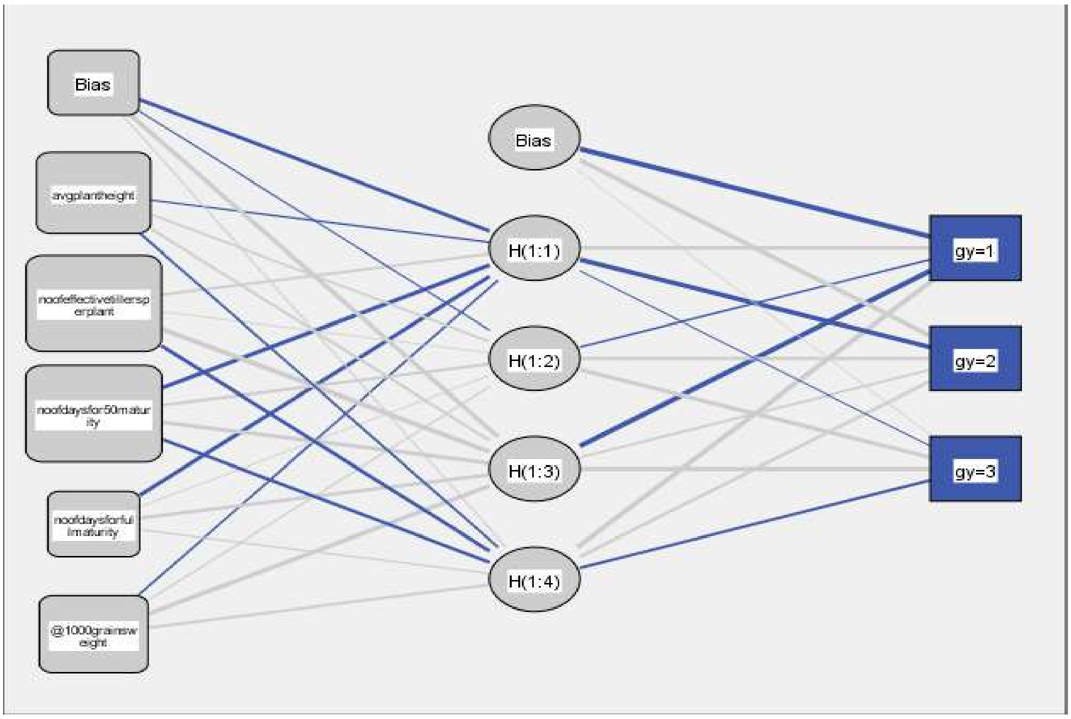


**Figure 1: Architecture of MLPNN for yield of rice genotypes**

**Table 3: Classification table of MLPNN for yield of rice genotypes**

| Sample | Observed (Number of genotypes) | Predicted (Number of genotypes) | | | |
|---|---|---|---|---|---|
| | | Low | Medium | High | % Correct |
| Training (80 %) | Low (9) | 7 | 2 | 0 | 77.8 |
| | Medium (50) | 7 | 39 | 4 | 78.0 |
| | High (21) | 1 | 17 | 3 | 14.3 |
| | Overall % | | | | 61.3 |
| Testing (20 %) | Low (5) | 4 | 1 | 0 | 80.0 |
| | Medium (13) | 0 | 13 | 0 | 100.0 |
| | High (2) | 0 | 2 | 0 | 0.0 |
| | Overall % | | | | 85.0 |

Figure 2 represents the importance of independent variable through MLP neural network for classification of rice genotypes for different classes of yield and depicts that number of days for 50 percent maturity is the most important independent variable for classification (100 percent) followed by number of effective tillers per plant (99.1 percent), average plant height (76 percent), 1000 grains weight (69.8 percent) and number of days for full maturity (51.2 percent).

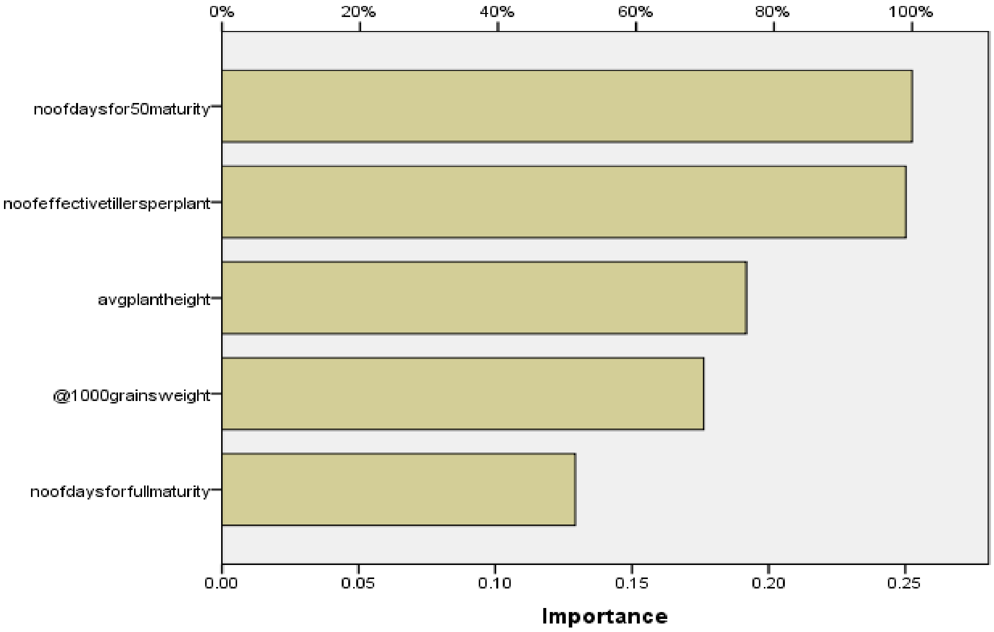


**Figure 2: Normalized independent variable importance**

**Table 4: Classification ability measure**

| Criteria | Measures | Discriminant Analysis | Multilayer Perceptron |
|---|---|---|---|
| Classification Ability | Accuracy Rate | 75 | 85 |
| | Kappa Statistics | 0.59 | 0.65 |

Table 4 represents the value of different classification ability measures and these measures will help to select the best model for classification. The accuracy rate for Discriminant Analysis is 75 percent whereas for MLPNN it is 85 percent. Also, value of kappa statistics for discriminant analysis is 0.59 but 0.65 for MLPNN.

## 4. Conclusion

The MLPNN method performed better as compare to Discriminant Analysis method for classification of rice genotypes for different classes of yield of rice genotypes as it has larger values of classification ability measures. The important attributing character for classification of rice genotypes through MLPNN is number of days for 50 percent maturity followed by number of effective tillers per plant, average plant height, 1000 grains weight and number of days for full maturity whereas for Discriminant Analysis important attributing variable is number of days for full maturity followed by number of days for 50 percent maturity, average plant height, number of effective tillers and 1000 grains weight for classification of rice genotypes for yield.

## References

Fisher, R. A. (1936). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc.publication.

Galdon, B. R., Mendez, E. M., Havel, J., and Diaz, C. (2010). Cluster analysis and artificial neural networks multivariate classification of onion Varieties. *Journal of Agricultural and Food Chemistr*y, **58**, 11435–11440.

Halagundegowda, G. R., Singh, Abhishek, and Meenakshi, H. K. (2017). Discriminant analysis for prediction and classification of farmers based on adoption of drought coping mechanisms. *Agriculture Update*, **12**, 635-640.

Khan, M., and Hooda, B. K. (2021). Potential of artificial neural networks as compared to discriminant analysis in the classification of mustard accessions using grain yield. *International Journal of Statistics and Applied Mathematics*, **6**, 20-23.

Kumar, R., and Verma, R. (2012). Classification algorithms for data mining: A survey. *International Journal of Innovative Engineering Technology (IJIET)*, **1**, 7-14.

Nagraja M. S., and Singh, Abhishek (2018). Statistical models for classification of genotypes for yield of little Millet. *International Journal of Agriculture Sciences*, **10**, 5593-5597.

Nagraja, M. S., and Singh, A. (2018). Use of ordinal logistic regression and multiclass discriminant model for classification of genotypes for maturity of little millet. *International Journal of Pure Applied Biosciences*, **6**, 248-258.

Pazoki, A. R., Farokhi, F., and Pazoki, Z. (2014). Classification of rice grain varieties using two artificial neural networks (mlp and neuro-fuzzy). *The Journal of Animal & Plant Sciences*, **24**, 336-343.

Praveen, S., and Gayatri, V. (2005). Discriminant analysis for rice-wheat system having the same attributing characters towards grain yield. *Indian Journal Agricultural Research*, **39**, 203-207.

Savakar, D. (2012). Identification and classification of bulk fruits images using artificial neural networks. *International Journal of Engineering and Innovative Technology*, **1**, 36-40.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 139–150

# On Zero-inflated Generalized Alternative Hyper-Poisson Distribution and its Properties

**C. Satheesh Kumar and Rakhi Ramachandran**
*Department of Statistics*
*University of Kerala, Trivandrum, Kerala-695581*

Received: 26 July 2023; Revised: 06 August 2023; Accepted: 08 August 2023

## Abstract

A generalized version of the zero-inflated alternative hyper-Poisson distribution of Kumar and Ramachandran (*Statistica*, 2021) is introduced and study some of its important statistical properties such as mean, variance, recursion relations for probabilities, raw moments and factorial moments. The estimation of the parameters of this distribution is considered and the distribution has been fitted to a well-known data set. Further a generalized likelihood ratio test procedure is applied for testing the significance of the inflation parameter.

*Key words:* Confluent hypergeometric series; Count data modeling; Generalized likelihood ratio test; Model selection; Simulation; Zero-inflated Hermite distribution.

**AMS Subject Classifications:** 62K05, 05B05

## 1. Introduction

Bardwell and Crow (1964) considered a generalized version of the Poisson family of distributions through the following probability mass function (p.m.f.), for x= 0, 1, 2, ..., $\lambda > 0$ and $\theta > 0$.

$$f(x) = P(X = x) = \frac{1}{\phi(1;\lambda;\theta)} \frac{\theta^x}{(\lambda)_x}, \tag{1}$$

where

$$\phi(a;b;z) = \sum_{k=0}^{\infty} \frac{(a)_k z^k}{(b)_k k!}$$

is the confluent hypergeometric series, in which

$(a)_k = a(a+1)(a+2)...(a+k-1) = \dfrac{\Gamma(a+k)}{\Gamma(a)}$, for k=1, 2, ... and $(a)_0$=1. The distribution with p.m.f. (1) is known in the literature as the hyper-Poisson distribution (HPD).

Corresponding Author: C. Satheesh Kumar
Email: drcsatheeshkumar@gmail.com

Kumar and Nair (2012) considered an alternative form of the hyper-Poisson distribution (AHPD). The p.m.f.. of the AHPD is the following, for y=0, 1, 2, ... .

$$P(Y=y) = \frac{\gamma^y}{(\rho)_y}\phi(1+y;\rho+y;-\gamma), \tag{2}$$

in which $\gamma > 0$ and $\rho > 0$. The Poisson distribution is the special case of the AHPD when $\rho = 1$. Moreover over dispersion and under dispersion in cases of $\rho > 1$ and $\rho < 1$ is also one of the important characteristics of the AHPD. Kumar and Ramachandran (2021) introduced a zero-inflated version of the alternative hyper-Poisson distribution (ZIAHPD) whose p.m.f.. is given by

$$f(z) = \begin{cases} \omega + (1-\omega)\phi(1;\rho;-\gamma), \ z=0 \\ (1-\omega)\frac{\gamma^z}{(\rho)_z}\phi(1+z;\rho+z;-\gamma), \ z=1, \ 2, \ \dots , \end{cases} \tag{3}$$

in which $\omega \in [0,1]$, $\rho > 0$ and $\gamma > 0$. When $\rho = 1$, the ZIAHPD reduces to the zero-inflated Poisson distribution.

Through this paper we develop further a generalized version of the zero-inflated alternative hyper-Poisson distribution (ZIAHPD) of Kumar and Ramachandran (2021) which we call "the zero-inflated generalized alternative hyper-Poisson distribution (ZIGAHPD)" and discuss some of its important statistical properties. In section 2, we present the definition of the ZIGAHPD and obtain its probability generating function, expressions for its mean and variance, and recursion formulae for probabilities, raw moments and factorial moments. Further, the estimation of the parameters of the model is discussed in section 3 and a test procedure is discussed in section 4. In section 5 both procedures discussed in sections 3 and 4 are illustrated with its relevence with the help of a real life data set.

We need the following series representations in the sequel.

$$\sum_{x=0}^{\infty}\sum_{r=0}^{\infty} A(r,x) = \sum_{x=0}^{\infty}\sum_{r=0}^{x} A(r,x-r) \tag{4}$$

$$\sum_{x=0}^{\infty}\sum_{r=0}^{\infty} A(r,x) = \sum_{x=0}^{\infty}\sum_{r=0}^{[\frac{x}{m}]} A(r,x-rm). \tag{5}$$

## 2. Definition and Properties

We present the definition of the ZIGAHPD and discuss some of its properties.

**Definition 1:** A discrete random variable $M$ is said to follow "the zero-inflated generalized alternative hyper-Poisson distribution or in short ZIGAHPD" with parameters $\omega$, $\lambda$, $\theta_1$, $\theta_2$ and $\theta_3$ if its p.m.f. is

$$\begin{aligned} f(m) &= P(M=m) \qquad (6) \\ &= \begin{cases} \omega + (1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)], & m=0 \\ (1-\omega)\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} \frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)]\frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^k}{k!}\frac{\theta_3^j}{j!}, & m=1,2,\dots \\ 0, otherwise \end{cases} \end{aligned}$$

in which $\omega \in [0,1)$, $\lambda > 0$, $\theta_1 > 0$, $\theta_2 \geq 0$ and $\theta_3 \geq 0$.

Important special cases of the ZIGAHPD includes the following cases.

1. when $\omega = 0$, ZIGAHPD → generalized alternative hyper-Poisson distribution (GAHPD) of Kumar and Sandeep (2022).

2. when $\theta_2 = \theta_3 = 0$, ZIGAHPD → the ZIAHPD of Kumar and Ramachandran (2021) with p.m.f.. (3).

3. when $\theta_2 = \theta_3 = 0$ and $\lambda = 1$, ZIGAHPD → ZIPD of Lambert (1992).

4. when $\theta_3 = 0$ and $\lambda = 1$, ZIGAHPD → zero-inflated Hermite distribution (ZIHD) of Kumar and Ramachandran (2020).

5. when $\omega = 0$, $\theta_3 = 0$ and $\lambda = 1$, ZIGAHPD → Hermite distribution (HD) of Kemp and Kemp (1965).

6. when $\omega = 0$ and $\theta_3 = 0$, ZIGAHPD → modified alternative hyper-Poisson distribution (MAHPD) of Kumar and Nair (2013).

7. when $\omega = 0$, $\theta_2 = 0$ and $\theta_3 = 0$, ZIGAHPD → alternative hyper-Poisson distribution (AHPD) of Kumar and Nair (2012).

Now we obtain the following results.

**Result 1:** The probability generating function (p.g.f) G(t) of the ZIGAHPD with p.m.f. (6) is the following.

$$G(t) = \omega + (1-\omega)\ \phi[1;\lambda;\theta_1(t-1) + \theta_2(t^2-1) + \theta_3(t^3-1)]. \tag{7}$$

**Proof:** By definition, the p.g.f of the ZIGAHPD having p.m.f. (6) is given by

$$\begin{aligned} G(t) &= \sum_{m=0}^{\infty} f(m)t^m \\ &= \omega + (1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)] + (1-\omega)\sum_{m=1}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} t^m \frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}} \\ &\times \ \phi[1+m-2j-k;\lambda+m-2j-k+z;-(\theta_1+\theta_2+\theta_3)]\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(m-3j-2k)!k!j!} \\ &= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]} t^m \frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}} \frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(m-3j-2k)!k!j!} \\ &\times \ \phi[1+m-2j-k;\lambda+m-2j-k+z;-(\theta_1+\theta_2+\theta_3)]. \end{aligned}$$

In the light of the following result $(\lambda)_x(\lambda+x)_r = (\lambda)_{x+r}$, we have

$$
\begin{aligned}
G(t) &= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(m-2j-k)!(m-3j-k)!}{(m-3j-k)!j!k!(m-3j-2k)}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}} \qquad (8)\\
&\times \ \phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)]t^m\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}}\\
&\times \ \phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)]t^m\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}}t^m\\
&\times \ \sum_{r=0}^{\infty}\frac{(1+m-2j-k)_r}{(\lambda+m-2j-k)_r}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{r!}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}}t^m\\
&\times \ \frac{(1)_{m-2j-k}}{(m-2j-k)!(\lambda)_{m-2j-k}}\sum_{r=0}^{\infty}\frac{(1+m-2j-k)_r}{(\lambda+m-2j-k)_r}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{r!}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}(m-2j-k)!}t^m\\
&\times \ \sum_{r=0}^{\infty}(1)_{m-2j-k}\frac{(1+m-2j-k)_r}{(\lambda+m-2j-k)_r}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{r!}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\theta_1^{m-3j-2k}\theta_2^k\theta_3^j t^m\\
&\times \ \sum_{r=0}^{\infty}(1)_{m-2j-k+r}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{(\lambda)_{m-2j-k+r}r!(m-2j-k)!}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\theta_1^{m-3j-2k}\theta_2^k\theta_3^j t^m\\
&\times \ \sum_{r=0}^{\infty}\binom{m-2j-k+r}{r}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{(\lambda)_{m-2j-k+r}}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{r=0}^{\infty}\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\binom{m-2j-k+r}{r}\binom{m-2j-k}{j}\binom{m-3j-k}{k}\\
&\times \ \frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k+r}}[-(\theta_1+\theta_2+\theta_3)]^r t^m.
\end{aligned}
$$

Using inequality (4), we obtain

$$\begin{aligned}
G(t) &= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{r=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{[\frac{m}{2}]}\binom{m+2j+r-k}{r}\binom{m+j-k}{j}\binom{m-k}{k} \qquad (9)\\
&\times \frac{(\theta_1 t)^{m-2k}(\theta_2 t^2)^k(\theta_3 t^3)^j}{(\lambda)_{m+r+j-k}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{r=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\binom{m+j+r-k}{r}\binom{m+j+k}{j}\binom{m+k}{k}\\
&\times \frac{(\theta_1 t)^{m}(\theta_2 t^2)^k(\theta_3 t^3)^j}{(\lambda)_{m+r+j+k}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{r=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{m}\binom{m+j+r}{r}\binom{m+j}{j}\binom{m}{k}\\
&\times \frac{(\theta_1 t)^{m-k}(\theta_2 t^2)^k(\theta_3 t^3)^j}{(\lambda)_{m+r+j}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\binom{m+j+r}{r}\binom{m+j}{j}\frac{(\theta_1 t+\theta_2 t^2)^m(\theta_3 t^3)^j}{(\lambda)_{m+r+j}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{j=0}^{m}\binom{m+r}{r}\binom{m}{j}\frac{(\theta_1 t+\theta_2 t^2)^{m-j}(\theta_3 t^3)^j}{(\lambda)_{m+r+j}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\binom{m+r}{r}\frac{(\theta_1 t+\theta_2 t^2+\theta_3 t^3)^m}{(\lambda)_{m+r}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\binom{m}{r}\frac{(\theta_1 t+\theta_2 t^2+\theta_3 t^3)^{m-r}}{(\lambda)_{m}}[-(\theta_1+\theta_2+\theta_3)]^r\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\frac{[\theta_1 t+\theta_2 t^2+\theta_3 t^3-(\theta_1+\theta_2+\theta_3)]^m}{(\lambda)_m}\\
&= \omega + (1-\omega)\sum_{m=0}^{\infty}\frac{[\theta_1(t-1)+\theta_2(t^2-1)+\theta_3(t^3-1)]^m}{(\lambda)_m}
\end{aligned}$$

which on simplification gives (7). □

**Result 2:** The mean and variance of the ZIGAHPD with p.g.f (7) are

$$Mean = \frac{(1-\omega)}{\lambda}\,(\theta_1+2\theta_2+3\theta_3)$$

and

$$Variance \;=\; \left\{\left(\frac{2}{\lambda+1}-\frac{1-\omega}{\lambda}\right)(\theta_1+2\theta_2+3\theta_3)^2+(\theta_1+4\theta_2+9\theta_3)\right\}\frac{(1-\omega)}{\lambda}.$$

Next we derive certain recursion formulae for the probabilities, raw moments and factorial moments of the ZIGAHPD through the following results. Hereafter, for the convenience of the notation, we write $f_m(\lambda^{(j)})$ for the probability mass function $f(m)$ as given in (6) where $\lambda^{(j)} = (1+j, \lambda+j)$ for $j = 0, 1, 2, \ldots$ .

**Result 3:** A simple recursion formula for probabilities $f_m(\lambda^{(j)})$ of the ZIGAHPD is the following

$$f_1(\lambda^{(0)}) = \frac{\theta_1}{\lambda}\left(f_0(\lambda^{(1)}) - \omega\right) \tag{10}$$

and

$$(m+1)f_{m+1}(\lambda^{(0)}) = \frac{1}{\lambda}\left[\theta_1 f_m(\lambda^{(1)}) + 2\theta_2 f_{m-1}(\lambda^{(1)}) + 3\theta_3 f_{m-2}(\lambda^{(1)})\right],\ for\ m \geq 1. \tag{11}$$

**Proof:** The p.g.f of the ZIGAHPD given in (7) can be written as

$$\begin{aligned} G(t) &= \omega + (1-\omega)\ \phi[1;\lambda;\theta_1(t-1) + \theta_2(t^2-1) + \theta_3(t^3-1)] \\ &= \sum_{m=0}^{\infty} t^m f_m(\lambda^{(j)}). \end{aligned} \tag{12}$$

On differentiating (12) with respect to t, we obtain the following.

$$\sum_{m=0}^{\infty}(m+1)f_{m+1}(\lambda^{(0)})t^m = \frac{(1-\omega)}{\lambda}(\theta_1+2\theta_2 t+3\theta_3 t^2)\ \phi[2;\lambda+1;\theta_1(t-1)+\theta_2(t^2-1)+\theta_3(t^3-1)]. \tag{13}$$

Also from (12), we have

$$(1-\omega)\phi[2;\lambda+1;\theta_1(t-1)+\theta_2(t^2-1)+\theta_3(t^3-1)] = \sum_{m=0}^{\infty} f_m(\lambda^{(1)})t^m - \omega. \tag{14}$$

Combining relations (13) and (14) we obtain

$$\sum_{m=0}^{\infty}(m+1)f_{m+1}(\lambda^{(0)})t^m = \frac{(\theta_1 + 2\theta_2 t + 3\theta_3 t^2)}{\lambda}\left[\sum_{m=0}^{\infty} f_m(\lambda^{(1)})t^m - \omega\right]. \tag{15}$$

Now, on equating the coefficients of $t^0$ on both sides of (15), we get (10), and on equating the coefficients of $t^y$ on both sides of (15), we get (11). □

**Result 4:** For $r \geq 0$, a recursion formula for raw moments $\mu_r(\lambda^{(0)})$ of the ZIGAHPD is

$$\mu_{[r+1]}(\lambda^{(0)}) = \frac{1}{\lambda}\sum_{k=0}^{r}\left(\begin{array}{c} r \\ k \end{array}\right)(\theta_1 + 2^{k+1}\theta_2 + 3^{k+1}\theta_3)\mu_{[r-k]}(\lambda^{(1)}). \tag{16}$$

**Proof:** For any $t \in \Re = (-\infty, \infty)$ and $i = \sqrt{-1}$, the characteristic function of the ZIGAHPD is

$$\begin{aligned} H(t) &= G(e^{it}) \\ &= \omega + (1-\omega)\ \phi[1;\lambda;\theta_1(e^{it}-1) + \theta_2(e^{2it}-1) + \theta_3(e^{3it}-1)] \\ &= \sum_{r=0}^{\infty}\mu_r(\lambda^{(0)})\frac{(it)^r}{r!}. \end{aligned} \tag{17}$$

Differentiating (17) with respect to $t$, we get

$$\sum_{r=0}^{\infty}\mu_{[r+1]}(\lambda^{(0)})\frac{(it)^r}{r!} = \frac{(1-\omega)}{\lambda}\left(\theta_1 e^{it} + 2\theta_2 e^{2it} + 3\theta_3 e^{3it}\right)\times \\ \phi[2;\lambda+1;\theta_1(e^{it}-1)+\theta_2(e^{2it}-1)+\theta_3(e^{3it}-1)]. \tag{18}$$

Also, from (17) we have

$$(1-\omega)\phi[2;\lambda+1;\theta_1(e^{it}-1)+\theta_2(e^{2it}-1)+\theta_3(e^{3it}-1)] = \sum_{r=0}^{\infty}\mu_{[r]}(\lambda^{(1)})\frac{(it)^r}{r!} - \omega. \tag{19}$$

Combining (18) and (19), we get

$$\sum_{r=0}^{\infty}\mu_{[r+1]}(\lambda^{(0)})\frac{(it)^r}{r!} = \frac{1}{\lambda}\left(\theta_1 e^{it} + 2\theta_2 e^{2it} + 3\theta_3 e^{3it}\right)\left(\sum_{r=0}^{\infty}\mu_{[r]}(\lambda^{(1)})\frac{(it)^r}{r!} - \omega\right). \tag{20}$$

On expanding the exponential functions in the right hand side expression of (20) and equating the coefficients of $\frac{(it)^r}{r!}$ on both sides we get (16). □

**Result 5:** For $r \geq 1$ a simple recursion formula for factorial moments $\mu_{[r]}(\lambda^{(0)})$ of the ZIGAHPD is

$$\mu_{[r+1]}(\lambda^{(0)}) = \frac{\theta_1}{\lambda}\mu_{[r]}(\lambda^{(1)}) + \frac{2\theta_2 r}{\lambda}\mu_{[r-1]}(\lambda^{(1)}) + \frac{3\theta_3 r(r-1)}{\lambda}\mu_{[r-2]}(\lambda^{(1)}). \tag{21}$$

**Proof:** The factorial moment generating function F(t) of the ZIGAHPD with p.g.f. (7) is the following

$$\begin{aligned} F(t) &= G(1+t) \\ &= \omega + (1-\omega)\phi[1;\lambda;\theta_1 t + \theta_2\{(1+t)^2-1\} + \theta_3\{(1+t)^3-1\}] \\ &= \sum_{r=0}^{\infty}\mu_{[r]}(\lambda^{(0)})\frac{t^r}{r!}. \end{aligned} \tag{22}$$

On differentiating (22) with respect to t, we get

$$\frac{1}{\lambda}(1-\omega)(\theta_1 t + \theta_2\{(1+t)^2-1\}) + \theta_3\{(1+t)^3-1)\}\times \\ \phi[2;\lambda+1;\theta_1 t + \theta_2\{(1+t)^2-1\} + \theta_3\{(1+t)^3-1\}] = \sum_{r=0}^{\infty}\mu_{[r+1]}(\lambda^{(0)})\frac{t^r}{r!}. \tag{23}$$

From (22), we obtain

$$(1-\omega)\phi[2;\lambda+1;\theta_1 t + \theta_2\{(1+t)^2-1\} + \theta_3\{(1+t)^3-1\}] = \sum_{r=0}^{\infty}\mu_{[r]}(\lambda^{(1)})\frac{t^r}{r!} - \omega. \tag{24}$$

Equations (23) and (24) together lead to

$$\sum_{r=0}^{\infty}\mu_{[r+1]}(\lambda^{(0)})\frac{t^r}{r!}=\frac{\theta_1+2\theta_2(1+t)+3\theta_3(1+t)^2}{\lambda}\left(\sum_{r=0}^{\infty}\mu_{[r]}(\lambda^{(1)})\frac{t^r}{r!}-\omega\right). \tag{25}$$

On equating the coefficients of $\frac{t^r}{r!}$, we get (21). □

## 3. Maximum likelihood estimation

Here we consider the estimation of the parameters $\omega$, $\lambda$, $\theta_1$, $\theta_2$ and $\theta_3$ of the ZIGAHPD by the method of maximum likelihood. For any $m = 0, 1, 2, ...$, let $A(m)$ be the observed frequency of $m$ events and let $z$ be the highest value of $m$ observed. Then the likelihood function of the sample is given by

$$L(\Theta; m) = \prod_{m=0}^{z}[f(m)]^{A(m)},$$

where $f(m)$ is the p.m.f. of the ZIGAHPD given in (6).
Now $L(\Theta; m)$ can be written as

$$L(\Theta; m) = (f(0))^s\prod_{m=1}^{z}(f(m))^{A(m)},$$

where $s = A(0)$.

Then the log-likelihood function can be written as

$$\begin{aligned}\ln L(\theta; m) &= s\ln\left[\omega+(1-\omega)\phi(1;\lambda;\theta_1+\theta_2+\theta_3)\right]+\sum_{m=1}^{z}A(m)\\ &\times \ln\left[(1-\omega)\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)]\right.\\ &\times \left.\frac{\theta_1^{m-3j-2k}\ \theta_2^k\ \theta_3^j}{(m-3j-2k)!\ k!\ j!}\right]\end{aligned} \tag{26}$$

Assume that $\hat{\omega}$, $\hat{\lambda}$, $\hat{\theta_1}$, $\hat{\theta_2}$ and $\hat{\theta_3}$ be the maximum likelihood estimators of the parameters $\omega$, $\lambda$, $\theta_1$, $\theta_2$ and $\hat{\theta_3}$ of the ZIGAHPD. Now, on differentiating the log-likelihood function (26) with respect to $\omega$, $\lambda$, $\theta_1$, $\theta_2$ and $\theta_3$ and equating to zero, we obtain the following likelihood equations:

$$\frac{\partial\ln L}{\partial\omega}=0$$

which implies

$$\frac{s\left[1-\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]\right]}{\omega+(1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]}-\sum_{m=1}^{z}\frac{A(m)}{(1-\omega)}=0, \tag{27}$$

$$\frac{\partial \ln L}{\partial \lambda} = 0$$

which implies

$$\begin{aligned}
&\frac{s(1-\omega)}{\omega+(1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]}\sum_{r=0}^{\infty}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{(\lambda)_r}[\psi(\lambda)-\psi(\lambda+r)] \\
&+\sum_{m=1}^{z}A(m)\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{[(\lambda)_{m-2j-k}]^2}\frac{\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(m-3j-2k)!k!j!} \\
&\times\left\{\frac{1}{(\lambda)_{m-2j-k}}\sum_{r=0}^{\infty}\frac{[-(\theta_1+\theta_2+\theta_3)]^r}{r!(\lambda+m-2j-k)_r}(1+m-2j-k)_r\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]\right. \\
&\times[\psi(\lambda+m-2j-k)-\psi(\lambda+m-2j-k+r)]-\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)] \\
&\left.\times\,\frac{1}{(\lambda)_{m-2j-k}}[\psi(\lambda+m-2j-k)-\psi(\lambda+m-2j-k+r)]\right\}=0,
\end{aligned}\tag{28}$$

$$\frac{\partial \ln L}{\partial \theta_1} = 0$$

which implies

$$\begin{aligned}
&\frac{-s^{(1-\omega)}/_{\lambda}\phi[2;\lambda+1;-(\theta_1+\theta_2+\theta_3)]}{\omega+(1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]}+\sum_{m=0}^{z}A(m)\frac{1}{\xi}\left\{\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\right. \\
\times\quad&\frac{\theta_1^{m-3j-2k-1}}{(m-3j-2k-1)!}\frac{\theta_2^k\theta_3^j}{k!j!}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)] \\
&-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^k\theta_3^j}{k!j!}\frac{1}{\lambda+m-2j-k} \\
&\left.\times\,\phi[2+m-2j-k;\lambda+1+m-2j-k;-(\theta_1+\theta_2+\theta_3)]\right\}=0,
\end{aligned}\tag{29}$$

$$\frac{\partial \ln L}{\partial \theta_2} = 0$$

which implies

$$\begin{aligned}
&\frac{-s^{(1-\omega)}/_{\lambda}\phi[2;\lambda+1;-(\theta_1+\theta_2+\theta_3)]}{\omega+(1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]}+\sum_{m=0}^{z}A(m)\frac{1}{\xi}\left\{\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\right. \\
\times\quad&\frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^{k-1}\theta_3^j}{(k-1)!j!}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)] \\
&-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^k\theta_3^j}{k!j!}\frac{1}{\lambda+m-2j-k} \\
&\left.\times\,\phi[2+m-2j-k;\lambda+1+m-2j-k;-(\theta_1+\theta_2+\theta_3)]\right\}=0
\end{aligned}\tag{30}$$

and

$$\frac{\partial \ln L}{\partial \theta_3} = 0$$

which implies

$$\frac{-s{}^{(1-\omega)}\!/_{\lambda}\phi[2;\lambda+1;-(\theta_1+\theta_2+\theta_3)]}{\omega+(1-\omega)\phi[1;\lambda;-(\theta_1+\theta_2+\theta_3)]} + \sum_{m=0}^{z} A(m)\frac{1}{\xi}\left\{\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\right. \tag{31}$$
$$\times \frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^k\theta_3^{j-1}}{k!(j-1)!}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)]$$
$$-\sum_{j=0}^{[\frac{m}{3}]}\sum_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}}{(\lambda)_{m-2j-k}}\frac{\theta_1^{m-3j-2k}}{(m-3j-2k)!}\frac{\theta_2^k\theta_3^j}{k!j!}\frac{1}{\lambda+m-2j-k}$$
$$\left.\times\,\phi[2+m-2j-k;\lambda+1+m-2j-k;-(\theta_1+\theta_2+\theta_3)]\right\},$$

in which $\psi(\lambda) = \frac{\partial}{\partial\lambda}log\;\Gamma(\lambda)$ and
$\xi = \sum\limits_{j=0}^{[\frac{m}{3}]}\sum\limits_{k=0}^{[\frac{m}{2}]}\frac{(1)_{m-2j-k}\theta_1^{m-3j-2k}\theta_2^k\theta_3^j}{(\lambda)_{m-2j-k}(m-3j-2k)!j!k!}\phi[1+m-2j-k;\lambda+m-2j-k;-(\theta_1+\theta_2+\theta_3)].$

On solving the likelihood equations (27), (28), (29), (30) and (31) with the help of some mathematical softwares, say *Mathematica*, one can obtain the maximum likelihood estimators of the parameters of the proposed distribution.

## 4. Testing

In order to test the significance of the inflation parameter $\omega$ of the ZIGAHPD, we adopt the following generalized likelihood ratio test (GLRT) procedure. Here the null hypothesis is

$$H_0 : \omega = 0 \textit{ against the alternative hypothesis } H_1 : \omega \neq 0.$$

The test statistic suggested in the case of GLRT is given by

$$-2\ln\psi = 2\left(\iota_1 - \iota_2\right), \tag{32}$$

where, $\iota_1 = \ln L(\hat{\theta}; m)$, where $\hat{\theta}$ is the maximum likelihood estimator for $\theta = (\omega, \lambda, \theta_1, \theta_2, \theta_3)$ with no restrictions, and $\iota_2 = lnL(\hat{\theta}^*; m)$, in which $\hat{\theta}^*$ is the maximum likelihood estimator for $\theta$ under the null hypothesis $H_0$. The test statistic defined in (32) is asymptotically distributed as $\chi^2$ with one degree of freedom.

## 5. Applications

In this section we illustrate all the procedures discussed in sections 3 and 4 with the help of a real life data set.

The data here considered is a biological data based on the distribution of European Corn borer Pyrausta Naubilalis in field corn (Avi *et al.* (2008)). We have fitted the ZIGAHPD to the data set and considered the fitting of the models - ZIAHPD, ZIHD, ZIPD, ZIMAHPD and GAHPD for comparison. For comparing the models we computed the values of $\chi^2$, AIC, BIC and AICc. The numerical results obtained are presented in Tables 1. Based on the computed values of $\chi^2$, AIC, BIC and AICc as presented in Table 1, one can observe that the ZIGAHPD gives a better fit to the data set while all other models such as ZIAHPD, ZIHD, ZIPD, ZIMAHPD and GAHPD are not appropriate.

We have also calulated the values of the test statistic. The value of the test statistic for $\ln L(\hat{\theta}^*; m) = -169.1$ and $\ln L(\hat{\theta}; m) = -144.3$ is given by 49.6. The critical value of the test having 5% level of significance and degree of freedom one is 3.84, so that the null hypothesis is rejected in all the cases. Thus, we conclude that the additional parameter $\omega$ in the model is significant.

**Table 1: Distribution of the spread of European Corn borer Pyrausta Naubilalis in field corn (Rodriguez et.al., 2008) and the expected frequencies computed using ZIAHPD, ZIHD, ZIPD, ZIMAHPD, GAHPD and ZIGAHPD.**

| *Count* | *Observed frequency* | *ZIAHPD* | *ZIHD* | *ZIPD* | *ZIMAHPD* | *GAHPD* | *ZIGAHPD* |
|---|---|---|---|---|---|---|---|
| 0 | 206 | 265.73 | 200.4 | 252.15 | 213.65 | 245.4 | 200.3 |
| 1 | 143 | 148.4 | 100.515 | 151.43 | 112.51 | 157.6 | 142.6 |
| 2 | 128 | 144.2 | 137.73 | 140.6 | 179.2 | 151.4 | 119.8 |
| 3 | 107 | 127.1 | 110.6 | 118.35 | 119.6 | 128.048 | 100.8 |
| 4 | 71 | 80.361 | 90.7 | 75.78 | 93.2 | 73.42 | 80.4 |
| 5 | 36 | 7.29 | 59.8 | 23.9 | 35.3 | 8.5 | 38.5 |
| 6 | 32 | 4.6 | 38.6 | 7.1 | 19.08 | 7.4 | 39.4 |
| 7 | 17 | 2.3 | 21.3 | 5.4 | 5.9 | 3.6 | 12.6 |
| 8 | 14 | 1.25 | 11.5 | 4.02 | 2.58 | 2.45 | 19.2 |
| 9 | 7 | 0.5 | 5.7 | 1.25 | 0.67 | 1.98 | 5.2 |
| 10 | 7 | 0.25 | 2.7 | 1.6 | 0.23 | 0.87 | 7.9 |
| 11 | 2 | 0.0024 | 1.52 | 0.0015 | 0.05 | 0.50 | 1.2 |
| 12 | 3 | 0.01 | 0.61 | 0.35 | 0.018 | 0.23 | 1.8 |
| 13 | 3 | 0.006 | 0.2 | 0.021 | 0.004 | 0.20 | 2.3 |
| 14 | 1 | 0.00003 | 0.08 | 0.006 | 0.00815 | 0.35 | 2.2 |
| 15 | 1 | 0.00009 | 0.03 | 0.0002 | 0.00025 | 0.05 | 1.1 |
| 16 | 1 | 0.000006 | 0.011 | 0.00004 | 0.00007 | 0.002 | 2.7 |
| 17 | 2 | 0.000007 | 0.003 | 0.041 | 0.000013 | 0.00035 | 2.5 |
| 18 | 1 | 0.000009 | 0.0013 | 0.000008 | 0.0000021 | 0.000007 | 1.5 |
| *Total* | 782 | 782 | 782 | 782 | 782 | 782 | 782 |
| *df* | | 3 | 7 | 6 | 3 | 3 | 6 |
| *Estimates* | | $\lambda$=12.09<br>$\omega$=0.59<br>$\theta$=7.0009 | $\lambda$=1.11<br>$\omega$=0.14<br>$\theta$=0.89 | $\lambda$=3.95<br>$\omega$=0.17 | $\lambda$=0.1<br>$\omega$=0.8<br>$\theta_1$=0.32<br>$\theta_2$=0.36 | $\lambda$=0.15<br>$\omega$=0.29<br>$\theta$=0.60 | $\lambda$=0.63<br>$\omega$=0.26<br>$\theta_1$=0.22<br>$\theta_2$=0.55<br>$\theta_3$=0.025 |
| $\chi^2$*-value* | | 880.21 | 880.21 | 188.64 | 297.9 | 418.01 | 7.48 |
| *P-value* | | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.2787 |
| *AIC* | | 1733.5 | 3344.1 | 910.6 | 3766.18 | 1220.5 | 840.25 |
| *BIC* | | 1734.3 | 3346.9 | 911.8 | 3769.9 | 1224.3 | 841.25.5 |
| *AICc* | | 1739.6 | 3345.8 | 915.3 | 3769.4 | 1224.6 | 845.7 |

## References

Avi, J. R., Jimenez, M. O., Sanchez, A. C., and Castillo, A. S. (2008). The 3 f 2 with complex parameters as generating function of discrete distribution. *Communications in Statistics—Theory and Methods*, **37**, 3009–3022.

Bardwell, G. E. and Crow, E. L. (1964). A two-parameter family of hyper-poisson distributions. *Journal of the American Statistical Association*, **59**, 133–141.

Kemp, C. and Kemp, A. W. (1965). Some properties of the 'hermite'distribution. *Biometrika*, **52**, 381–394.

Kumar, C. S. and Nair, B. U. (2012). An alternative hyper-poisson distribution. *Statistica*, **72**, 357–369.

Kumar, C. S. and Nair, B. U. (2013). Modified alternative hyper-poisson distribution. *Collection of Recent Statistical Methods and Applications*, **2**, 97–109.

Kumar, C. S. and Sandeep, S. (2022). Generalized alternative hyper-poisson distribution. *In Proceedings International Conference on Advances in Mathematicaland Statistical Sciences*, **1**, 127–136.

Kumar, S. and Ramachandran, R. (2020). On some aspects of a zero-inflated overdispersed model and its applications. *Journal of Applied Statistics*, **47**, 506–523.

Kumar, S. and Ramachandran, R. (2021). On zero-inflated alternative hyper-poisson distribution. *Statistica*, **81**, 423–446.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp151-168

# First Collision Time of Three Independent Random Walks

**Arighna Dey[1], Kumarjit Saha[2], and Anish Sarkar[3]**
[1] *Indian Statistical Institute, Kolkata, West Bengal, India*
[2] *Department of Mathematics, Ashoka University, Sonepat, Haryana, India*
[3] *Indian Statistical Institute, Delhi, India*

Received: 29 July 2023; Revised: 07 August 2023; Accepted: 10 August 2023

**Abstract**

Random walks are mathematical objects for modelling random trajectories where the future of the trajectory does not depend on the past. We take three simple random walk where the increments are distributed as $+1, -1$ valued random variables with probabilities $p$ and $1 - p$. We study the expected first collision time of three such random walks. This work is an extension of the work of Coupier *et. al.* (2020) where they studied the case of $p = 1/2$.

*Key words:* Random walks; First collision time; Martingale.

**AMS Subject Classifications:** 62D05

## 1. Introduction

A random walk, denoted by RW, represents a trajectory or collection of trajectories that consists of taking successive random steps, each of which are independent and identically distributed. The most studied example of random walk is the walk on the integers $\mathbb{Z}$, which starts at an integer point and at each step moves by $+1$ or $-1$. This is known as the simple random walk (SRW). When the probabilities of moving to $+1$ and to $-1$ are identical, we call it the simple symmetric random walk (SSRW).

Random walks originate in almost all sciences quite naturally and find applications in various branches of mathematics, computer science, biology, chemistry, physics. In Physics, random walks are used to model the movement of particles in a random environment. The limiting process of the random walk yields the Brownian motion which is central to almost many predictive models. This has connected various branches of Mathematics and physics through the application of random walk.

In biological science, the genetic drift is modelled using random walks, which provide a general idea of the statistical processes involved. In physics, we can random walks to describe an ideal chains of polymers. The concepts of random work has been very crucially used in several fields such as psychology, finance, ecology. In Economics Stock market modelling and pricing are done through the Brownian motion. It is possible to describe fluctuations in the stock market with the random walk concepts. This has resulted several Nobel prizes

Corresponding Author: Anish Sarkar
Email: anish.sarkar@gmail.com

in Economics. Random walks also find application in the Google search engine algorithms, namely the page rank algorithm.

A simple way to construct the random walk is to flip a coin, and if the toss results in a HEAD, move to right by single step, whereas if the toss results in a TAIL, move to left by a single step. To define this walk formally, we take a sequence of independent random variables independent and identically distributed random variables, called the increment sequence, $\{I_k : k \in \mathbb{N}\}$ and an initial state $x \in \mathbb{Z}$. The random walk, starting from $x$, is defined as follows:

$$S_0 = x \text{ and } S_n = x + \sum_{k=1}^{n} I_k \text{ for } n \in \mathbb{N}.$$

This sequence $\{S_n : n \geq 0\}$ is called the random walk on $\mathbb{Z}$.

In this article we deal with three independent simple random walks. Therefore, we will consider three starting points. We note that if the starting positions of two random walks are of different parity, they will never be at the same position at any time point. Thus, we need to consider all starting positions of same parity. Since the intersection times and collision times will not change when we translate all the processes by same amount, we may choose the starting positions so that one random walk starts below the origin (the left random walk), one at the origin (the middle random walk) and the other above the origin (the right random walk). More precisely, we choose $a$ and $b$ positive even numbers and start the random walks at $-a, 0$ and $b$ respectively. We also consider three independent sequences of independent and identically distributed increment random variables $\left\{I_k^{(L)} : k \geq 1\right\}$, $\left\{I_k^{(M)} : k \geq 1\right\}$ and $\left\{I_k^{(R)} : k \geq 1\right\}$ with

$$\mathbb{P}\Big(I_k^{(s)} = +1\Big) = p = 1 - \mathbb{P}\Big(I_k^{(s)} = -1\Big) \tag{1}$$

where $p \in (0,1)$ and $s \in \{L, M, R\}$. Now, we consider the random walks represented by

$$S_n^{(L)} = -a + \sum_{k=1}^{n} I_k^{(L)}, \qquad S_n^{(M)} = \sum_{k=1}^{n} I_k^{(M)} \qquad \text{and} \qquad S_n^{(R)} = b + \sum_{k=1}^{n} I_k^{(R)}.$$

By construction, these three random walks $S_n^{(L)}$, $S_n^{(M)}$ and $S_n^{(R)}$, starting from $-a$, 0 and $+b$ respectively, are independent. We define the first collision time of these three random walks by

$$\tau_c = \inf\Big\{n \geq 1 : (S_n^{(M)} - S_n^{(L)})(S_n^{(R)} - S_n^{(M)})(S_n^{(L)} - S_n^{(R)}) = 0\Big\}. \tag{2}$$

In this article we compute the expectation of $\tau_c$. Coupier et. al. (2020) studied the behavior of $\tau_c$ in the case of simple symmetric random walks, i.e., the increment random variables are distributed as random variables taking values $+1$ with probability $\frac{1}{2}$ and $-1$ with probability $\frac{1}{2}$. We extend the result of Coupier et. al. (2020) for any value of $p \in (0,1)$.

## 2. Collision of two random walks

In Spitzer (1964) it is shown that the first hitting time of a random walk to a state where increment random variables are independent and identically distributed having mean 0 and finite variance is finite almost surely.

We observe that the expectation of the increment random variables and the expectation of the square of the increment random variables are given by : for $s \in \{L, M, R\}$,

$$\mathbb{E}\Big(I_k^{(s)}\Big) = p - (1-p) = 2p - 1 \text{ and}$$
$$\mathbb{E}\left(\Big(I_k^{(s)}\Big)^2\right) = p + 1 - p = 1.$$

Therefore, we have

$$\mathbb{V}\text{ar}\Big(I_k^{(s)}\Big) = \mathbb{E}\left(\Big(I_k^{(s)}\Big)^2\right) - \Big(\mathbb{E}\Big(I_k^{(s)}\Big)\Big)^2 = 4p(1-p).$$

In our particular case, we consider first the collision times of the left random walk and the middle random walk, i.e., set

$$\tau_{L,M} = \inf\Big\{n \geq 1 : S_n^{(L)} = S_n^{(M)}\Big\} = \inf\Big\{n \geq 1 : S_n^{(L)} - S_n^{(M)} = 0\Big\}. \tag{3}$$

Similarly, we may define the first collision time of the middle random walk and the right random walk by

$$\tau_{M,R} = \inf\Big\{n \geq 1 : S_n^{(M)} = S_n^{(R)}\Big\} = \inf\Big\{n \geq 1 : S_n^{(M)} - S_n^{(R)} = 0\Big\}. \tag{4}$$

We consider the collision time of the left and the middle random walk. We set the difference of the two walks by

$$X_n = S_n^{(M)} - S_n^{(L)} \tag{5}$$

for all $n \geq 0$. Similarly set

$$Y_n = S_n^{(R)} - S_n^{(M)} \tag{6}$$

for all $n \geq 0$. Hence, we observe that $X_0 = a$ and $Y_0 = b$.

We may now rephrase the first collision time of two random walks as follows:

$$\tau_{L,M} = \inf\Big\{n \geq 1 : X_n = 0\Big\} \quad \text{and} \quad \tau_{M,R} = \inf\Big\{n \geq 1 : Y_n = 0\Big\}. \tag{7}$$

We observe that, for $n \geq 1$,

$$X_n = S_n^{(M)} - S_n^{(L)} = a + \sum_{k=1}^{n}\Big[I_k^{(M)} - I_k^{(L)}\Big] = a + \sum_{k=1}^{n} D_k^{(M,L)}$$

where $D_k^{(M,L)} = I_k^{(M)} - I_k^{(L)}$ for any any $k \geq 1$. Note that $\mathbb{E}\Big(D_k^{(M,L)}\Big) = \mathbb{E}\Big(I_k^{(M)}\Big) - \mathbb{E}\Big(I_k^{(M)}\Big) = 0$ and $\mathbb{V}\text{ar}\Big(D_k^{(M,L)}\Big) = \mathbb{V}\text{ar}\Big(I_k^{(M)}\Big) + \mathbb{V}\text{ar}\Big(I_k^{(L)}\Big) = 8p(1-p)$. Thus, it is clear that the difference process $\{X_n : n \geq 0\}$ can also be be presented as a random walk with increments having mean 0 with finite variance. Therefore, using the result of Spitzer (1964), we may conclude that is finite almost surely. However, we will provide a direct argument and will actually compute the generating function of the collision time of the middle random walk and left random walk.

**Theorem 1:** Under the Assumption, we have

$$\tau_{L,M} < +\infty \text{ almost surely.}$$

Note that there is nothing special about the middle and left random walks. The result may be applied to any pair of random walks. So, as a corollary, we also have

**Corollary 1:** Under the Assumption, we have

$$\tau_{M,R} < +\infty \text{ almost surely.}$$

We will prove the result using martingale method. The method is inspired by the results in Williams (1991).Let us define the filtration $\left\{\mathcal{F}_n^{(M,L)} : n \geq 0\right\}$, where

$$\mathcal{F}_n^{(M,L)} = \sigma\Big(I_k^{(L)}, I_k^{(M)} : k \leq n\Big) = \sigma\Big(S_k^{(L)}, S_k^{(M)} : k \leq n\Big)$$

is the $\sigma$-algebra generated by the increment random variables of the middle random walk and the left random walk up to time $n$. Also, this is same as the $\sigma$-algebra generated by the middle random walk and the left random walk up to time $n$. This is the natural filtration associated with two random walks we are studying.

We have already observed that

$$X_n = a + \sum_{k=1}^{n} \Big(I_k^{(M)} - I_k^{(L)}\Big)$$

for $n \geq 0$. The random variables $\{I_k^{(M)} - I_k^{(L)} : k \geq 1\}$ is a sequence of independently and identically distributed random variables with common distribution being the same as of a random variable taking values $+2$ with probability $p(1-p)$, $-2$ with probability $p(1-p)$ and 0 with probability $1 - 2p(1-p)$. Let us set $\alpha = p(1-p)$.

For $\lambda \in \mathbb{R}$, let us define, the Laplace transform of the common increment distribution by

$$f(\lambda) = \mathbb{E}\Big[\exp\Big(-\lambda\Big(I_1^{(M)} - I_1^{(L)}\Big)\Big)\Big] = \alpha\Big(e^{2\lambda} + e^{-2\lambda}\Big) + (1 - 2\alpha). \tag{8}$$

Clearly, we have

$$f(\lambda) = \alpha\Big(e^{2\lambda} + e^{-2\lambda} - 2\Big) + 1 = \alpha\Big(e^{\lambda} - e^{-\lambda}\Big)^2 + 1.$$

This implies that $f(\lambda) > 1$ for $\lambda \in \mathbb{R}$ and $f(\lambda) = 1$ for $\lambda = 0$. Also, by continuity of $f$ at 0, $f(\lambda) \downarrow 1$ as $\lambda \to 0$.

Let us define, for $n \geq 0$,

$$Z_n = \exp\left(-\lambda X_n\right) \Big(f(\lambda)\Big)^{-n}. \tag{9}$$

We first show

**Proposition 1:** The sequence $\{Z_n : n \geq 0\}$ is an $\mathcal{F}_n^{(M,L)}$-martingale.

**Proof:** Clearly $Z_0 = \exp\Big(-\lambda X_0\Big) = \exp(-\lambda a)$. We observe that the $X_n$ is $\mathcal{F}_n^{(M,L)}$ adapted by definition. Since $Z_n$ is a measurable function of $X_n$, $Z_n$ is also $\mathcal{F}_n^{(M,L)}$ adapted. It is easy to check that for each $n \geq 0$, we have $|Z_n| \leq \exp\Big(|\lambda|(a+n)\Big)$ and hence $\mathbb{E}(|Z_n|) < \infty$ for all $n \geq 1$.

Now, to show $\{Z_n : n \geq 0\}$ is a martingale with respect to $\mathcal{F}_n^{(M,L)}$, we note that $X_n$ is measurable with respect to $\mathcal{F}_n^{(M,L)}$. We have

$$\begin{aligned}
&\mathbb{E}\Big(Z_{n+1} \mid \mathcal{F}_n^{(M,L)}\Big) \\
&= \mathbb{E}\Big[\exp\left(-\lambda X_{n+1}\right)\Big(f(\lambda)\Big)^{-n-1} \mid \mathcal{F}_n^{(M,L)}\Big] \\
&= \mathbb{E}\Big[\exp\Big(-\lambda\Big(X_n + I_{n+1}^{(M)} - I_{n+1}^{(L)}\Big)\Big)\Big(f(\lambda)\Big)^{-n-1} \mid \mathcal{F}_n^{(M,L)}\Big] \\
&= \exp\Big(-\lambda X_n\Big)\Big(f(\lambda)\Big)^{-n-1}\mathbb{E}\left[\exp\Big(-\lambda\Big(I_{n+1}^{(M)} - I_{n+1}^{(L)}\Big)\Big)\right] \\
&= \exp\Big(-\lambda X_n\Big)\Big(f(\lambda)\Big)^{-n-1} f(\lambda) = \exp\Big(-\lambda X_n\Big)\Big(f(\lambda)\Big)^{-n} = Z_n.
\end{aligned}$$

This completes the proof of the proposition. □

Now we prove Theorem 1.

**Proof:** We note that

$$\{\tau_{L,M} = n\} = \{X_0 = a > 0, X_1 > 0, \ldots, X_{n-1} > 0, X_n = 0\}$$

and hence $\{\tau_{L,M} = n\} \in \mathcal{F}_n^{(M,L)}$. Thus, $\tau_{L,M}$ is a stopping time relative to $\{\mathcal{F}_n^{(M,L)}\}$. Hence, the family $\Big\{Z_{n\wedge\tau_{L,M}} : n \geq 0\Big\}$ is also a $\mathcal{F}_n^{(M,L)}$-martingale. Therefore, we obtain

$$\begin{aligned}
&\mathbb{E}\Big(\exp(-\lambda X_{n\wedge\tau_{L,M}})\Big(f(\lambda)\Big)^{n\wedge\tau_{L,M}}\Big) = \mathbb{E}\left(Z_{n\wedge\tau_{L,M}}\right) \\
&= \mathbb{E}\left(Z_{0\wedge\tau_{L,M}}\right) = \mathbb{E}\left(Z_0\right) = \exp(-\lambda a). \qquad (10)
\end{aligned}$$

Now, we specialize to the case of $\lambda > 0$ and take limit as $n \to \infty$ in equation (10). We have already noted that $f(\lambda) > 1$ for $\lambda \in \mathbb{R}$, in particular for $\lambda > 0$.

- On the event $\{\tau_{L,M} = +\infty\}$, clearly $\Big(f(\lambda)\Big)^{-n\wedge\tau_{L,M}} \to 0$ as $n \to \infty$.
- On the event $\{\tau_{L,M} < \infty\}$, we have $X_{n\wedge\tau_{L,M}} \to X_{\tau_{L,M}} = 0$. Thus, $\exp(-\lambda X_{n\wedge\tau_{L,M}}) \to 1$ as $n \to \infty$ and $\Big(f(\lambda)\Big)^{-n\wedge\tau_{L,M}} \to \Big(f(\lambda)\Big)^{-\tau_{L,M}}$ as $n \to \infty$.

Combining, we have

$$\exp\left(-\lambda X_{n\wedge\tau_{L,M}}\right)\left(f(\lambda)\right)^{-\left(n\wedge\tau_{L,M}\right)} \to \mathbb{I}\left(\tau_{L,M} < \infty\right)\left(f(\lambda)\right)^{-\tau_{L,M}}$$

as $n \to \infty$. Further, we observe that

- For all $n \geq 0$, $X_{n\wedge\tau_{L,M}} \geq 0$. For $\lambda > 0$, this implies that
$$\exp\left(-\lambda X_{n\wedge\tau_{L,M}}\right) \leq 1.$$
- Since $f(\lambda) > 1$ for $\lambda > 0$ and $n \geq 0$, we have
$$(f(\lambda))^{-(n\wedge\tau_{L,M})} \leq 1.$$

Thus, we have
$$\exp(-\lambda X_{n\wedge\tau_{L,M}})\Big(f(\lambda)\Big)^{n\wedge\tau_{L,M}} \leq 1.$$
Thus, we can use DCT in equation (10) to obtain, for all $\lambda > 0$,
$$\mathbb{E}\bigg(\mathbb{I}\left(\tau_{L,M} < \infty\right)\Big(f(\lambda)\Big)^{-\tau_{L,M}}\bigg) = \exp(-\lambda a). \tag{11}$$

Now, we will take limit by letting $\lambda \downarrow 0$ in equation (11). On the event $\{\tau_{L,M} < \infty\}$, using continuity of $f$, we get $\Big(f(\lambda)\Big)^{-\tau_{L,M}} \to 1$ as $\lambda \downarrow 0$. Therefore, we have
$$\mathbb{I}\left(\tau_{L,M} < \infty\right)\Big(f(\lambda)\Big)^{-\tau_{L,M}} \to \mathbb{I}\left(\tau_{L,M} < \infty\right).$$
Furthermore, we have
$$\mathbb{I}\left(\tau_{L,M} < \infty\right)\Big(f(\lambda)\Big)^{-\tau_{L,M}} \leq 1$$
as $f(\lambda) > 1$ for any $\lambda > 0$. Thus, by apply DCT in (11), we have
$$\begin{aligned}\mathbb{P}\Big(\tau_{L,M} < \infty\Big) &= \mathbb{E}\Big(\mathbb{I}\left(\tau_{L,M} < \infty\right)\Big)\\ &= \lim_{\lambda\downarrow 0}\mathbb{E}\bigg(\mathbb{I}\left(\tau_{L,M} < \infty\right)\Big(f(\lambda)\Big)^{-\tau_{L,M}}\bigg)\\ &= \lim_{\lambda\downarrow 0}\exp(-\lambda a) = 1.\end{aligned}$$
This proves that $\tau_{L,M} < \infty$ with probability 1. □

The result in (11) yields more information. Indeed, we may calculate the probability generating function of $\tau_{L,M}$, in in turn provides more information.

**Corollary 2:** The probability generating function of $\tau_{L,M}$ is given by
$$\mathbb{E}\Big(s^{\tau_{L,M}}\Big) = \frac{1}{(2\sqrt{\alpha})^a}\left[\sqrt{\frac{1}{s} - 1 + 4\alpha} - \sqrt{\frac{1}{s} - 1}\right]^a \tag{12}$$
for $-1 < s \leq 1$.

**Proof:** Since $\tau_{L,M} < \infty$ almost surely, we can rewrite equation (11), for all $\lambda > 0$
$$\mathbb{E}\left((f(\lambda))^{-\tau_{L,M}}\right) = \exp(-\lambda a).$$

This formula may be used to get the probability generating function of $\tau_{L,M}$. Letting $s = \Big(f(\lambda)\Big)^{-1}$ for $\lambda > 0$ and solving $\lambda$ in terms of $s$, we have

$$\mathbb{E}\left(s^{\tau_{L,M}}\right) = \exp(-\lambda a) = \frac{1}{(2\sqrt{\alpha})^a}\left[\sqrt{\frac{1}{s} - 1 + 4\alpha} - \sqrt{\frac{1}{s} - 1}\right]^a .$$

This proves the corollary. □

This may be used to show that the expectation is infinite. Indeed, we have

$$\frac{d}{ds}\mathbb{E}\left(s^{\tau_{L,M}}\right) = \frac{a}{(2\sqrt{q})^a}\left[\sqrt{\frac{1}{s} - 1 + 4q} - \sqrt{\frac{1}{s} - 1}\right]^{a-1} \times \frac{1}{2s^2}\left[\frac{1}{\sqrt{\frac{1}{s} - 1}} - \frac{1}{\sqrt{\frac{1}{s} - 1 + 4q}}\right].$$

So, when $s \uparrow 1$, the right hand side diverges to $\infty$. Thus, $\mathbb{E}\left(\tau_{L,M}\right) = \infty$. Similarly we can also prove that $\mathbb{E}\left(\tau_{M,R}\right) = \infty$. We may also obtain the tail behaviour of the stopping time.

## 3. Collision time of three random walks : simulation

Before we go into the theoretical derivation, we carry out some simulation studies. Here we use a cutoff, to stop the process the process if the the simulation has not resulted in a value. Our cutoff is 10000000 and we have simulated for 10000000 times. We have also taken different values of $a$ and $b$ where $a$ and $b$ are both even positive integers. We have carried out the simulation using 3 different values of $p$, which are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{5}{7}$ respectively.

For $p = \frac{1}{2}$, $y_1$ is the observed mean of the first collision time of three random walks after simulating it 10000000 times, For $p = \frac{1}{3}$, $y_2$ is the observed mean of the first collision time of three random walks after simulating it 10000000 times, For $p = \frac{5}{7}$, $y_3$ is the observed mean of the first collision time of three random walks after simulating it 10000000 times. Now we will look at the scatter plots of $(ab, y_1)$, $(ab, y_2)$ and $(ab, y_3)$ and also we will find and plot regression lines of $y_1$ on $ab$, $y_2$ on $ab$ and $y_3$ on $ab$. Here $S_0^{(L)}$, $S_0^{(M)}$ and $S_0^{(R)}$ are $-a$, 0 and $+b$ respectively.

### Simulation output

**Table 1: Simulation of expected collision times**

| $-a$ | $+b$ | $y_1$ | $y_2$ | $y_3$ | $ab$ |
|---|---|---|---|---|---|
| -2 | 2 | 3.9987 | 4.4956 | 4.8801 | 4 |
| -2 | 4 | 7.9961 | 9.0174 | 9.7983 | 8 |
| -2 | 6 | 12.0102 | 13.4858 | 14.6516 | 12 |
| -2 | 8 | 15.9895 | 17.9811 | 19.6139 | 16 |
| -2 | 10 | 20.0246 | 22.5134 | 24.4733 | 20 |
| -2 | 12 | 24.0198 | 27.0171 | 29.3998 | 24 |
| -2 | 14 | 28.0139 | 31.4881 | 34.3256 | 28 |
| -2 | 16 | 31.9907 | 35.9944 | 39.2114 | 32 |
| -2 | 18 | 35.9821 | 40.4913 | 44.0897 | 36 |

**Table 1: Simulation of expected collision times**

| $-a$ | $+b$ | $y_1$ | $y_2$ | $y_3$ | $ab$ |
|---|---|---|---|---|---|
| -2 | 20 | 40.0912 | 44.9591 | 48.9771 | 40 |
| -4 | 4 | 15.9931 | 17.5619 | 19.5812 | 16 |
| -4 | 6 | 23.9978 | 27.0127 | 29.3665 | 24 |
| -4 | 8 | 31.9914 | 36.0223 | 39.1997 | 32 |
| -4 | 10 | 40.0297 | 45.0136 | 49.0315 | 40 |
| -4 | 12 | 47.9956 | 53.9889 | 58.7969 | 48 |
| -4 | 14 | 55.9992 | 62.9156 | 68.5899 | 56 |
| -4 | 16 | 63.9958 | 71.9929 | 78.3878 | 64 |
| -4 | 18 | 72.0154 | 81.0147 | 88.2156 | 72 |
| -4 | 20 | 80.0083 | 90.0396 | 97.9089 | 80 |
| -6 | 6 | 35.9841 | 40.5069 | 44.0989 | 36 |
| -6 | 8 | 47.9892 | 53.9574 | 58.7899 | 48 |
| -6 | 10 | 59.9946 | 67.4998 | 73.5017 | 60 |
| -6 | 12 | 71.9839 | 80.9758 | 88.1898 | 72 |
| -6 | 14 | 84.0629 | 94.5195 | 102.8761 | 84 |
| -6 | 16 | 95.9779 | 108.0251 | 117.6112 | 96 |
| -6 | 18 | 108.0022 | 121.5245 | 132.2893 | 108 |
| -6 | 20 | 119.9141 | 134.9596 | 146.9674 | 120 |
| -8 | 8 | 63.9951 | 72.0212 | 78.3894 | 64 |
| -8 | 10 | 79.9917 | 90.0018 | 97.9825 | 80 |
| -8 | 12 | 95.9679 | 107.9786 | 117.5997 | 96 |
| -8 | 14 | 112.0091 | 125.9925 | 137.2119 | 112 |
| -8 | 16 | 127.9899 | 143.9512 | 156.7898 | 128 |
| -8 | 18 | 143.9769 | 161.9213 | 176.2996 | 144 |
| -8 | 20 | 160.0998 | 179.9621 | 196.0176 | 160 |
| -10 | 10 | 100.0518 | 112.5185 | 122.4886 | 100 |
| -10 | 12 | 119.9371 | 135.0121 | 147.0259 | 120 |
| -10 | 14 | 139.9145 | 157.4852 | 171.4966 | 140 |
| -10 | 16 | 159.9159 | 179.9597 | 195.9979 | 160 |
| -10 | 18 | 180.0263 | 202.5096 | 220.3999 | 180 |
| -10 | 20 | 199.9564 | 224.9917 | 244.9732 | 200 |
| -12 | 12 | 143.9768 | 161.9129 | 176.3993 | 144 |
| -12 | 14 | 168.0459 | 189.0432 | 205.7915 | 168 |
| -12 | 16 | 192.0091 | 216.0278 | 235.2112 | 192 |
| -12 | 18 | 215.9316 | 242.9841 | 264.5889 | 216 |
| -12 | 20 | 239.9089 | 269.9124 | 294.0113 | 240 |
| -14 | 14 | 195.9989 | 220.5398 | 240.1376 | 196 |
| -14 | 16 | 223.9388 | 251.9492 | 274.2998 | 224 |
| -14 | 18 | 251.9164 | 283.4919 | 308.6779 | 252 |
| -14 | 20 | 279.9936 | 315.0154 | 342.9547 | 280 |
| -16 | 16 | 255.9989 | 287.9754 | 313.6291 | 256 |
| -16 | 18 | 287.9669 | 324.0478 | 352.7959 | 288 |
| -16 | 20 | 319.9799 | 359.9954 | 391.9286 | 320 |

**Table 1: Simulation of expected collision times**

| $-a$ | $+b$ | $y_1$ | $y_2$ | $y_3$ | $ab$ |
|---|---|---|---|---|---|
| -18 | 18 | 323.9193 | 364.3991 | 396.8777 | 324 |
| -18 | 20 | 359.9899 | 404.9145 | 441.1223 | 360 |
| -20 | 20 | 399.9918 | 449.7982 | 489.8979 | 400 |

The scatter plots of the above data is very instructive as they clearly bring out the relation between $ab$ and the expected time of the first collision time $\tau_c$.

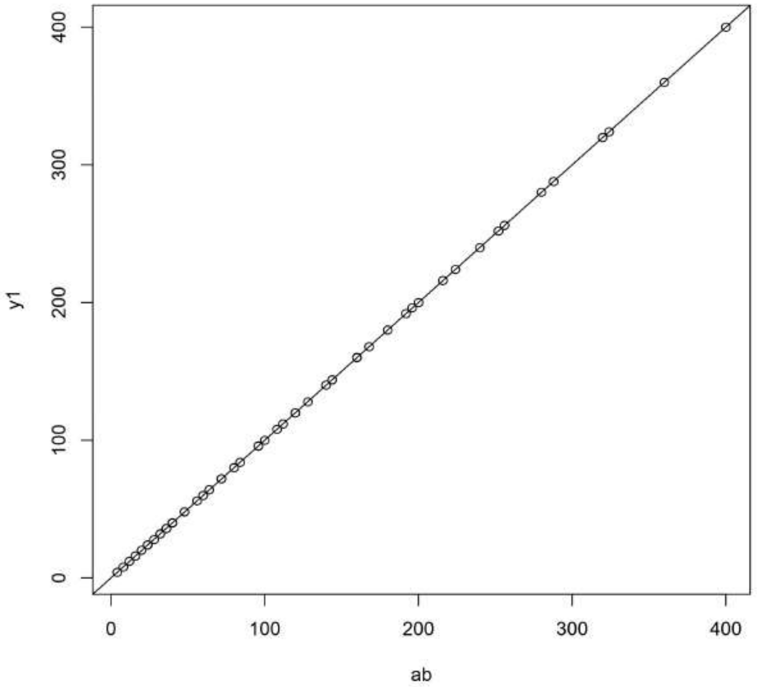


**Figure 1: Scatter plot of $(ab,y_1)$ and regression line of $y_1$ on $ab$**

**Table 2: Summary statistics of simulation**

| Statistics | Estimate | T statistics | P value |
|---|---|---|---|
| $Constant_1$ | 0.00727396818 | 0.8787564386 | 0.3835000647 |
| $Slope_1$ | 0.9998608551 | 19192.1238453652 | 0 |
| $Constant_2$ | -0.0096277426 | -0.6188423746 | 0.5386710900 |
| $Slope_2$ | 1.1248997399 | 11488.2891168571 | 0 |
| $Constant_3$ | -0.0077084624 | -0.9309990213 | 0.3560755895 |
| $Slope_3$ | 1.2249623703 | 23506.6195548538 | 0 |

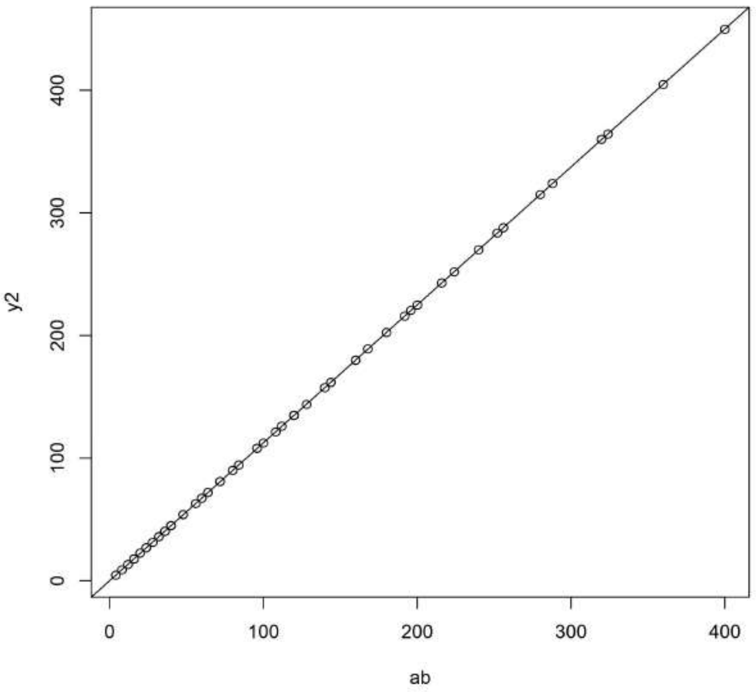


**Figure 2: Scatter plot of ($ab$,$y_2$) and regression line of $y_2$ on $ab$**

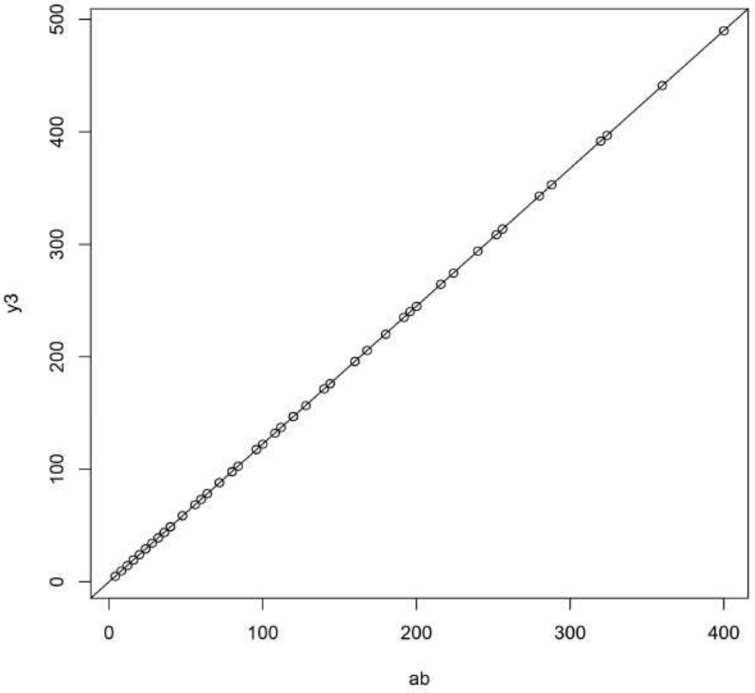


**Figure 3: Scatter plot of ($ab$,$y_3$) and regression line of $y_3$ on $ab$**

The regression lines on $ab$ are for different values of $p$:

$$\begin{aligned}
\widehat{y_1} &= 0.00727396818 + 0.9998608551 \times ab \\
\widehat{y_2} &= -0.0096277426 + 1.1248997399 \times ab \\
\widehat{y_3} &= -0.0077084624 + 1.2249623703 \times ab.
\end{aligned}$$

The correlation coefficients are 0.9999999281, 0.9999997992, 0.9999999952 respectively. In each of the three cases the correlation coefficient is very close to $+1$, so here we can observe near perfect positive correlation.

The summary statistics of the above data, which from the above scatter plots is quite expected, clearly shows that there should be a linear relationship between the expected time and the product of the initial distances $ab$. In each of the three cases the estimate of the constant is very close to 0 and the estimate of the slope is very close to $\Big(4p(1-p)\Big)^{-1}$. Also in each of the three cases the p-value of the intercept is greater than 0.05, so the intercept is not significant. From these observations we postulate that the expectation of $\tau_c$ should be $ab\Big(4p(1-p)\Big)^{-1}$. In the next section we derive these theoretical results.

## 4. Theoretical results

We first note that we are working with random walks having steps size of $\pm 1$ with the starting points are on even lattice. Therefore, these independent random walks do not cross each other before intersecting. So, we can write the first collision time of these three random walks $\tau_c$ as,

$$\tau_c = \min\Big\{\tau_{L,M}, \tau_{M,R}\Big\}. \tag{13}$$

As an immediate consequence of Theorem 1, we have

$$\tau_c < +\infty \text{ with probability 1.}$$

Further from the above observation, it is easy to conclude that at $\tau_c$ either the pair of left random walk and the middle random walk collides or the pair of middle random walk and the right random walk collides. So, we can rephrase the definition of $\tau_c$ (see equation (2)) as follows:

$$\begin{aligned}
\tau_c &= \inf\Big\{n \geq 1 : (S_n^{(M)} - S_n^{(L)})(S_n^{(R)} - S_n^{(M)})(S_n^{(L)} - S_n^{(R)}) = 0\Big\} \\
&= \inf\Big\{n \geq 1 : (S_n^{(M)} - S_n^{(L)})(S_n^{(R)} - S_n^{(M)}) = 0\Big\} \\
&= \inf\Big\{n \geq 1 : X_n Y_n = 0\Big\}. \qquad (14)
\end{aligned}$$

We will use this identification to justify these results.

We will again use the martingale method. Let us define the filtration $\Big\{\mathcal{F}_n : n \geq 0\Big\}$, where

$$\mathcal{F}_n = \sigma\Big(I_k^{(L)}, I_k^{(M)}, I_k^{(R)} : k \leq n\Big) = \sigma\Big(S_k^{(L)}, S_k^{(M)}, S_k^{(R)} : k \leq n\Big)$$

is the $\sigma$-algebra generated by the increment random variables of all the random walks. Also, this is same as the $\sigma$-algebra generated by the all the random walk up to time $n$. This is the natural filtration associated with all three random walks we are studying.

**Proposition 2:** The family $\{X_nY_n + 4np(1-p) : n \geq 0\}$ is an $\mathcal{F}_n$-martingale.

**Proof:** It is easy to see that random variable $X_nY_n+4np(1-p)$ is $\mathcal{F}_n$-adapted for any $n \geq 0$. Further, for any $n \geq 0$,

$$|X_nY_n| \leq (a+2n)(b+2n)$$

Thus, we have $\mathbb{E}\Big(|X_nY_n + 4np(1-p)|\Big) < \infty$ for all $n \geq 0$.

Now, we have

$$\begin{aligned}
&X_{n+1}Y_{n+1} + 4(n+1)p(1-p) \\
&= \Big(X_n + \big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\big)\Big)\Big(Y_n + \big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\big)\Big) + 4(n+1)p(1-p) \\
&= X_nY_n + X_n\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) + Y_n\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big) \\
&\quad + \Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big) + 4(n+1)p(1-p).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&(X_{n+1}Y_{n+1} + 4(n+1)p(1-p)) - (X_nY_n + 4np(1-p)) \\
&= X_n\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) + Y_n\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big) + \Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) + 4p(1-p).
\end{aligned}$$

Note that $X_n$ and $Y_n$ are $\mathcal{F}_n$-measurable and the random variables $\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big)$, $\Big(I^{(M)}_{n+1}-I^{(L)}_{n+1}\Big)$ are independent of $\mathcal{F}_n$ with expectation 0. Further, the random variables $I^{(L)}_{n+1}, I^{(M)}_{n+1}$ and $I^{(R)}_{n+1}$ are also independent of $\mathcal{F}_n$ and are independent with expectation $2p-1$ and variance $4p(1-p)$.

Now, we take conditional expectation with respect to $\mathcal{F}_n$. Observe that

- $\mathbb{E}\Big[X_n\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = X_n\mathbb{E}\Big[\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = X_n\mathbb{E}\Big[\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big)\Big] = 0$ where we have used the fact that $X_n$ is $\mathcal{F}_n$-measurable and the increments random variables are independent of $\mathcal{F}_n$.

- Similarly we have $\mathbb{E}\Big[Y_n\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = 0$ .

- Finally, using the fact that the increments are independent of $\mathcal{F}_n$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) \mid \mathcal{F}_n\Big] \\
&= \mathbb{E}\Big[\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big] \\
&= \mathbb{E}\Big[\Big((I^{(M)}_{n+1} - (2p-1)) - (I^{(L)}_{n+1} - (2p-1)\Big)\Big((I^{(R)}_{n+1} - (2p-1)) - (I^{(M)}_{n+1} - (2p-1))\Big)\Big] \\
&= -\mathbb{V}\mathrm{ar}\Big(I^{(M)}_{n+1}\Big) = -4p(1-p).
\end{aligned}$$

Combing the above and the fact that $X_nY_n$ is measurable with respect to $\mathcal{F}_n$, we now have

$$\mathbb{E}\Big(X_{n+1}Y_{n+1} + 4(n+1)p(1-p) \mid \mathcal{F}_n\Big) = X_nY_n + 4np(1-p).$$

This proves the proposition. □

Next we proves another similar proposition.

**Proposition 3:** The family $\{X_nY_n\,(X_n + Y_n) : n \geq 0\}$ is an $\mathcal{F}_n$-martingale.

**Proof:** The adaptedness of $X_nY_n(X_n+Y_n)$ with respect $\mathcal{F}_n$ is again straightforward. Further, it is also obvious that $|X_nY_n(X_n+Y_n)| \leq (a+2n)(b+2n)(a+b+4n)$ and hence $\mathbb{E}\Big(|X_nY_n(X_n+Y_n)|\Big) < \infty$ for any $n \geq 0$.

As in the earlier proposition, we have

$$\begin{aligned}
&X_{n+1}Y_{n+1}\,(X_{n+1} + Y_{n+1})\\
&= \Big(X_n + \Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big)\Big(Y_n + \Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big)\Big(X_n + Y_n + \Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big)\Big)\\
&= X_nY_n\,(X_n + Y_n) + X_nY_n\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) + X_n\,(X_n + Y_n)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\\
&\qquad + Y_n\,(X_n + Y_n)\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big) + X_n\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big)\\
&\qquad + Y_n\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) + (X_n + Y_n)\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\\
&\qquad \Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big).
\end{aligned}$$

As in the previous proposition, we have $X_n$ and $Y_n$ are $\mathcal{F}_n$-measurable and the random variables $\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)$, $\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)$ and $\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big)$ are independent of $\mathcal{F}_n$ with expectation 0. Thus, same arguments as above, apply to show that

- $\mathbb{E}\Big[X_nY_n\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = \mathbb{E}\Big[X_n(X_n+Y_n)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = \Big[Y_n(X_n+Y_n)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = 0.$

- Same arguments as above, yield

$$\begin{aligned}
(a)\quad & \mathbb{E}\Big[X_n\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = X_n\mathbb{V}\mathrm{ar}\Big(I^{(R)}_{n+1}\Big) = 4p(1-p)X_n\\
(b)\quad & \mathbb{E}\Big[Y_n\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(L)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = Y_n\mathbb{V}\mathrm{ar}\Big(I^{(L)}_{n+1}\Big) = 4p(1-p)Y_n\\
(c)\quad & \mathbb{E}\Big[\Big(X_n + Y_n\Big)\Big(I^{(M)}_{n+1} - I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1} - I^{(M)}_{n+1}\Big) \mid \mathcal{F}_n\Big] = -(X_n + Y_n)\mathbb{V}\mathrm{ar}\Big(I^{(M)}_{n+1}\Big)\\
& \qquad = -4p(1-p)(X_n + Y_n).
\end{aligned}$$

- We also have

$$\begin{aligned}
&\mathbb{E}\Big[\Big(I^{(M)}_{n+1}-I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big)\Big(I^{(R)}_{n+1}-I^{(L)}_{n+1}\Big)\mid \mathcal{F}_n\Big]\\
&=\mathbb{E}\Big[\Big(I^{(M)}_{n+1}-I^{(L)}_{n+1}\Big)\Big(I^{(R)}_{n+1}-I^{(M)}_{n+1}\Big)\Big(I^{(R)}_{n+1}-I^{(L)}_{n+1}\Big)\Big]\\
&=\mathbb{E}\Big[\Big((I^{(M)}_{n+1}-(2p-1))-(I^{(L)}_{n+1}-(2p-1))\Big)\Big((I^{(R)}_{n+1}-(2p-1))-(I^{(M)}_{n+1}-(2p-1))\Big)\\
&\qquad\times\Big((I^{(R)}_{n+1}-(2p-1))-(I^{(L)}_{n+1}-(2p-1))\Big)\Big]=0
\end{aligned}$$

by independence of the random variables and the fact that they have expectation 0.

Combining the above and the fact that $X_nY_n(X_n+Y_n)$ is $\mathcal{F}_n$-measurable, we have

$$\mathbb{E}\Big(X_{n+1}Y_{n+1}\Big(X_{n+1}+Y_{n+1}\Big)\mid \mathcal{F}_n\Big)=X_nY_n\Big(X_n+Y_n\Big).$$

This completes the proof. □

Now, we are in a position to state and prove our main result.

**Theorem 2:** We have

$$\mathbb{E}\Big(\tau_c\Big)=ab\Big(4p(1-p)\Big)^{-1}. \tag{15}$$

**Proof:** We observe that, from equation (13), that

$$\{\tau_c=n\}=\Big\{X_0Y_0>0, X_1Y_1>0,\ldots,X_{n-1}Y_{n-1}>0,X_nY_n=0\Big\}.$$

Clearly $\{\tau_c=n\}\in\mathcal{F}_n$, which implies that $\tau_c$ is also stopping time relative to $\{\mathcal{F}_n\}$.

By using Proposition 2, we get that, $\Big\{X_{n\wedge\tau_c}Y_{n\wedge\tau_c}+4p(1-p)\,(n\wedge\tau_c):n\geq 0\Big\}$ is a martingale and hence for any $n\geq 1$,

$$\begin{aligned}
&\mathbb{E}\Big(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}+4p(1-p)\,(n\wedge\tau_c)\Big)\\
&=\mathbb{E}\Big(X_{0\wedge\tau_c}Y_{0\wedge\tau_c}+4p(1-p)\,(0\wedge\tau_c)\Big)\\
&=\mathbb{E}\Big(X_0Y_0\Big)=ab
\end{aligned}\tag{16}$$

since $\tau_c\geq 0$.

Now, we will take limit in equation (16) as $n\to\infty$. Since $\tau_c<\infty$ almost surely, $n\wedge\tau_c\uparrow\tau_c$ as $n\to\infty$. By MCT, we obtain

$$\mathbb{E}\Big(n\wedge\tau_c\Big)\to\mathbb{E}\Big(\tau_c\Big)$$

as $n\to\infty$.

To complete the proof we show that $\mathbb{E}\Big(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big) \to 0$ as $n \to \infty$. Since $\tau_c < \infty$ almost surely, we have that

$$X_{n\wedge\tau_c}Y_{n\wedge\tau_c} \to X_{\tau_c}Y_{\tau_c} = 0 \tag{17}$$

as $n \to \infty$.

In order to show that the expected value also converges to 0, we will use Theorem 26.13 of Billingsley (1986). For this we require to show that the sequence of random variable $\{X_{n\wedge\tau_c}Y_{n\wedge\tau_c} : n \geq 0\}$ is an uniformly integrable family. A sufficient condition for a family of random variables to be uniformly integrable (see Billingsley (1986)) is given by

$$\sup_{n\geq 0}\mathbb{E}\Big[\Big(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big)^{1+\epsilon}\Big] < \infty$$

for some $\epsilon > 0$.

By using Proposition 3, we get that $\Big\{X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\left(X_{n\wedge\tau_c} + Y_{n\wedge\tau_c}\right) : n \geq 0\Big\}$ is also a martingale. Hence, for any $n \geq 1$,

$$\begin{aligned}
&\mathbb{E}\Big[X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big(X_{n\wedge\tau_c} + Y_{n\wedge\tau_c}\Big)\Big]\\
&= \mathbb{E}\Big[X_{0\wedge\tau_c}Y_{0\wedge\tau_c}\Big(X_{0\wedge\tau_c} + Y_{0\wedge\tau_c}\Big)\Big]\\
&= \mathbb{E}\Big[X_0Y_0\Big(X_0 + Y_0\Big)\Big]\\
&= ab(a+b).
\end{aligned}$$

For non-negative $u, v \geq 0$, using AM-GM inequality, we have $(uv)^{3/2} \leq \frac{1}{2}uv(u+v)$. Since $X_{n\wedge\tau_c}$ and $Y_{n\wedge r_c}$ are both non negative, we have, for any $n \geq 0$

$$\mathbb{E}\Big[\Big(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big)^{3/2}\Big] \leq \frac{1}{2}\mathbb{E}\Big[X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big(X_{n\wedge\tau_c} + Y_{n\wedge\tau_c}\Big)\Big] = \frac{1}{2}ab(a+b).$$

Therefore,

$$\sup_{n\geq 0}\mathbb{E}\Big[\Big(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Big)^{1+1/2}\Big] \leq \frac{1}{2}ab(a+b) < \infty.$$

Hence, we conclude that $\{X_{n\wedge\tau_c}Y_{n\wedge\tau_c} : n \geq 0\}$ is an uniformly integrable family. Therefore, we have

$$\mathbb{E}\Bigg(X_{n\wedge\tau_c}Y_{n\wedge\tau_c}\Bigg) \to 0 \text{ as } n \to \infty.$$

This completes the proof of the Theorem. □

## Acknowledgements

Most of the work was completed when Arighna Dey was pursuing the M. Stat. degree from the Indian Statistical Institute and he wishes to the thank ISI for this opportunity. Kumarjit Saha wishes to thank the Ashoka University for the support. Anish Sarkar wishes to the Indian Statistical Institute for support and also the organizers of the conference for giving the an opportunity to present the work.

## References

Billingsley, P. (1986). *Probability and Measure.* John Wiley and Sons.

Coupier, D., Saha, K., Sarkar, A., and Tran, V. C. (2020). Collision times of random walks and applications to the Brownian web. *Genealogies of interacting particle systems, Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, World Scientific ISBN - 978-981-120-608-5.*

Spitzer, F. (1964). *Principles of Random Walk.* Van Nostrand.

Williams, D. (1991). *Probability with Martingales.* Cambridge University Press.

## ANNEXURE

### R code for simulation

```
Increment<-function (uniform, p)
{
    # uniform := uniform variable
    # p := probability of increment of +1,

    # output := the increment with probability distribution

    if (uniform <= p)
    {
        return (1)
    }
    return (-1)

}

FindCollision<-function(
    startright,
    startmid,
    startleft,
    p,
    cutoff)
{
    # startright := starting position of right random walk
    # startmid := starting position of mid random walk
    # startleft := starting position of left random walk
    # cutoff := the max length of random walk to be considered

    # output := First Collision time of 3 random walks

    # initialize starting positions
    rightpos = startright
    midpos = startmid
    leftpos = startleft

    # set time for collision to cutoff + 1
    time = cutoff+1

    # run the loop until cutoff time
    for (i in 1:cutoff)
    {

        # Get three uniforms
        uniforms = runif(3)

        # update the random walk positions
        rightpos = rightpos + Increment(uniforms[1], p)
        midpos = midpos + Increment(uniforms[2], p)
        leftpos = leftpos + Increment(uniforms[3], p)

        # Check for collision
        if ( (rightpos - midpos)*(midpos-leftpos) == 0 )
        {
            # Collision has happened
            # set time to this collision time
            time = i

            # stop the simulation
            break
        }
    }

    # return the time
    return (time)
}

RW<-function (
    startright,
    startmid,
    startleft,
    p,
    cutoff,
    num)
{
    # output := mean of First Collision times of num repeatation
    W = rep (0, num)

    # run loop for repeatations of times
    for (i in 1:num)
    {
        W[i] = FindCollision(
             startright,
             startmid,
             startleft,
             p,
             cutoff)
    }
    ava = c(mean(W))
    return (ava)
}
```

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 169-203

# Analysis of Large Medical Databases: Addressing the Clinical Relevance of Statistically Significant Findings

**Rachana Lele[1a,b], Anand Seth[2], Sameer Patel[1d], Jianmin Pan[1a-c], Marepalli B. Rao[1a], and Shesh N. Rai[1a-c, 3]**

*[1a] Division of Biostatistics and Bioinformatics, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Ohio*
*[1b]Cancer Data Science Center, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Ohio*
*[1c]Biostatistics and Informatics Shared Resources, University of Cincinnati Cancer Center, Ohio*
*[1d]Department of Surgery, College of Medicine, University of Cincinnati, Ohio*
*[2]Research and Development, SK Patent Associates, LLC, Ohio*
*[3]Division of Environmental Cardiology, School of Medicine, University of Louisville*

Received: 01 August 2023; Revised: 13 August 2023; Accepted: 16 August 2023

## Abstract

The analysis of studies using large medical databases has gained popularity due to their ability to provide extensive and diverse samples. However, in the currently published literature, the selection of samples in such studies often relies on inclusion criteria based solely on the study's objectives, rather than utilizing formal sample size calculation techniques. Also, inferences are predominantly drawn based on *p*-values, which tend to be highly significant due to large samples but may lack clinical relevance. In this article, we explore the issue of statistically significant *p*-values but with limited clinical relevance when analyzing large databases. We propose the incorporation of effect sizes, a concept well-established in the literature, to supplement *p*-values in assessing the practical significance of research findings. To address the unique challenges of analyzing large samples using logistic regression, we present a novel effect size measure specifically tailored for this context. Moreover, we introduce conventions for interpreting effect sizes when analyzing large databases, thus providing researchers with a standardized approach for evaluating the magnitude of the observed associations. To validate the proposed effect size measure, we employ state-of-the-art machine learning techniques on the same datasets and demonstrate its robustness and utility in large-scale medical studies. To illustrate the statistical challenges and the application of our novel effect size measure, we present a compelling case study utilizing breast cancer data from the National Cancer Database (NCDB). Our findings shed light on the potential pitfalls of relying solely on *p*-values in large database studies and highlight the significance of incorporating effect sizes to better understand the clinical implications of research results. By emphasizing the importance of effect sizes in addition to *p*-values, this study aims to improve the accuracy and clinical relevance of statistical analyses for large medical databases. Implementing our suggested approach can lead to more informative and meaningful insights, thereby contributing to the advancement of evidence-based medicine and patient care.

Corresponding Author: Shesh N. Rai
Email: raise@ucmail.uc.edu

*Key words:* National cancer database; NCDB; Breast cancer; *p*-values; Effect sizes; Machine learning.

## 1. Introduction

Most traditional statistical analysis methods, such as linear regression, *t*-test, ANOVA, *etc.*, require the assumption of normality and randomized study sample selection. While conducting clinical research to test the safety and efficacy of new drugs, randomization is an essential component as well. Such studies require careful selection of a representative sample which is achieved using formal randomization techniques.

A relatively new branch of study, often referred as Real-World Data (RWD) and Real World Evidence (RWE), involves analyzing databases maintained by various government and private institutions to discover new insights related to public health which were previously underutilized, Breckenridge *et al.* (2019). Since data collection is lengthy and expensive, readily available databases provide an excellent alternative to conducting research and help save time, cost, and resources and complement evidence obtained from randomized clinical studies. However, large databases (which may also be referred as 'big data') are repositories of majority of the actual observed cases; hence, these are not randomized. Being extremely large, normality assumption is unrealistic and often unmet for most of these databases. Hence, using traditional parametric tests for such databases tend to produce highly significant results which may have no clinical relevance. The traditional statistical tests run using these large databases tend to produce highly significant *p*-values and inferences based on *p*-values alone and could lead to misleading or incorrect conclusions. Several articles such as Sullivan and Feinn (2012) and Solla *et al.* (2018) explore the alternative of using effect sizes and provide criticism for the use of *p*-values alone talk about the use of effect sizes and confidence intervals in addition to using p-values. They note highly significant *p*-values with small effect sizes may be clinically irrelevant as suggested by Ranstam *et al.* (2012) that confidence interval is a better alternative to using *p*-values.

Cohen (1988) and Cohen (1992) introduced the concept of effect sizes and defined it as the discrepancy between the null and the alternative hypotheses. He suggested formulae for effect sizes using normally distributed outcomes as well as proportions which were respectively popularized as Cohen's *d* and Cohen's *h*. A strength of the effect size measure is that it does not directly depend on the sample size and hence, is unaffected by large sample sizes. Consequently, for big data analysis involving large databases, effect size could be a better inferential measure than the traditional *p*-values. However, in the case of a logistic regression in which we are comparing effects of two treatments within two different categories of a variable, the Cohen's *h* effect size cannot be used in the present form.

In this paper, we argue that *p*-values alone can provide misleading results for extremely large sample sizes since *p*-value calculations depend on sample size. We compare the p-values obtained from an overall test and those obtained from individual tests as well as Bonferroni adjusted *p*-values. As an alternative to using *p*-values, we propose a modification/extension of Cohen's *h* effect size estimator for logistic regression. We validate our results obtained using the new Cohen's *h* measure using machine learning techniques such as Association Rule Mining (ARM) and Naïve Bayes classifier.

The organization of the paper is described as follows. In Section 2, we introduce statistical formulation of the issue of obtaining highly significant *p*-values for large sample sizes. We formally introduce the concept of alpha adjustment for multiple comparisons and introduce the theory of Bonferroni method of multiplicity adjustment. Furthermore, we introduce the theory of logistic regression, and we introduce the concept of effect size and explore the theory of different effect size measures. Lastly, we provide the theory for ARM and Naïve Bayes classifier. In Section 3, we describe our proposed modification/extension of Cohen's $h$ measure which can be applied to logistic regression analysis. In Section 4, we describe different statistical issues in the analysis of large databases using National Cancer Database (NCDB) as an example and present a literature review of articles published using NCDB. In Section 5, we present a case study using breast cancer data from NCDB to demonstrate the central issue of p-values addressed in this paper and how our proposed modified Cohen's $h$ effect size will lead to 'meaningful statistical significance' as opposed to clinically irrelevant statistical significance. We provide a strong support for our arguments by using ARM and Naïve Bayes classifier methods.

## 2. Statistical methods

### 2.1. Wald test

When analyzing data using statistical tests, $p$-values are often used to draw inferences. We will illustrate the effect of large sample size on $p$-values using the Wald test as established by Wald (1974). The traditional $t$-test statistic, binomial test statistic, Poisson test statistic *etc.* are special cases of the Wald test statistic. We will illustrate the dependence of the test statistic on the sample size using the simple case of Wald test for Bernoulli random variable. In the case of Bernoulli test as described by Klotz (1973), we observe independent binary responses, and we wish to draw inferences about the probability of an event in the population.

Suppose we sample $n$ individuals from a pre-specified population and the probability of occurrence of an event in this population is the same for an individual, say, $p$.

Let $Y_i$ denote the occurrence of an event for each individual $i$. Here, we define $Y_i = 1$ if an event occurs and $Y_i = 0$, otherwise. Thus, the observed data would be given by $Y_1$, $Y_2$, …, $Y_n$.

The maximum likelihood estimate (MLE) of $p$ is given by

$$\hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{1}$$

Now, suppose we are testing the hypothesis $H_0\colon p = p_0$ *vs.* $H_1\colon p \neq p_0$.

The Wald test statistic ($W$) is given by a difference in the MLE estimate of $p$ and the hypothesized value, normalized by the MLE estimate of the standard deviation.

Thus, we have

$$W = \frac{(\hat{p}-p_0)^2}{\hat{p}(1-\hat{p})/n} \tag{2}$$

This test can be extended for the case of logistic regression to test the significance of regression coefficients.

For $E(Y_i) = \pi_i$, we have

$logit(\pi_i) = \beta_0 + \beta_0 x_{i1} + \cdots + \beta_p x_{ip} = x_i{}'\beta$, where $x_i = \left(1, x_{i1}, \ldots, x_{ip}\right)'$ and $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right)'$.

To test a single $\beta$ coefficient value, the Wald test statistic will be given by

$$Z = \frac{\widehat{\beta}_j - \beta_{j0}}{\widehat{se}(\widehat{\beta})} \sim N(0,1) \tag{3}$$

where $\widehat{se}(\hat{\beta})$ is calculated by taking the inverse of the estimated information matrix.

From equation (2), we observe that $W$ statistic depends on the sample size, $n$. Note as $n$ becomes large in equation (2), *i.e.*, as $n \rightarrow \infty$, $W \rightarrow \infty$. Similarly, for the case of logistic regression, using equation (3), as $n \rightarrow \infty$, $Z \rightarrow \infty$.

The $p$-value may be defined as the probability of observing a test statistic as extreme as the one observed if the null hypothesis were true. Alternatively, $p$-value is the observed risk of rejecting $H_0$. For the Wald test, we have $p$-value, $p' = P(|Z| > |T_{obs}|)$, where $T_{obs}$ is the observed value of the test statistic.

Thus, using (3) and definition of $p'$, as $Z \rightarrow \infty$, $p' \rightarrow 0$.

As described above, increasing the sample size leads to a significant increase in the value of test statistics, resulting in a very low $p$-value. This is considered highly significant in statistical terms. However, it's important to note that simply increasing the sample size does not guarantee clinical relevance. In fact, using an infinitely large sample can lead to significant results even if there is no real clinical difference, as is often the case with studies that use large databases. Therefore, it is important to reconsider the use of $p$-values when analyzing large databases to ensure that clinical relevance is accurately assessed. Thus, $p$-values alone cannot provide reliable results when sample size becomes extremely large. In addition to the use of the Wald test statistic, multiple comparisons and multiplicity adjustment are discussed in the next section.

## 2.2. Multiple comparisons and multiplicity adjustment

In exploratory analyses on large datasets, many hypotheses are evaluated. Sometimes when an experiment is conducted to answer a research question, multiple hypotheses may need to be tested, thereby requiring multiple comparisons to be performed. If all comparisons are simultaneously performed with an error rate of 0.05, the actual error rate gets inflated to a quantity equal to 0.05 times the number of hypothesis tests. This would reduce the reliability of the results and hence, we require an appropriate statistical inferential procedure to handle such a situation. Therefore, multiple comparison adjustments have been suggested in the literature which help in maintaining the allowable error rate at 5%. Consider a family of $k$ independent null hypotheses being tested at level $\alpha$. In this case, the family wise error rate (FWER) described by Ranstam *et al.* (2012) would be $1\text{-}(1\text{-}\ \alpha)^k$. Some commonly used multiplicity adjustment techniques are Bonferroni test, Tukey test, and Scheffé test as shown by Lee and Lee (2018). The Bonferroni method offers a higher level of rigor compared to the Tukey test, which is more permissive toward Type I errors. It also provides more leniency compared to the highly conservative Scheffé's method as indicated by Lee and Lee (2018). For a detailed description of the Bonferroni method and its application in this study, please refer to the next section.

### 2.2.1. Bonferroni test

When working with a family of hypotheses and their corresponding $p$-values, the Bonferroni correction can be used to control the FWER. The FWER is the probability of incorrectly rejecting at least one true null hypothesis$(H_i)$. The Bonferroni correction involves rejecting the null hypothesis for each $p$-value that is less than or equal to alpha divided by the total number of hypotheses as described by Lee and Lee (2018). This approach effectively controls the FWER at a level of $\alpha$. Boole's inequality as shown by Khrennikov (2008) provides proof that this control is achieved as follows:

$$FWER = P\left\{\cup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq \alpha \tag{3}$$

where $m$ = total number of null hypotheses, $H_1, \ldots., H_m$ are a family of hypotheses and $p_1, \ldots., p_m$ are corresponding $p$-values.

This control method is very versatile and flexible, though conservative, as it doesn’t rely on any assumptions about the relationships between $p$-values or how many of the null hypotheses are actually true.

## 2.3. Logistic regression

There are different types of logistic regression, including simple, ordinal, and multiple versions of logistic and ordinal regression as described by McNulty (2021). Simple logistic regression is used when the outcome variable is binary, while ordinal regression is used when the outcome variable has multiple ordered categories. Multiple versions of logistic and ordinal regression are used depending on the complexity of the data and research question. When researchers conduct multiple statistical tests within these regression models, they may encounter multiple comparisons, which can lead to false positives. A logistic regression model, also known as the logit model, estimates the probability of occurrence of an event, such as treatment was beneficial or not, based on certain set of independent variables as described by McNulty (2021). In logistic regression, a logit transformation is performed on the odds, *i.e.*, the probability of success divided by the probability of failure. The logistic function is written as follows

$$logit\ (p_i) = \frac{1}{1+e^{-p_i}}\,. \tag{4}$$

The logistic regression model is written as follows.

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_kX_k \tag{5}$$

Here, $logit\ (p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$ is the dependent variable and $\underline{X} = (X_1, X_2, \ldots, X_k)^T$ is the vector of independent variables. Generally, the maximum likelihood estimation (MLE) method is used to estimate the beta coefficients of the logistic model.

## 2.4. Effect size

Effect sizes represent quantitative measures of the relationships between variables. While the term "effect size" has historically been associated with various specific measures, it is now commonly used to denote any index indicating the relationship between variables.

Effect sizes serve as a means to convey the magnitude of the relationship observed between variables in a scientific study performed by Hedges *et al.* (2008). Effect size helps in quantifying the difference between the comparison groups as described by Grissom and Kim (2005). It gives an idea about the actual difference between groups and does not directly depend on the sample size. We describe some of the common effect size measures in the following subsections.

### 2.4.1. Cohen's *d*

The effect size using Cohen's $d$ presented by Cohen (1988) and Cohen (1992) is calculated as follows:

$$d = \frac{\mu_1 - \mu_2}{s} \tag{6}$$

Here, $\mu_1$and $\mu_2$ are the means of the two comparison groups and *s* is the pooled standard deviation. Cohen's *d* is used for continuous outcomes and follows a general convention that *d* = 0.2 implies small effect, *d* = 0.5 implies medium effect and *d* = 0.8 implies large effect.

### 2.4.2. Glass's Δ

The effect size using Glass's Δ presented by Rosenthal *et al.* (1994) is calculated as follows.

$$\Delta = \frac{\mu_1 - \mu_2}{s_C} \tag{7}$$

Here, $\mu_1$and $\mu_2$ are the means of the two comparison groups and $s_c$ is the standard deviation of the control group. The same convention is followed for Glass's Δ effect size estimates as that for Cohen's *d* described above with respect to interpretation based on cutoff values.

### 2.4.3. Cohen's *h*

In case of categorical outcomes, Cohen's *d,* or Glass's Δ cannot be used. In such cases, difference between proportions is tested instead of means presented by Cohen (1988) and Rosenthal *et al.* (1994).

Suppose $p_1$ and $p_2$ represent two proportions. Cohen's $h$ effect size measure is represented by

$$h = \varphi_1 - \varphi_2 \tag{8}$$

$$\text{where } \varphi_i = 2\arcsin\left(\sqrt{p_i}\right) \tag{9}.$$

The same convention is followed for Cohen's *h* effect size estimates as that for Cohen's *d* described above with respect to interpretation based on cutoff values as shown by Cohen (1988), Cohen (1992), Hedges *et al.* (2008) and Grissom and Kim (2005).

### 2.4.4. Odds ratio (OR)

OR is used to assess degree of association between binary outcomes and is interpreted as follows as reported by Chinn (2000). *OR* = 1.5 indicates weak association, *OR* = 2 indicates medium association and *OR* = 3 indicates strong association.

Consider the following 2x2 table.

**Table 1: 2×2 Contingency table**

| | **Event** | |
|---|---|---|
| **Exposure** | **Yes** | **No** |
| **Yes** | *a* | *b* |
| **No** | *c* | *d* |

$$Odds\ ratio\ (OR) = \frac{odds\ of\ the\ event\ in\ the\ exposed\ group}{odds\ of\ the\ event\ in\ the\ non{-}exposed\ group} \tag{10}$$

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} \tag{11}$$

### 2.5. Machine learning techniques

Multiple Linear Regression (MLR) is a powerful statistical technique commonly used in large data set analyses. MLR aims to model the relationship between a dependent variable and multiple independent variables by estimating the best-fitting regression equation as described by Ayyadevara (2018). In scenarios where data sets are large and complex, MLR serves as a valuable tool to identify and quantify the effects of multiple predictors on the outcome of interest. By incorporating multiple independent variables simultaneously, MLR allows researchers to understand the collective influence of various factors on the dependent variable, enabling them to uncover complex patterns and associations within the data. Although there are multiple MLR models, the choice of MLR model depends on the nature of the data, the research question, and the assumptions underlying the analysis as described by Ayyadevara (2018). In this work, we use ARM and the Naïve Bayes Classifier within Multiple Linear Regression (MLR) to provide valuable insights and enhance the interpretability and statistical analysis of big datasets.

### 2.5.1. Association rule mining (ARM)

To discover interesting relations between variables in large databases, ARM can be used as denoted by Ayyadevara (2018). ARM is a rule-based machine learning approach. The main concept in ARM is to discover rules that govern how certain sets of variables relate to each other. To find the degree of these relations, different measures such as lift (*L*), support (*S*) and confidence (*C*) can be used as described below.

In order to distinguish a trivial rule from a non-trivial rule, a measure used in ARM called the lift (*L*) can be calculated as follows as denoted by Geurts *et al.* (2003).

$$L = \frac{s(X \Rightarrow Y)}{s(X) \cdot s(Y)} \tag{12}$$

*X* is known as the antecedent of the rule and *Y* is known as the consequent. The numerator *s(X⇒Y)* measures the observed frequency of the items in *X* and *Y* occurring together and the denominator *s(X)* • *s(Y)* measures the expected frequency of the items in *X* and *Y* occurring together under the assumption of conditional independence as denoted by Geurts *et al.* (2003).

If *L* has a value greater than 1, we conclude that there is positive interdependence between *X* and *Y*. If the value of *L* is less than 1, we conclude that there is negative interdependence between *X* and *Y*. Lastly, if $L = 1$, *X* and *Y* are said to be conditionally independent. The greater the value of lift *L*, the stronger is the dependence between *X* and *Y*.

Two other important parameters for the ARM are the support (*S*) and confidence (*C*) of a rule by means of which the algorithm to produce a set of rules describing the underlying patterns in the data. Support of a rule indicates the frequency with which a rule occurs in a dataset and confidence measures the reliability of an association rule as indicated by Geurts *et al.* (2003). Suppose we are studying the association of different predictor variables with different surgery types for breast cancer.

**Table 2: Interpretation of lift values**

| Outcome | Interpretation of lift (L) |
|---|---|
| L < 1 | Negative interdependence between X and Y |
| L = 1 | Conditional independence between X and Y |
| L > 1 | Positive interdependence between X and Y |

Then,

$$S\,\{X\} = \frac{\textit{number of patients receiving surgery type X}}{\textit{total number of patients}} \text{ for a rule } \{X \Rightarrow Y\} \tag{13}$$

$$C\,\{X \geq Y\} = \frac{\textit{number of patients receiving surgery type X in predictor variable category Y}}{\textit{total number of patients receiving surgery type X}} \tag{14}$$

### 2.5.2. Naïve Bayes classifier

Naïve Bayes classifier is a machine learning algorithm based on Bayes' theorem that follows a probabilistic approach for solving classification problems. In real-world scenarios, variables have some correlations and are not entirely independent. However, the algorithm is called 'Naïve' Bayes classifier because it assumes independence between predictor variables as described by Zhang (2016) and Ayyadevara (2018).

The equation for Bayes' theorem is given as

$$P(A \mid B) = \frac{P(B \mid A) * P\,(A)}{P(B)} \tag{15}$$

Here, *P(A | B)*: Conditional probability of an event *A*, given the event *B*,
*P(A)*: Probability of event *A*
*P(B)*: Probability of event B
*P(B|A)*: Conditional probability of an event *B*, given the event *A*

The equation (15) represents a case with a single predictor. However, in real-world scenarios, there are more than one predictor variables and for a classification problem, there are multiple output classes. Let us represent these classes as $C_1, C_2, \ldots, C_k$ and the predictor variables as $x_1, x_2, \ldots, x_n$.

The objective of a Naïve Bayes algorithm is to estimate the conditional probability that an event with a feature vector $x_1, x_2, \ldots, x_n$ belongs to a particular class $C_i$.

Given these conditions, the equation for Naïve Bayes' classifier can be written as follows.

$$P(C_i \mid x_1, x_2, \ldots, x_n) \propto \left(\prod_{j=1}^{n} P(x_j \mid C_i)\right) \cdot P(C_i) \text{ for } 1 < i < k \tag{16}$$

Two statistical measures, namely, misclassification and accuracy, can be calculated for the Naïve Bayes classifier, based on which the model performance can be evaluated. *Misclassification* is the percentage of times a classifier incorrectly classifies an item into a class or category. *Accuracy* is the percentage of times a classifier correctly classifies an item into a class or category.

Other classification techniques include the Random Forest method that combines multiple decision trees to improve accuracy and reduce overfitting as explained by Ayyadevara 2018. The Neural Networks classification approach uses deep learning models with multiple layers of interconnected nodes that could be used for complex classification tasks. Another classification technique is the K-Nearest Neighbors (KNN), a non-parametric method that assigns class labels based on the majority class of the k-nearest data points. While other classification methods like Random Forest, Neural Networks, and K-Nearest Neighbors also offer their respective advantages, we have chosen the Naïve Bayes Classifier for this study as shown by Zhang (2016) and Ayyadevara (2018). The decision to use this method is based on factors such as interpretability, computational efficiency, and the specific characteristics of the NCDB dataset and the research question.

## 3. Modification/Extension of Cohen's *h* for logistic regression

Consider the following notations for effect size calculation.

**Table 3: Variables and notations for effect size calculations**

| | | **Outcome variable** | | | |
|---|---|---|---|---|---|
| **Variable 1** | **Total** | **Category 1** | **Category 2** | **…** | **Category *n*'** |
| **Category 1** | $n_1$ | $n_{11}$ | $n_{12}$ | | $n_{1n'}$ |
| **Category 2** | $n_2$ | $n_{21}$ | $n_{22}$ | | $n_{2n'}$ |
| **…** | | | | | |
| **Category *n*** | $n_n$ | $n_{n1}$ | $n_{n2}$ | | $n_{nn'}$ |

**Notations**

- $n_{ij}$: Number of patients in category *i* of variable 1 and category *j* of *variable 2*; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n'$.
- $n_i$: Total number of patients in $i^{th}$ category of variable 1.
- $p_{ij}$: Prevalence for category *i* of variable 1 and category *j* of *variable 2*. We calculate $p_{ij}$ as
  $p_{ij} = n_{ij}/n_i$

Cohen's *h* effect size for the $i^{th}$ category will be given by

$$h_i = \varphi_{i1} - \varphi_{i2};\ i = 1, 2 \tag{17}$$

$$\varphi_{ij} = 2\, logit\left(\sqrt{p_{ij}}\right);\ i = 1, 2;\ j = 1, 2 \tag{18}$$

These are defined for comparing two categories within two variables.

Here, instead of the traditional *arcsin* transformation used for Cohen's $h$ shown by Catarino *et al.* (2011), we use *logit* transformation described by Collins *et al.* (1992). A *logit* transformation is more appropriate in the case of logistic regression.

For our case we will calculate $h_1$ and $h_2$ corresponding to the two groups that we are comparing using equation (17). Then the effect size $h$ is given by

$$h = h_1 - h_2 \tag{19}$$

Here, we are comparing the effect sizes for one category of predictor variable with another category of predictor variable based on different categories of outcome variable. Since we are comparing the two categories of a predictor variable, a difference of differences is proposed. This difference, $h$ defined using equations (17), (18) and (19) is the novel effect size measure which would help in determining the meaningfully significant differences.

The convention used for the interpretation of the effect sizes is described in the table below. For large sample sizes such as the NCDB database, effects show up quickly due to the large sample. Hence, the convention that we have suggested considers an effect of approximately 93% as a small effect, 99.3% as medium effect and 99.9% as large effect. We suggest using this convention owing to the large sample size and using the guidelines suggested by Cohen for determining small, medium and large effect sizes as detailed by Cohen (1988) and Cohen (1992).

Thus, in the case of modified Cohen's $h$ – 1.5: small effect, 2.5: medium effect, 3: large effect (Table 4).

In this paper, using a case study from the National Cancer Database (NCDB), we have presented how the proposed novel effect size measure above can be utilized to help in obtaining meaningfully significant results. In addition, we have also validated the novel effect size measure using machine learning techniques.

**Table 4: Convention for modified Cohen's *h***

| Relative size | Effect size | Difference between the comparison groups |
|---|---|---|
| | 0.0 | 50% |
| **Small** | 1.5 | 93.3% |
| **Medium** | 2.5 | 99.3% |
| **Large** | 3 | 99.9% |
| | 5.5 | 100% |

The next section describes a brief literature review which is followed by the case study which demonstrates the statistical issues in the analysis of large databases, particularly expanding on $p$-values, and illustrates the use of the novel effect size measure.

## 4. Literature review

In this section, we provide a short literature review that comprises of 15 research articles that were carefully selected using the flowchart presented in Figure 1. Our main objective was to gain insights into the prevailing statistical issues surrounding sample selection, missing data imputation techniques, commonly used statistical methods and inference measures used. Our search revealed that a total of 3,331 articles were published between 2004 and 2014 using the NCDB, out of which 257 were focused on female breast cancer. To maintain uniformity, since the case study presented in this paper is focused on the association of surgery types with different demographic predictor variables, we only included articles that dealt with 'mastectomy' and 'lumpectomy' surgeries. This led us to 22 articles, and after removing duplicates, we were left with 15 articles that were used for our literature review.

### 4.1. Article search protocol

In the current article, we have presented a case study to examine association of surgery types with different demographic predictor variables to demonstrate statistical issues while analyzing large databases using NCDB as an example. The three surgery types for this study included from the NCDB were ‘lumpectomy’, ‘mastectomy without reconstruction’ and ‘mastectomy with reconstruction’. Hence, we designed the literature to identify and demonstrate statistical issues in the analysis of large medical databases. We identified articles published using female breast cancer data from NCDB and we performed keyword search using PubMed, MEDLINE (Web of Science), and Embase databases. We used the following keywords to search relevant articles: ‘NCDB’, ‘National Cancer Database’, ‘Breast Cancer’, ‘surgery’, ‘mastectomy’, ‘lumpectomy’, and ‘female’ and narrowed down to 15 articles that were most relevant.

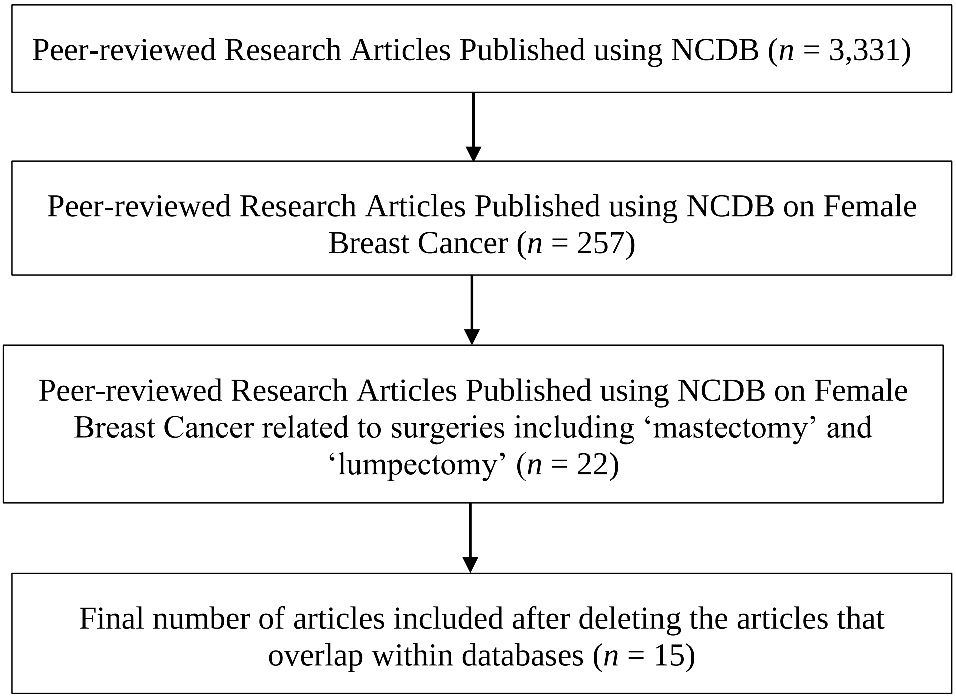

Peer-reviewed Research Articles Published using NCDB ($n$ = 3,331)
Peer-reviewed Research Articles Published using NCDB on Female Breast Cancer ($n$ = 257)
Peer-reviewed Research Articles Published using NCDB on Female Breast Cancer related to surgeries including ‘mastectomy’ and ‘lumpectomy’ ($n$ = 22)
Final number of articles included after deleting the articles that overlap within databases ($n$ = 15)

**Figure 1: Schematic representation for selection of articles**

### 4.2. Literature review results

Table 5 presents an overview of the research articles with respect to important statistical considerations.

**Table 5: General overview of research articles**

| **Article reference (Year)** | | **Details of the article** |
|---|---|---|
| Hotsinpiller *et al.* (2021) | Objective | Describe rates and predictors of positive margins for invasive breast cancers in the NCDB |
| | Sample Size | 707,798 |
| | Missing/Imputation | None |
| | Statistical Methods | Two-sided *t*-test; Chi-square test; Multivariable logistic regression |
| | Inference Measures | Odds ratios with 95% CI; *p*-values |
| Wrubel *et al.* (2021) | Objective | Compare BCT with mastectomy for treatment of early-stage breast cancer |
| | Sample Size | 202,236 |
| | Missing/Imputation | None |
| | Statistical Methods | Chi-square test; Kaplan-Meier analysis; Log-rank test |
| | Inference Measures | Kaplan-Meier survival curves; Overall survival (%); *p*-values |
| Weiser *et al.* (2021) | Objective | Identify sub-groups of node-positive patients with low to intermediate RS who still benefit from adjuvant chemotherapy |
| | Sample Size | 28,591 |
| | Missing/Imputation | None |
| | Statistical Methods | *t*-test; Chi-square test; Multivariable logistic regression; Kaplan-Meier method; Log-rank test; Multivariable Cox proportional hazards model |
| | Inference Measures | Hazard ratios with 95% CI; Odds ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| Lehrberg *et al.* (2021) | Objective | Evaluate the outcomes and predictors for patients receiving BCS treatment outside of the standard NCCN guidelines, compared with patients receiving standard MRM treatment |
| | Sample Size | 10,610 |
| | Missing/Imputation | None |
| | Statistical Methods | *t*-test; Chi-square test; Cochran-Armitage trend test; Multivariate Cox proportional hazards model |
| | Inference Measures | Adjusted hazards ratios; *p*-values |
| Pratt *et al.* (2021) | Objective | Examine the association between the time interval from time of diagnosis to completion of all acute breast cancer treatment modalities (surgery, chemotherapy, and radiation therapy) and survival |
| | Sample Size | 50,720 |
| | Missing/Imputation | None |

| **Article reference (Year)** | | **Details of the article** |
|---|---|---|
| | Statistical Methods | Univariate and multivariate Cox proportional hazards model; Log-rank test; Kaplan-Meier method; Chi-square test; Fisher's exact test; Two-sample *t*-test |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; 5-year survival (%); *p*-values |
| Lewis *et al.* (2019) | Objective | Determine the clinical characteristics, outcomes, and propensity for lymph node metastasis of patients with IMPC of the breast recorded in the NCDB |
| | Sample Size | 2660 |
| | Missing/Imputation | None |
| | Statistical Methods | Log-rank test; Cox proportional hazards model; Kaplan-Meier method |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| Mazor *et al.* (2019) | Objective | Assess patterns and outcomes of BCT for T3 tumors |
| | Sample Size | 37,268 |
| | Missing/Imputation | None |
| | Statistical Methods | Sensitivity analysis; Chi-square test; Wilcoxon rank sum test; Multivariable logistic regression; Cochran-Armitage trend test; Spearman's correlation; Kaplan-Meier method; Cox proportional hazards model |
| | Inference Measures | Odds ratios with 95% CI; Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| Zhu *et al.* (2019) | Objectives | Study clinicopathological features, treatment patterns and prognosis of SCC; Investigate whether SCC (*vs.* IEDC) is associated with poor clinicopathological characteristics, different treatment patterns and worse survival; Perform exploratory analysis of the benefits of systematics therapy for SCC patients |
| | Sample Size | 3,430 |
| | Missing/Imputation | None |
| | Statistical Methods | Chi-square test; Kaplan-Meier analysis; Cox regression model |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| Landercasper *et al.* (2019) | Objective | Determine if there were differences in the OS of matched breast cancer patients undergoing lumpectomy *vs.* mastectomy in the NCDB |
| | Sample Size | 845,136 |
| | Missing/Imputation | None |

| **Article reference (Year)** | | **Details of the article** |
|---|---|---|
| | Statistical Methods | Kaplan-Meier method; Propensity score matched analysis; Cox proportional hazards model; Subgroup analysis |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| McClelland *et al.* (2019) | Objective | To assess trends in patterns of care and clinical outcomes to manage localized breast angiosarcoma |
| | Sample Size | 826 |
| | Missing/Imputation | None |
| | Statistical Methods | Chi-square test; Cochran-Armitage trend test; Univariate and multivariate logistic regression analysis; Univariate and adjusted Cox models; Survival analysis |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; Forest plots; *p*-values |
| Mills *et al.* (2018) | Objective | Utilize data from NCDB to complete investigation of the prognostic importance of histology within TMBC |
| | Sample Size | 89,220 |
| | Missing/Imputation | None |
| | Statistical Methods | Kaplan-Meier method; Log-rank test; Multivariate Cox proportional hazards model |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |
| Chiba *et al.* (2017) | Objective | Evaluate national trends in NET use in relation to conduct of Z1031 trial and impact of NET on the rates of BCS |
| | Sample Size | 77,272 |
| | Missing/Imputation | None |
| | Statistical Methods | Cochran-Armitage trend test; Chi-square test; Two-sample *t*-test; Multivariable logistic regression |
| | Inference Measures | Odds ratios with 95% CI; *p*-values |
| Landercasper *et al.* (2017 | Objective | Investigate whether the receipt of NAC is associated with fewer reoperations |
| | Sample Size | 71,627 |
| | Missing/Imputation | None |
| | Statistical Methods | Cochran-Armitage trend test; Chi-square test; Multivariable logistic regression; Propensity score matching |
| | Inference Measures | Odds ratios with 95% CI; Forest plots; *p*-values |
| Rusthoven *et al.* (2016) | Objective | Evaluate the impact of PMRT and RNI for women with clinically node positive breast cancer treated with NAC |
| | Sample Size | 15,315 |
| | Missing/Imputation | None |

| Article reference (Year) | | Details of the article |
|---|---|---|
| | Statistical Methods | Kaplan-Meier method; Log-rank test; Multivariate Cox models; Propensity score matched analysis |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; Forest plot; *p*-values |
| Chen *et al.* (2015) | Objective | Compare efficacy of BCS with RT and mastectomy using NCDB. |
| | Sample Size | 160,880 |
| | Missing/Imputation | None |
| | Statistical Methods | Kaplan-Meier method; Cox regression; Propensity score analysis |
| | Inference Measures | Hazard ratios with 95% CI; Kaplan-Meier survival curves; *p*-values |

BCS= breast conserving surgery, RT= radiotherapy, CI= Confidence interval, PMRT= Postmastectomy radiotherapy, RNI= Regional nodal irradiation, NAC = Neoadjuvant chemotherapy, NET = Neoadjuvant endocrine therapy, IMPC= invasive micropapillary carcinoma, BCT= Breast conservation therapy, SCC= Squamous cell carcinoma, IDC= Infiltrating ductal carcinoma, RS= recurrence score assay, OS=overall survival, NCCN= National comprehensive cancer network, MRM= Modified radical mastectomy, TMBC= Triple negative breast cancer

### 4.3. Observations from the literature review

After conducting a comprehensive literature review, we found that there is a commonality between the sample size selection techniques, missing value imputation methods, statistical methods, and inference measures used in published research studies. Table 5 provides a helpful summary of these various approaches, which we organized according to the study objective and statistical considerations. Our analysis of this information yielded some interesting findings, which we will now discuss in more detail in next sections.

#### 4.3.1. Sample selection techniques

Based on our analysis, appropriate sample size selection is a crucial aspect of statistical research. We have found that most of the studies that we reviewed worked with large sample sizes. In fact, we observed that the largest sample size among the 15 studies was an impressive 707,798 as observed in Hotsinpiller *et al.* (2021) while the smallest sample size was just 826 as observed in McClelland *et al.* (2019).

Interestingly, we also found that none of the studies that we reviewed described any formal sample size calculation techniques used for sample selection. Instead, it appears that samples were selected primarily based on data availability and filtering based on the study objective. This means that convenience/purposive sampling was used, and analysis methods designed for randomized data were employed, which could lead to misleading results and conclusions.

#### 4.3.2. Missing data imputation

It is important to note that missing data can be a common occurrence, especially in large datasets like the NCDB. If missing data is simply deleted without any imputation, it can lead to a biased sample with biased results. Unfortunately, many of the studies such as Landercasper *et al.* (2017), Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021) included in the literature review did not describe or perform any missing data imputation. Instead, they chose to perform a complete case analysis. However, it is worth noting that some studies, like Rusthoven *et al.* (2016), conducted a sensitivity analysis before and after excluding unknown variable values and still obtained similar results. Others, like Chiba *et al.* (2017), reported missing values in their table for tumor characteristics but did not mention whether these values were imputed or deleted before analysis. Lastly, Landercasper *et al.* (2019) pointed out that all but one of the studies they cited did not include any missing data imputation.

### 4.3.3. Statistical analysis methods

The choice of analysis methods depends upon the study objective. However, every statistical technique involves certain assumptions and if these assumptions are not satisfied, the analysis may not result in reliable conclusions. This is a common issue with statistical analysis of the NCDB. For example, application of a common Cox proportional hazard model to non-randomized studies (case-control and databases) results in unreliable estimate of hazard ratio (relative risk) due to heterogeneity, time-varying exposure, corelated risk factors, and confounding *etc.* as presented by Moolgavkar *et al.* (2018).

From the studies included in the present literature review, we observed that the most common statistical analysis methods used were univariable as shown in Weiser *et al.* (2021) and multivariable logistic regression as used in Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), Hotsinpiller *et al.* (2021), and Weiser *et al.* (2021) and survival analysis as shown in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Baseline characteristics are commonly compared using either a *t*-test (continuous variables) as indicated in Chiba *et al.* (2017), Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et al.* (2021), and Weiser *et al.* (2021) or a chi-square test (categorical variables) as indicated in (Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen 2019, Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et* al. (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Propensity score matching methods shown in Chen *et al.* (2015), Rusthoven *et al.* (2016), Landercasper *et al.* (2017), and Landercasper *et al.* (2019) and the Cochran-Armitage trend test as shown in Chiba *et al.* (2017), Landercasper *et al.* (2017), Mazor *et al.* (2019), McClelland *et al.* (2019), and Lehrberg *et al.* (2021) are also popular techniques for analyzing NCDB.

### 4.3.4. Inference measures

Different statistical analysis methods involve different inference measures based on which we draw conclusions. The most common inference measures in the cancer studies are hazard ratio, odds ratio, Kaplan-Meier survival curve, and the ubiquitous *p*-value which is used for making conclusions in most of the analysis procedures.

In the articles included in the literature review as well, *p*-values were the most common statistical inference measure were used in Chen *et al.* (2015), Rusthoven *et al.* (2016), Chiba *et al.* (2017), Landercasper *et al.* (2017), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Hotsinpiller *et al.* (2021), Lehrberg *et al.* (2021), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021). Odds ratios was used in Chiba et al. (2017), Landercasper *et al.* (2019), Mazor *et al.* (2019), Hotsinpiller *et al.* (2021), and Weiser *et al.* (2021), hazard ratios was used in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.*

(2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Lehrberg *et al.* (2021), Pratt *et al.* (2021), and Weiser *et al.* (2021), and their 95% confidence intervals, Kaplan-Meier survival curves was shown in Chen *et al.* (2015), Rusthoven *et al.* (2016), Mills *et al.* 2018, Landercasper *et al.* (2019), Lewis *et al.* (2019), Mazor *et al.* (2019), McClelland *et al.* (2019), Zhu and Chen (2019), Pratt *et al.* (2021), Weiser *et al.* (2021), and Wrubel *et al.* (2021) were also a common choice. Some studies such as Rusthoven *et al.* (2016), Landercasper *et al.* (2017), and McClelland *et al.* (2019) used forest plots.

#### 4.3.5. Statistical issues with large databases – Summary

After conducting the literature review and carefully studying the published literature, we observed the following statistical issues with large databases.

Based on our comprehensive literature review and analysis of large databases, we have identified several design and statistical analysis issues that need to be addressed. Firstly, databases such as NCDB have extremely large sample sizes, which can create challenges in analyzing the data effectively. Secondly, data from large databases are not randomized, and therefore include close to all observed cases, making it difficult to control for bias. There is also a risk of duplicate records if a patient gets treated at multiple facilities and gets included in the databases that many times they were treated at different locations, which could significantly confound the results.

Another issue is that since databases such as NCDB are huge, the assumption of normality fails, which can impact the validity of parametric statistical analysis, most of which is built on the assumption of normality of data. Furthermore, due to extremely large sample size, analysis results in highly significant *p*-values, which may have no clinical relevance and could lead to false conclusions. Lastly, Simpson's paradox is a concern when using the entire database, as we may get highly significant results that might get reversed if we use a smaller sample selected using formal sample size selection techniques as reported by Hernán *et al.* (2011), and Pearl (2022).

## 5. Case study

### 5.1. Study design

Using the general study considerations outlined previously and based on the most common statistical analysis methods, we designed a case study using female breast cancer data from the NCDB collected between 2004 and 2014. This case study was designed and presented to demonstrate statistical issues related to analyzing large databases using NCDB as an example and suggesting alternative inference techniques that could describe real-world scenarios better than the current methods. Additionally, using the case study, we demonstrate the effectiveness of the novel modified Cohen's *h* effect size estimator. We also present category-wise comparisons which result in multiple *p*-values and show Bonferroni multiplicity adjustment fails to produce meaningful results as reported by Leon (2004). Note that the Bonferroni method is the most conservative method for adjusting for multiple comparisons.

The objective of the case study was to examine whether there is an association between surgery types and different demographic predictor variables. The dependent and independent variables and their levels/categories is given in Table 6.

The study used participants that satisfied certain inclusion and exclusion criteria based on the type of malignancy, diagnosis year, cancer stage, surgery type, *etc*. This outline has been explained in detail in Figure 2. After choosing participants, we investigated the data for missingness. In the current design, we deleted the missing observations as the proportion of missing values was low (approximately 1%). Details about the analysis have been discussed in section.

As described in Table 6, the dependent variable used in the case study was surgery type and the independent variables were age, race, insurance status, facility type, stage (of cancer) and great circle distance (distance between the medical facility and patient's residence).

Figure 2 presents the flowchart for selection of study participants. We included nine primary tumor sites as shown in Figure 2. The initial sample size was 2,445,870 for subjects who had tumors detected at the given primary sites. We, then, included patients who had a single malignant primary tumor, invasive or microinvasive breast cancer behavior, were diagnosed between 2004 and 2014, experienced breast cancer stages I, II or III and were female subjects. Later, we excluded the patients who did not undergo any surgery or had surgeries other than lumpectomy, mastectomy without reconstruction or mastectomy with reconstruction. The final sample size was 1,158,387 after applying all the inclusion and exclusion criteria. This is a large sample size that fits the definition of big data and hence, analysis using traditional approaches poses issues that need a robust solution. We will demonstrate the statistical issues and the effectiveness of our novel modified Cohen's *h* effect size estimator in the following sections.

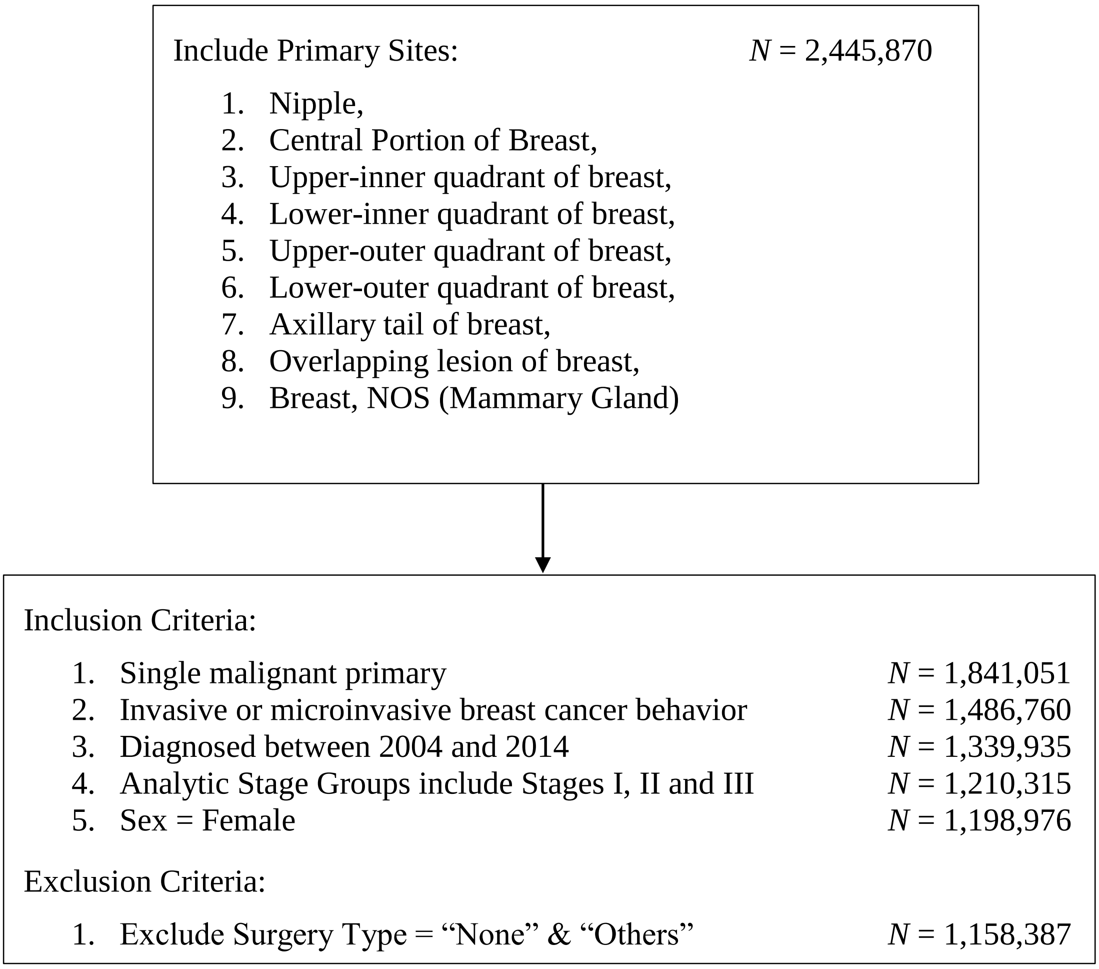
**Table 6: Dependent and independent variables for the case study**

| | **Variable** | **Levels/Categories** |
|---|---|---|
| **Dependent Variable** | Surgery Type | 1. Lumpectomy<br>2. Mastectomy without reconstruction<br>3. Mastectomy with reconstruction |
| **Independent Variables** | Age | 1. < 40 years<br>2. 40 – 50 years<br>3. 50 – 65 years<br>4. > 65 years |
| | Race | 1. White,<br>2. Black<br>3. Others |
| | Insurance status | 1. Medicare + Other Govt.<br>2. Private<br>3. Medicaid<br>4. Not insured |
| | Facility type | 1. Academic<br>2. Community Cancer Center<br>3. Comprehensive Community Center<br>4. Integrated Network |
| | Stage (of cancer) | 1. Stage I<br>2. Stage II<br>3. Stage III |
| | Great Circle Distance (Distance from the medical facility and patient's residence) | 1. < 50 miles<br>2. 50 – 150 miles<br>3. > 150 miles |

### 5.2. Analysis plan

The study sample was divided into three groups based on the surgery types, and baseline characteristics. Values were tabulated for each of the three groups with respect to the different categories of predictor variables (see Table 7). We compared the baseline characteristics of the study participants using chi-square test, and the results are presented in Table 7.

To examine the association between surgery types and different predictor variables with multiple levels/categories for each, we applied multinomial multivariable logistic regression. We presented results along with odds ratios, their 95% confidence intervals (CI) and *p*-values in Table 9. In addition, we calculated the effect sizes using modified Cohen's *h* and presented the results for comparison in Table 10. A schematic of the study design has been presented in Figure 3.



Include Primary Sites: *N* = 2,445,870

1. Nipple,
2. Central Portion of Breast,
3. Upper-inner quadrant of breast,
4. Lower-inner quadrant of breast,
5. Upper-outer quadrant of breast,
6. Lower-outer quadrant of breast,
7. Axillary tail of breast,
8. Overlapping lesion of breast,
9. Breast, NOS (Mammary Gland)

Inclusion Criteria:

1. Single malignant primary *N* = 1,841,051
2. Invasive or microinvasive breast cancer behavior *N* = 1,486,760
3. Diagnosed between 2004 and 2014 *N* = 1,339,935
4. Analytic Stage Groups include Stages I, II and III *N* = 1,210,315
5. Sex = Female *N* = 1,198,976

Exclusion Criteria:

1. Exclude Surgery Type = "None" & "Others" *N* = 1,158,387

**Figure 2: Study flowchart**

Furthermore, we used ARM and NBC to understand the underlying associations and identify clinically relevant variables. Both ARM and NBC were used to validate the results obtained using our novel effect size estimator and to identify the important variables. This was especially important given the large and complex datasets we were working with.

### 5.3. Methods

A detailed protocol stating the study objectives, methods, analysis plan *etc*. was submitted for IRB approval in order to gain access to the data from NCDB for the present research. The protocol was approved by the University of Louisville and James Graham Brown Cancer Center Cancer Committee which is a CoC Accredited Cancer Center. After receiving the IRB approval (IRB Number: 20.0049, we submitted access request to NCDB, which was also approved and access to NCDB was then granted to the authors of this manuscript. IRB approval request at the University of Cincinnati is pending. This study falls under IRB exemption since it is a retrospective study that looks at and utilizes NCDB data for analyses but does not involve working with human subjects.

## 5.4. Results

This section presents and describes the results obtained from analyzing our retrospective study. Table 7 shows the frequency distribution of breast cancer patients' three most common surgery types. For analyzing these data, we deleted the missing observations and compared Lumpectomy *vs.* Mastectomy without reconstruction and Mastectomy with reconstruction using the complete cases available. We present an overall *p*-value along with *p*-values for individual comparisons for each category of dependent variable. These *p*-values were then adjusted for multiple comparison using Bonferroni's test and the results are presented in Table 8. Majority of the studies present an overall *p*-value without looking at individual level comparisons. Since logistic regression compares difference between surgery types within categories of predictor variables, it is useful to present individual level *p*-values to help direct comparison of results.

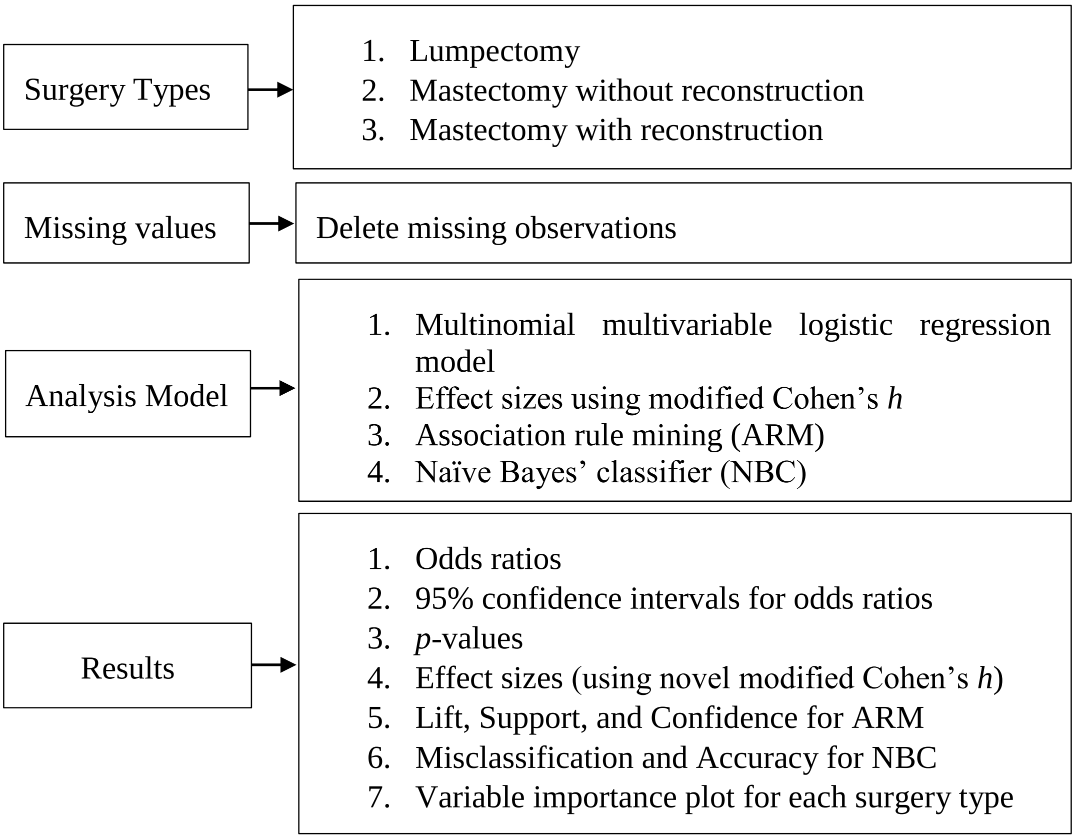

Surgery Types
1. Lumpectomy
2. Mastectomy without reconstruction
3. Mastectomy with reconstruction
Missing values
Delete missing observations
Analysis Model
1. Multinomial multivariable logistic regression model
2. Effect sizes using modified Cohen's *h*
3. Association rule mining (ARM)
4. Naïve Bayes' classifier (NBC)
Results
1. Odds ratios
2. 95% confidence intervals for odds ratios
3. *p*-values
4. Effect sizes (using novel modified Cohen's *h*)
5. Lift, Support, and Confidence for ARM
6. Misclassification and Accuracy for NBC
7. Variable importance plot for each surgery type

**Figure 3: Study design**

### 5.4.1. Baseline characteristics

Table 7 presents the baseline characteristics (lumpectomy (LT) *vs.* mastectomy without reconstruction (MTnR)/mastectomy with reconstruction (MTR)). It can be observed that *p*-values for all the variables shown in Table 7 are <$10^{-5}$ due to the large sample size and such a situation where almost all comparisons appear to be statistically significant, regardless of their practical importance or real-world significance. Therefore, we provide an alternative approach as presented in Table 8.

Table 8 presents the raw *p*-values in addition to Bonferroni-adjusted *p*-values to aid in comparing logistic regression results with baseline characteristics results.

**Table 7: Baseline characteristics (lumpectomy (LT) *vs.* mastectomy without reconstruction (MTnR)/mastectomy with reconstruction (MTR))**

| **Variable** | | **Surgery type** | | | ***p*-value** |
|---|---|---|---|---|---|
| | **Total (*N* = 1157322)** | **LT (*N* = 692564)** (59.84%) | **MTnR (*N* = 325104)** (28.09%) | **MTR (*N* = 139654)** (12.07%) | |
| | **Frequency (%)** | | | | |
| **Age** | | | | | < $10^{-5}$ |
| < 40 years | 65404 (5.65) | 23133 (3.34) | 21327 (6.56) | 20944 (15.00) | -- |
| 40 – 50 years | 217557 (18.8) | 110796 (15.99) | 57470 (17.68) | 49291 (35.30) | Ref |
| 50 – 65 years | 483166 (41.75) | 302250 (43.64) | 123114 (37.87) | 57802 (41.39) | 9.9x$10^{-5}$ |
| > 65 years | 391195 (33.80) | 256385 (37.02) | 123193 (37.89) | 11617 (8.32) | 9.9x$10^{-5}$ |
| **Race** | | | | | < $10^{-5}$ |
| White | 971210 (83.92) | 587089 (84.77) | 265520 (81.67) | 118601 (84.92) | Ref |
| Black | 124883 (10.79) | 71636 (10.34) | 40003 (12.30) | 13244 (9.48) | 9.9x$10^{-5}$ |
| Others | 48959 (4.23) | 26181 (3.78) | 16482 (5.07) | 6296 (4.51) | 9.9x$10^{-5}$ |
| **Insurance Status** | | | | | < $10^{-5}$ |
| Medicare + Other Govt | 399491 (34.52) | 256126 (36.98) | 127127 (39.10) | 16238 (11.63) | Ref |
| Private | 643302 (55.59) | 378917 (54.71) | 152246 (46.83) | 112139 (80.30) | 9.9x$10^{-5}$ |
| Medicaid | 70271 (6.07) | 35047 (5.06) | 27374 (8.42) | 7850 (5.62) | 9.9x$10^{-5}$ |
| Not Insured | 23876 (2.06) | 11782 (1.70) | 10169 (3.13) | 1925 (1.38) | 9.9x$10^{-5}$ |
| **Facility Type** | | | | | < $10^{-5}$ |
| Academic | 321773 (27.80) | 196015 (28.30) | 82863 (25.49) | 42895 (30.72) | Ref |

| Variable | | Surgery type | | | *p*-value |
|---|---|---|---|---|---|
| | **Total (*N* = 1157322)** | **LT (*N* = 692564)** (59.84%) | **MTnR (*N* = 325104)** (28.09%) | **MTR (*N* = 139654)** (12.07%) | |
| Community Cancer | 116406 (10.06) | 72115 (10.41) | 37621 (11.57) | 6670 (4.78) | 9.9x$10^{-5}$ |
| Comprehensive Community | 532529 (46.01) | 327628 (47.31) | 151580 (46.63) | 53321 (38.18) | 9.9x$10^{-5}$ |
| Integrated Network | 121210 (10.47) | 73673 (10.64) | 31713 (9.75) | 15824 (11.33) | 0.004 |
| **Analytic Stage** | | | | | < $10^{-5}$ |
| I | 620662 (53.63) | 456319 (65.89) | 101909 (31.35) | 62434 (44.71) | Ref |
| II | 399934 (34.56) | 204739 (29.56) | 137980 (42.44) | 57215 (40.97) | 9.9x$10^{-5}$ |
| III | 136726 (11.81) | 31506 (4.55) | 85215 (26.21) | 20005 (14.32) | 9.9x$10^{-5}$ |
| **Great Circle Distance** | | | | | < $10^{-5}$ |
| < 50 miles | 1069529 (92.41) | 646664 (93.37) | 296028 (91.06) | 126837 (90.82) | Ref |
| 50 – 150 miles | 61414 (5.31) | 31426 (4.54) | 20605 (6.34) | 9383 (6.72) | 9.9x$10^{-5}$ |
| > 150 miles | 16632 (1.44) | 8656 (1.25) | 5375 (1.65) | 2601 (1.86) | 9.9x$10^{-5}$ |

**Table 8: Raw and Bonferroni-adjusted *p*-values for baseline characteristics**

| Variable | Raw *p*-value | Bonferroni adjusted *p*-value |
|---|---|---|
| **Age** | | |
| 40 – 50 years | Ref | |
| 50 – 65 years | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| > 65 years | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| **Race** | | |
| White | Ref | |
| Black | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| Others | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| **Insurance Status** | | |
| Medicare + Other Govt | Ref | |
| Private | 9.9x$10^{-5}$ | 30x$10^{-5}$ |
| Medicaid | 9.9x$10^{-5}$ | 30x$10^{-5}$ |
| Not Insured | 9.9x$10^{-5}$ | 30x$10^{-5}$ |
| **Facility Type** | | |
| Academic | Ref | |
| Community Cancer | 9.9x$10^{-5}$ | 30x$10^{-5}$ |
| Comprehensive Community | 9.9x$10^{-5}$ | 30x$10^{-5}$ |
| Integrated Network | 0.004 | 0.012 |
| **Analytic Stage** | | |

| **Variable** | **Raw *p*-value** | **Bonferroni adjusted *p*-value** |
|---|---|---|
| I | Ref | |
| II | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| III | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| **Great Circle Distance** | | |
| < 50 miles | Ref | |
| 50 – 150 miles | 9.9x$10^{-5}$ | 20x$10^{-5}$ |
| > 150 miles | 9.9x$10^{-5}$ | 20x$10^{-5}$ |

The variables that have three categories and two comparisons (age, race, analytical stage, and great circle distance) will be tested at a significance level of 0.05/2 = 0.025. The variables that have four categories and three comparisons (insurance status and facility type) will be tested at a significance level of 0.05/3 = 0.017.

### 5.4.2. Logistic regression: lumpectomy *vs.* mastectomy without reconstruction and mastectomy with reconstruction

Table 9 presents the results of multinomial multivariable logistic regression using surgery types as the dependent variable and same predictor variables as those presented in the baseline characteristics table.

**Table 9: Multinomial multivariable logistic regression results**

| | **Multiple logistic regression** | | |
|---|---|---|---|
| **Predictors** | **Odds ratio** | **95% CI** | ***p*-value** |
| **LT *vs.* MTnR** | | | |
| **Age** | | | |
| 40 – 50 years (Ref) | 1 | NA | NA |
| 50 – 65 years | 0.85 | (0.85, 0.87) | $< 10^{-5}$ |
| > 65 years | 0.97 | (0.95, 0.99) | 0.0007 |
| **Race** | | | |
| White (Ref) | 1 | NA | NA |
| Black | 1.05 | (1.03, 1.07) | $< 10^{-5}$ |
| Others | 1.41 | (1.38, 1.45) | $< 10^{-5}$ |
| **Insurance status** | | | |
| Medicare + Other Govt (Ref) | 1 | NA | NA |
| Private | 0.75 | (0.74, 0.76) | $< 10^{-5}$ |
| Medicaid | 1.17 | (1.14, 1.19) | $< 10^{-5}$ |
| Not Insured | 1.31 | (1.27, 1.36) | $< 10^{-5}$ |
| **Facility Type** | | | |
| Academic (Ref) | 1 | NA | NA |
| Community Cancer | 1.29 | (1.27, 1.31) | $< 10^{-5}$ |
| Comprehensive Community | 1.19 | (1.17, 1.20) | $< 10^{-5}$ |
| Integrated Network | 1.09 | (1.07, 1.11) | $< 10^{-5}$ |
| **Stage** | | | |
| I (Ref) | 1 | NA | NA |
| II | 3.05 | (3.02, 3.08) | $< 10^{-5}$ |
| III | 12.37 | (12.18, 12.57) | $< 10^{-5}$ |
| **Great Circle Distance** | | | |
| < 50 miles (Ref) | 1 | NA | NA |

| | **Multiple logistic regression** | | |
|---|---|---|---|
| **Predictors** | **Odds ratio** | **95% CI** | ***p*-value** |
| 50 – 150 miles | 1.48 | (1.46, 1.52) | $< 10^{-5}$ |
| > 150 miles | 1.22 | (1.18, 1.28) | $< 10^{-5}$ |
| | **LT *vs.* MTR** | | |
| **Age** | | | |
| 40 – 50 years (Ref) | 1 | NA | NA |
| 50 – 65 years | 0.45 | (0.44, 0.46) | $< 10^{-5}$ |
| > 65 years | 0.15 | (0.14, 0.152) | $< 10^{-5}$ |
| **Race** | | | |
| White (Ref) | 1 | NA | NA |
| Black | 0.73 | (0.71, 0.75) | $< 10^{-5}$ |
| Others | 0.92 | (0.89, 0.95) | $< 10^{-5}$ |
| **Insurance** | | | |
| Medicare + Other Govt (Ref) | 1 | NA | NA |
| Private | 1.49 | (1.45, 1.52) | $< 10^{-5}$ |
| Medicaid | 0.94 | (0.90, 0.97) | 0.0005 |
| Not Insured | 0.67 | (0.63, 0.71) | $< 10^{-5}$ |
| **Facility Type** | | | |
| Academic (Ref) | NA | NA | NA |
| Community Cancer | 0.48 | (0.47, 0.49) | $< 10^{-5}$ |
| Comprehensive Community | 0.80 | (0.78, 0.81) | $< 10^{-5}$ |
| Integrated Network | 1.04 | (1.02, 1.07) | 0.0001 |
| **Stage** | | | |
| I | NA | NA | NA |
| II | 1.83 | (1.80, 1.85) | $< 10^{-5}$ |
| III | 4.07 | (3.98, 4.16) | $< 10^{-5}$ |
| **Great Circle Distance** | | | |
| < 50 miles (Ref) | NA | NA | NA |
| 50 - 150 miles | 1.43 | (1.39, 1.47) | $< 10^{-5}$ |
| > 150 miles | 1.43 | (1.36, 1.50) | $< 10^{-5}$ |

### 5.4.3. Effect size

Table 10 presents the effect sizes using modified Cohen's *h* estimator. These results will be compared with the *p*-values obtained using logistic regression to identify the important associated variables for surgery types that the two procedures predict.

### 5.4.4. Association rule mining

Table 11 presented below shows the results obtained using the ARM procedure. Here, we present lift, support, and confidence for each of the associations obtained using this procedure. The ARM procedure was run to check associations between the surgery types and all the predictor variables. Conclusions were mainly drawn using lift values and support and confidence values were presented to demonstrate the strength of the lift values.

**Table 10: Effect sizes using modified Cohen's *h***

| Variable | Effect size using modified Cohen's *h* procedure | |
|---|---|---|
| | LT *vs.* MTnR | LT *vs.* MTR |
| **Age** | | |
| 40 – 50 years | Ref | |
| 50 – 65 years | -0.91 | -1.92 |
| > 65 years | -0.69 | -4.01 |
| **Race** | | |
| White | Ref | |
| Black | 0.57 | 0.01 |
| Others | 0.96 | 0.58 |
| **Insurance Status** | | |
| Medicare + Other Govt | Ref | |
| Private | -0.23 | 2.48 |
| Medicaid | 1.53 | 2.40 |
| Not Insured | 1.81 | 1.97 |
| **Facility Type** | | |
| Academic | Ref | |
| Community Cancer | 0.41 | -1.28 |
| Comprehensive Community | 0.16 | -0.48 |
| Integrated Network | 0.04 | -0.02 |
| **Analytic Stage** | | |
| I | Ref | |
| II | 3.22 | 2.28 |
| III | 7.16 | 4.32 |
| **Great Circle Distance** | | |
| < 50 miles | Ref | |
| 50 – 150 miles | 1.09 | 1.06 |
| > 150 miles | 0.94 | 1.04 |

**Table 11: Association rule mining – testing association between surgery types and predictor variables lift (*L*), support (*S*), and confidence (*C*)**

| Association | *L* | *S* | *C* |
|---|---|---|---|
| **Surgery types and Age** | | | |
| Lumpectomy and age group '> 65 years' | 1.09 | 0.24 | 0.66 |
| Lumpectomy and age group '50 – 65 years' | 1.04 | 0.24 | 0.62 |
| Lumpectomy and age group '40 – 50 years' | 0.85 | 0.10 | 0.51 |
| **Surgery types and Race** | | | |
| Mastectomy with Reconstruction and White | 1.01 | 0.10 | 0.86 |
| Lumpectomy and White | 1.01 | 0.51 | 0.86 |
| Mastectomy without Reconstruction and White | 0.97 | 0.23 | 0.82 |
| **Surgery types and Insurance status** | | | |
| Mastectomy with Reconstruction and Private | 1.43 | 0.10 | 0.81 |
| Lumpectomy and Medicare + Other Govt. | 1.07 | 0.22 | 0.64 |
| Lumpectomy and Private | 0.98 | 0.33 | 0.59 |
| **Surgery types and Facility type** | | | |
| Lumpectomy and Comprehensive Community Cancer Program | 1.03 | 0.28 | 0.62 |
| Lumpectomy and Academic/Research Program | 1.02 | 0.17 | 0.61 |

| **Association** | ***L*** | ***S*** | ***C*** |
|---|---|---|---|
| Lumpectomy and Integrated Network Cancer Program | 1.02 | 0.06 | 0.10 |
| Lumpectomy and Community Cancer Program | 0.88 | 0.08 | 0.52 |
| **Surgery types and Stage of cancer** | | | |
| Mastectomy with Reconstruction and Stage III | 2.22 | 0.07 | 0.62 |
| Lumpectomy with Stage I | 1.23 | 0.39 | 0.74 |
| Lumpectomy and Stage II | 0.86 | 0.18 | 0.51 |
| **Surgery types and Great circle distance** | | | |
| Lumpectomy and < 50 miles | 1.01 | 0.56 | 0.94 |
| Lumpectomy and > 150 miles | 0.87 | 0.01 | 0.52 |
| Lumpectomy and 50 – 150 miles | 0.86 | 0.03 | 0.51 |

### 5.4.5. Naïve Bayes classifier

To validate the results obtained using modified effect sizes, we used Naïve Bayes classifier. To identify important variables using Naïve Bayes classifier, the measures used were misclassification and accuracy as described in Section 2.5.2. The variable with the least misclassification and highest accuracy was concluded to have highest association with the dependent variable, surgery type. Furthermore, for the Naïve Bayes classifier, we plotted a variable importance plot by surgery type. The variable importance plot is presented in Figure 3. The variables highly associated with each surgery type is presented in the plot in the order of the strength of association. Each variable has an associated bar indicating the magnitude of its importance.

**Table 12: Naïve Bayes classifier – misclassification percentages along with accuracy**

| **Association of surgery types with** | **Misclassification** | **Accuracy** |
|---|---|---|
| Age | 40.3% | 59.7% |
| Race | 40.9% | 59.1% |
| Insurance Status | 40.3% | 59.7% |
| Facility Type | 40.04% | 59.96% |
| Stage | 35.4% | 64.6% |
| Great Circle Distance | 41.3% | 58.7% |

## 5.5. Interpretation

### 5.5.1. Baseline characteristics

Our selected study population had approximately 60% of the patients who received lumpectomy, 28% received mastectomy without reconstruction, 12% received mastectomy with reconstruction, and for around 1% of the subjects’ surgery information was missing.

Table 7 shows that the age group of 50 – 65 years receives the maximum proportion of breast cancer surgery. Among this group, lumpectomy is the most common surgery type, followed by mastectomy with reconstruction. The study also revealed that 84% of the patients were white, and 11% were black. The most common surgery type for white patients was lumpectomy and mastectomy without reconstruction, and for black patients, it was also lumpectomy and mastectomy without reconstruction. Private insurance was the preferred choice of the majority of the patients, and within this group, mastectomy with reconstruction was the most common surgery type. The study also found that about 46% of the patients received treatment at a comprehensive community center, and the maximum proportion of patients received lumpectomy. Among the patients with Stage I and Stage II breast cancer, 74% received lumpectomy and 51%, respectively, while most of the patients with Stage III received mastectomy without reconstruction (62%). The study also revealed that 92% of the patients lived within 50 miles of the facility where they received treatment, and within this group, 60% received lumpectomy followed by mastectomy without reconstruction (28%).

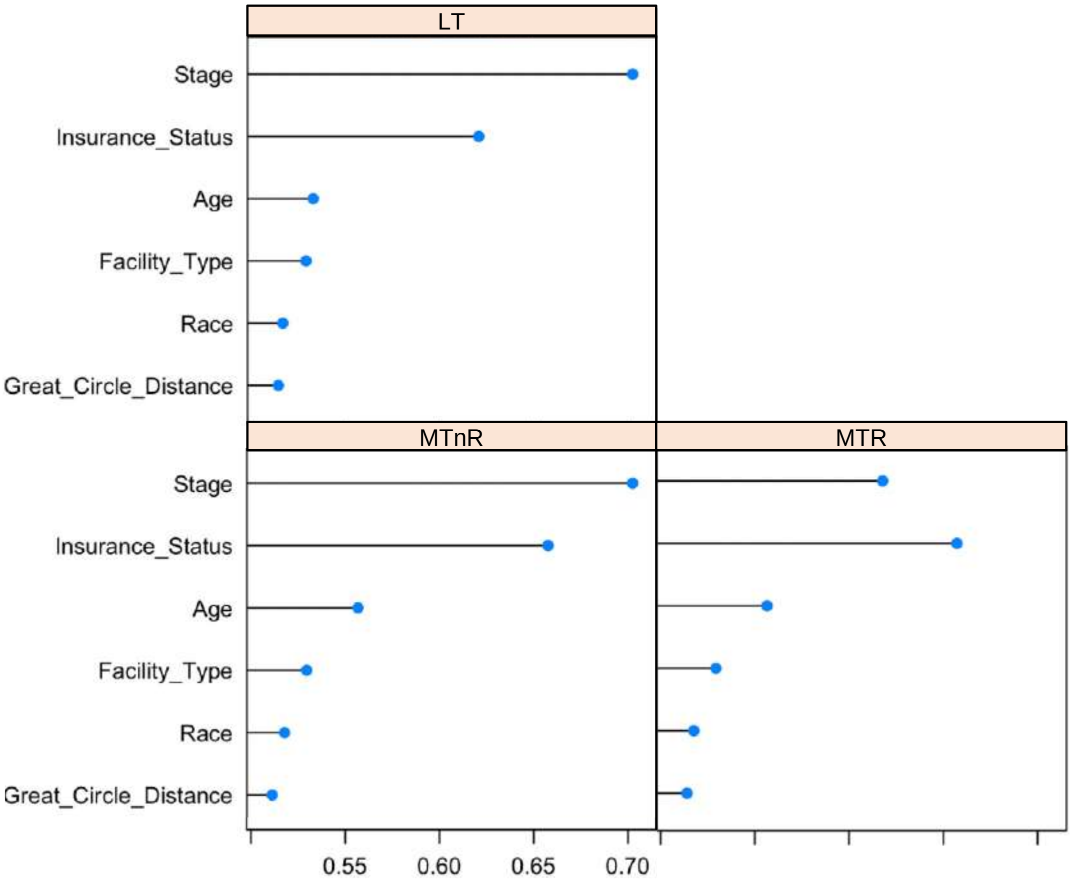


**Figure 4: Variable importance plot**

The chi-square tests run on the baseline characteristics reveal a highly significant association between the surgery types and all predictors, i.e., age, race, insurance status, facility type, analytic stage, and great circle distance. When we conducted pairwise comparisons using one of the categories as the reference category for each of the predictor variables, we obtained highly significant *p*-values for all the comparisons. Table 8 presents Bonferroni adjusted *p*-values in addition to the raw *p*-values. The Bonferroni adjusted *p*-values are highly significant for all variables as well. From tables 7 and 8, we conclude that surgery type has a highly significant association with all the predictor variables included in the study.

### 5.5.2. Logistic regression – *p*-values and odds ratios

It is interesting to note that Table 9 displays highly significant *p*-values ($< 10^{-5}$) for all the variables. However, the odds ratios tell a different story. According to Sullivan and Feinn (2012) an odds ratio around 1.5 indicates a small difference, around 2 indicates a medium difference, and around 3 indicates a significant difference. Except for the stage of cancer, none of the other variables have an odds ratio of >1.5, and thus, based on *p*-values, it can be concluded that stage is the only variable with a significant difference.

It should be noted that the statistical fallacy in this case is that the p-values are highly significant, with an odds ratio close to 1. This is demonstrated by the odds ratio of 0.97 and a *p*-value of 0.0007 for age group > 65 years *vs*. 40 – 50 years when comparing LT *vs*. MTnR. From the odds ratio, which is very close to 1, we would conclude that the age groups > 65 years and 40 – 50 years does not significantly differ from each other in terms of the surgery types. The 95% CI for this case is (0.95, 0.99) which almost cuts through the value of 1, thus hinting non-significant difference. This is a classic example of how *p*-values can be misleading, and concluding significant differences based on *p*-values alone would lead to misleading conclusions.

### 5.5.3. Effect sizes

In Table 10, the effect sizes calculated using modified Cohen's *h* formula are presented. To interpret the effect sizes, we use the cutoffs presented in Table 4 suggested by the authors of this paper. If the effect size is close to 1.5, it indicates a small effect, 2.5 indicates a medium effect, and 3 indicates a large effect. For a significant difference, we require a large effect size as it indicates a large underlying difference.

From Table 10, we can see that for LT *vs*. MTnR, we obtained an effect size of 3.22 for Stage II *vs*. Stage I and an effect size of 7.16 for Stage III *vs*. Stage I using the novel modified Cohen’s *h* effect size. This means that patients with Stage II breast cancer have a significantly higher chance of receiving MTnR than LT when compared with patients who have Stage I breast cancer. A similar interpretation applies to Stage III *vs*. Stage I. For all other comparisons, we do not obtain a high effect size and thus, we may conclude that except stage of cancer, the other predictors do not have a significant association with surgery type.

When comparing LT *vs*. MTR, a large effect size of -4.01 was obtained for age group > 65 years *vs*. 40 – 50 years. This indicates that the age group > 65 years has a significantly lower chance of receiving LT *vs*. MTR when compared with patients in the age group 40 – 50 years.

Similarly, we obtained a large effect size of 4.32 for comparing stage III *vs*. stage I breast cancer for LT *vs*. MTR. This also indicates that except age group and stage of cancer, none of the other predictors show a highly significant association with surgery types.

Based on odds ratios and effect sizes, we find only one significant association, *i.e*., between surgery types and stage of cancer. Even though the *p*-values are highly significant for all comparisons, odds ratios and effect sizes do not support this result. Thus, using effect sizes in addition to *p*-values when analyzing large datasets may be a more statistically sound approach.

### 5.5.4. Association rule mining

We observe from Table 11 that the values for lift are either very close to 1 or just below 1 for most associations except for the variable ‘stage’. Thus, we may say that the predictor variables age, race, insurance status, facility type, and great circle distance have a weak correlation with surgery types. There is, however, a strong association between stage of cancer and surgery types. From the lift value of 2.22, we may conclude that patients suffering from stage III breast cancer have a strong possibility of receiving mastectomy with reconstruction.

Also, patients suffering from stage I breast cancer have a slightly positive correlation with receiving lumpectomy which agrees with our conclusions from Table 7. The lift value for Mastectomy with reconstruction and private is 1.43 which is higher than 1. This implies a positive correlation between the two values with a confidence of 0.81.

### 5.5.5. Naïve Bayes classifier

Table 12 presents the results obtained using the Naïve Bayes classifier. The misclassification proportion is 35% for stage of cancer which is the lowest and the corresponding accuracy is 65% which is the highest. This indicates and supports all the previous results and arguments that the only significantly associated variable with surgery types is the stage of cancer.

From the variable importance plot, we see that stage is the most important variable associated with the surgery types LT and MTnR and the second most important variable associated for surgery type MTR.

From all the above results, we see that *p*-values indicate highly significant associations for surgery types with all the predictors. However, this statistical significance is not clinically relevant as indicated by the odds ratios that are close to 1. The modified effect sizes indicate a highly significant association between stage and surgery types and one of the age groups for LT *vs.* MTR comparison. The significant association between stage of cancer and surgery type is supported by ARM and NBC results as well.

## 6. Conclusion

From our analysis, we observed that *p*-values alone can lead to misleading conclusions since they are very sensitive to sample sizes. As sample size increases, *p*-values tend to decrease and produce highly significant but clinically irrelevant results. Thus, an alternative to *p*-values when analyzing extremely large datasets is crucial. For this purpose, we explored effect sizes. However, to the best of our knowledge, effect size measures have not been suggested for the case of a logistic regression when we are comparing effects of two treatments within two different categories of a variable. To handle such a situation, we suggested an extension to Cohen's *h* effect size measure and demonstrated its use with the help of a case study using NCDB as an example. We proved its utility using machine learning tools such as ARM and NBC. We suggest using this modified version of Cohen's *h* for large databases when using logistic regression and comparing multiple treatments across different categories of predictor variables.

## 7. Discussion

The study aimed to address critical gaps in the existing literature related to the analysis of large electronic health record (EHR) databases, sample selection methods for such databases, and the over-reliance on *p*-values for drawing clinical inferences. By doing so, it sought to provide valuable insights into the limitations of *p*-values in the context of large sample sizes and propose a novel effect size measure tailored for logistic regression. Furthermore, the study aimed to validate the effectiveness of the proposed effect size measure using machine learning techniques.

The first research question focused on the clinical relevance of *p*-values when dealing with sampling from large databases. It is well-known that large sample sizes can lead to statistically significant *p*-values, even when the observed effect sizes are trivial or lack clinical importance. The study's objective was to demonstrate this phenomenon when analyzing large medical databases, which is critical in guiding researchers to avoid misinterpreting significant *p*-values as clinically meaningful results. By highlighting the limitations of relying solely on *p*-values, the study encourages researchers to adopt a more comprehensive approach that includes effect sizes for a more nuanced interpretation of results. For large databases, in general, even if the large number of comparisons are subjected to the most stringent procedure of controlling Type I error *i.e.*, Bonferroni adjustment, let alone less stringent procedures like Holm, Hochberg, Hommel, and Benjamini-Hochberg, will result in highly significant *p*-values.

The second research question sought to explore the clinical relevance of effect sizes compared to *p*-values. Effect sizes provide a quantitative measure of the magnitude of an observed effect, indicating the practical importance of a finding. The study recognized the value of effect sizes in determining clinical significance, especially when dealing with large samples. By proposing a novel effect size measure specifically designed for logistic regression, the study aimed to overcome the limitations of *p*-values and offer a more meaningful and informative measure for interpreting results.

The study's primary objective was to propose and validate a novel effect size measure for logistic regression using machine learning techniques. Machine learning methods, specifically, association rule mining and Naïve Bayes classifier served as complementary tools to corroborate the findings obtained from the effect size measure. The validation process aimed to strengthen the credibility of the proposed measure and ensure its applicability in real-world scenarios.

The study's findings shed light on the importance of considering both *p*-values and effect sizes in data analysis. It emphasized that large sample sizes can lead to significant *p*-values without necessarily indicating clinical relevance. The proposed novel effect size measure offered a valuable alternative for assessing practical significance, particularly in logistic regression models. The validation through machine learning techniques provided additional support for the effectiveness and reliability of the novel effect size measure.

### 7.1. Baseline characteristics

The baseline characteristics of the selected study population provided valuable insights into the distribution of breast cancer surgery types and the patient demographics. The majority of patients (approximately 60%) underwent lumpectomy, followed by 28% who received mastectomy without reconstruction and 12% who underwent mastectomy with reconstruction. While the proportion of missing surgery information was minimal (around 1%), it is essential for future studies to address and minimize missing data to ensure the completeness and accuracy of the analysis.

One of the key findings of the study was the prominence of the age group between 50 and 65 years, as it received the highest proportion of breast cancer surgeries. Within this age group, lumpectomy was the most common surgery type, followed by mastectomy with reconstruction. This observation aligns with the current clinical guidelines, which often recommend lumpectomy as a preferred option for early-stage breast cancer in older patients due to its less invasive nature and potential for better cosmetic outcomes Pusic *et al.* (1999).

It is essential to recognize the limitations of the study, including potential selection bias and generalizability. The study population may not be fully representative of the broader breast cancer patient population, particularly in settings with different healthcare systems or demographics. Researchers should consider such factors when interpreting and applying the study's findings to diverse patient populations.

### 7.2. Logistics regression – odds ratio and *p*-value

The findings from the logistic regression analysis revealed highly significant *p*-values ($< 10^{-5}$) for all the variables under investigation. Such high significance levels might lead one to believe that all predictor variables have a strong impact on the outcome (surgery types). However, a closer examination of the odds ratios indicated that, except for the stage of cancer, none of the other variables demonstrated odds ratios greater than 1.5. This observation suggests that most of the predictor variables might not have a substantial effect on the choice of surgery types, except for the stage variable, which appears to be significantly associated with surgery types.

One notable concern arising from the analysis is the occurrence of a statistical fallacy when highly significant *p*-values are accompanied by odds ratios close to 1. This implies that, despite the statistical significance, the observed effect sizes might be minimal or practically negligible. In the context of the study, this phenomenon is particularly evident in the comparison between age groups and surgery types received. Although the *p*-values suggest a significant association, the odds ratios close to 1 indicate that age groups may not play a substantial role in determining the choice of surgery.

Relying solely on *p*-values to draw conclusions can be misleading, as highlighted by the findings. Focusing solely on the significance levels without considering the effect sizes might lead to incorrect interpretations of significant differences. It is essential to consider both the statistical significance and the practical significance (effect sizes) of the predictor variables to gain a comprehensive understanding of their impact on the outcome.

To avoid this statistical fallacy and ensure a more meaningful interpretation of the results, researchers should adopt a more holistic approach that considers both *p*-values and effect sizes. By considering the magnitude and direction of the effect sizes, researchers can better understand the clinical relevance of the predictor variables in relation to the outcome of interest.

Furthermore, the study underscores the importance of interpreting logistic regression results in the context of the research question and the clinical significance of the variables under investigation. While highly significant *p*-values are essential in identifying potential associations, they should not be the sole basis for decision-making or drawing conclusions. Instead, researchers should use them as a starting point to explore effect sizes and consider the practical implications of the findings.

### 7.3. Effect sizes

The analysis of effect sizes provided valuable insights into the magnitude and clinical relevance of the associations between predictor variables and surgery types in breast cancer patients. The effect sizes demonstrated that patients with Stage II and Stage III breast cancer had significantly higher chances of receiving mastectomy without reconstruction (MTnR) compared to lumpectomy (LT) when compared to patients with Stage I breast cancer. The large effect sizes of 3.22 and 7.16 for Stage II and Stage III, respectively, indicated a substantial impact of cancer stage on the choice of surgery type. These findings align with clinical practice, as more advanced stages of cancer often require more extensive surgical interventions.

However, apart from the stage of cancer, the analysis of effect sizes revealed that other predictor variables did not show a significant association with surgery types. Effect sizes not being high for these comparisons indicated that variables such as age group, race, insurance status, facility type, analytic stage, and great circle distance might not have a considerable impact on the choice of surgery type. It is crucial to consider these results when making treatment decisions and designing interventions, as they highlight the relative importance of different predictors in guiding surgical decisions for breast cancer patients.

An interesting finding emerged when comparing lumpectomy (LT) *vs.* mastectomy with reconstruction (MTR) within different age groups. The effect size of -4.01 for the age group > 65 years indicated a significantly lower chance of receiving lumpectomy compared to patients in the age group 40-50 years. This observation suggests that age plays a critical role in determining the choice of surgical treatment, and older patients are more likely to undergo mastectomy with reconstruction. These insights can help inform patient counseling and shared decision-making, enabling healthcare providers to better tailor treatment plans based on age-related preferences and concerns.

Moreover, the discussion emphasizes the added value of incorporating effect sizes alongside *p*-values in analyzing large datasets. While *p*-values indicate statistical significance, they might not fully convey the practical relevance of the findings. In contrast, effect sizes provide a quantitative measure of the strength of the associations, allowing researchers to assess the clinical significance of the predictor variables. The finding that only the stage of cancer showed significant associations based on both odds ratios and effect sizes suggests that effect sizes serve as a more robust tool for identifying clinically relevant relationships.

Using effect sizes in conjunction with *p*-values in the analysis of large datasets is suggested as a more statistically sound approach. By combining these measures, researchers can gain a more comprehensive understanding of the study results, identify meaningful associations, and avoid drawing conclusions based solely on statistical significance. This approach ensures that the reported findings have practical implications in clinical decision-making and can guide evidence-based practices.

The effect size measure proposed in this study was developed based on breast cancer data from the National Cancer Database (NCDB). To ensure its broader applicability, the robustness of this measure, along with the recommended conventions for interpretation, needs to be thoroughly investigated across various types of cancer and in diverse medical databases. Only through such comprehensive validation can the suggested novel effect size measure be established as a globally applicable and reliable metric for analyzing large medical datasets.

### 7.4. Naïve Bayes classifier (NBC)

The NBC results provide valuable insights into the relationship between predictor variables and surgery types in breast cancer patients. The findings show that the misclassification proportion for the stage of cancer is the lowest at 35%, and the corresponding accuracy is the highest at 65%. These results align with previous arguments that the stage of cancer is the most influential variable associated with surgery types, indicating that patients' cancer stage significantly impacts the choice of surgical treatment.

The variable importance plot further supports the importance of the stage of cancer in determining surgery types. The plot reveals that the stage variable is the most critical factor associated with lumpectomy (LT) and mastectomy without reconstruction (MTnR). Additionally, the stage variable is the second most important variable linked to mastectomy with reconstruction (MTR). These results emphasize the significance of cancer staging in guiding surgical decisions for breast cancer patients.

Interestingly, the *p*-values indicate highly significant associations for surgery types with all predictor variables. However, the odds ratios are close to 1, suggesting that the observed statistical significance might not translate into substantial clinical relevance. This discrepancy between statistical significance and practical significance can lead to the statistical fallacy discussed earlier, where highly significant *p*-values might not provide meaningful insights into the impact of predictor variables on surgery types. This highlights the importance of using effect sizes, as demonstrated in the modified effect sizes, to assess the clinical relevance of the associations.

The modified effect sizes demonstrate a highly significant association between the stage of cancer and surgery types, as well as one of the age groups in the comparison between LT and MTR. These effect sizes provide a more accurate and clinically meaningful measure of the associations, helping researchers understand the practical implications of the predictor variables in determining surgical choices for breast cancer patients.

The consistency of the significant association between the stage of cancer and surgery types across the Naïve Bayes Classifier, ARM, and other methods reinforces the robustness of the findings. These complementary techniques lend additional support to the conclusion that the stage of cancer is the most critical predictor variable influencing surgery types.

Overall, the study contributed to the growing body of literature on statistical analysis methods, offering insights into how to avoid misinterpretations and ensure more robust and clinically meaningful inferences. By addressing the gaps in the existing literature and proposing a novel effect size measure, this research provides valuable guidance to researchers, helping them make informed decisions and draw more accurate conclusions from large EHR databases.

## References

Ayyadevara, V. K. (2018). *Pro Machine Learning Algorithms: A Hands-On Approach To Implementing Algorithms In Python And R.* Apress.

Breckenridge, A. M., Breckenridge, R. A., and Peck, C. C. (2019). Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. *British Journal of Clinical Pharmacology*, **85**, 1874-1877.

Catarino, A., Churches, O., Baron-Cohen, S., Andrade, A., and Ring, H. (2011). Atypical EEG complexity in autism spectrum conditions: a multiscale entropy analysis. *Clinical Neurophysiology*, **122**, 2375-2383.

Chen, K., Liu, J., Zhu, L., Su, F., Song, E., and Jacobs, L. K. (2015). Comparative effectiveness study of breast-conserving surgery and mastectomy in the general population: a NCDB analysis. *Oncotarget*, **6**, 40127.

Chiba, A., Hoskin, T. L., Heins, C. N., Hunt, K. K., Habermann, E. B., and Boughey, J. C. (2017). Trends in neoadjuvant endocrine therapy use and impact on rates of breast conservation in hormone receptor-positive breast cancer: a national cancer data base study. *Annals of Surgical Oncology*, **24**, 418-424.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, **19**, 3127-3131.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates Publishers.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, **1**, 98-101.

Collins, J., Fanciulli, M., Hohlfeld, R., Finch, D., Sandri, G. V. H., and Shtatland, E. (1992). A random number generator based on the logit transform of the logistic variable. *Computers in Physics*, **6**, 630-632.

Geurts, K., Wets, G., Brijs, T., and Vanhoof , K. (2003). Profiling of high-frequency accident locations by use of association rules. *Transportation Research Record*, **1840**, 123-130.

Grissom, R. J. and Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach*. Lawrence Erlbaum Associates Publishers.

Hedges, S. B., Duellman, W. E., and Heinicke, M. P. (2008). New World direct-developing frogs (Anura: Terrarana): molecular phylogeny, classification, biogeography, and conservation. *Zootaxa*, **1737**, 181–182.

Hernán, M. A., Clayton, D., and Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, **40**, 780-785.

Hotsinpiller, W., Everett, A., Richman, J., Parker, C., and Boggs, D. (2021). Rates of margin positive resection with breast conservation for invasive breast cancer using the NCDB. *The Breast*, **60**, 86-89.

Khrennikov, A. (2008). Bell-Boole inequality: nonlocality or probabilistic incompatibility of random variables?. *Entropy*, **10**, 19-32.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *The Annals of Statistics*, **1**, 373-379.

Landercasper, J., Bennie, B., Parsons, B. M., Dietrich, L. L., Greenberg, C. C., Wilke, L. G., and Linebarger, J. H. (2017). Fewer reoperations after lumpectomy for breast cancer with neoadjuvant rather than adjuvant chemotherapy: a report from the national cancer database. *Annals of Surgical Oncology*, **24**, 1507-1515.

Landercasper, J., Ramirez, L. D., Borgert, A. J., Ahmad, H. F., Parsons, B. M., Dietrich, L. L., and Linebarger, J. H. (2019). A reappraisal of the comparative effectiveness of lumpectomy versus mastectomy on breast cancer survival: a propensity score–matched update from the National Cancer Data Base (NCDB). *Clinical Breast Cancer*, **19**, e481-e493.

Lee, S. and Lee, D. K. (2018). What is the proper way to apply the multiple comparison test?. *Korean Journal of Anesthesiology*, **71**, 353-360.

Lehrberg, A., Sebai, M., Finn, D., Lee, D., Karabon, P., Kiran, S., and Dekhne, N. (2021). Trends, survival outcomes, and predictors of nonadherence to mastectomy guidelines for nonmetastatic inflammatory breast cancer. *The Breast Journal*, **27**, 753-760.

Leon, A. C. (2004). Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. *Journal of Clinical Psychiatry*, **65**, 1511-1514.

Lewis, G. D., Xing, Y., Haque, W., Patel, T., Schwartz, M., Chen, A., Farach, A., Hatch, S. S., Butler, E. B., and Chang, J. (2019). Prognosis of lymphotropic invasive micropapillary breast carcinoma analyzed by using data from the National Cancer Database. *Cancer Communications*, **39**, 1-9.

Mazor, A. M., Mateo, A. M., Demora, L., Sigurdson, E. R., Handorf, E., Daly, J. M., A. A., Aggon, Anderson, P. R., Weiss, S. E., and Bleicher, R. J. (2019). Breast conservation versus mastectomy in patients with T3 breast cancers (> 5 cm): An analysis of 37,268 patients from the National Cancer Database. *Breast Cancer Research and Treatment,* **173**, 301-311.

McClelland, S., Hatfield, J., Degnin, C., Chen, Y., and Mitin, T. (2019). Extent of resection and role of adjuvant treatment in resected localized breast angiosarcoma. *Breast Cancer Research and Treatment*, **175**, 409-418.

McNulty, K. (2021). *Handbook of Regression Modeling in People Analytics: With Examples in R and Python*. CRC Press.

Mills, M. N., Yang, G. Q., Oliver, D. E., Liveringhouse, C. L., Ahmed, K. A., Orman, A. G., Laronga, C., Hoover, S. J., Khakpour, N., and Costa, R. L. (2018). Histologic heterogeneity of triple negative breast cancer: A National Cancer Centre Database analysis. *European Journal of Cancer*, **98**, 48-58.

Moolgavkar, S. H., Chang, E. T., Watson, H. N., and Lau, E. C. (2018). An Assessment of the Cox Proportional Hazards Regression Model for Epidemiologic Studies. *Risk Analysis*, **38**, 777-794.

Pearl, J. (2022). Comment: understanding Simpson's paradox. *Probabilistic and Causal Inference: The Works of Judea Pearl*. 399-412.

Pratt, D., Burneikis, T., Tu, C., and Grobmyer, S. (2021). Time to completion of breast cancer treatment and survival. *Annals of Surgical Oncology,* **28**, 8600-8608.

Pusic, A., Thompson, T. A., Kerrigan, C. L., Sargeant, R., Slezak, S., Chang, B. W., Helzlsouer, K. J., and Manson, P. N. (1999). Surgical options for early-stage breast cancer: factors associated with patient choice and postoperative quality of life. *Plastic and Reconstructive Surgery*, **104**, 1325-1333.

Ranstam, J. (2012). *Why The p-value Culture Is Bad And Confidence Intervals A Better Alternative*. Elsevier.

Rosenthal, R., Cooper, H., and Hedges, L. (1994). Parametric measures of effect size. *The Handbook of Research Synthesis,* **621**, 231-244.

Rusthoven, C., Rabinovitch, R., Jones, B., Koshy, M., Amini, A., Yeh, N., Jackson, M., and Fisher, C. (2016). The impact of postmastectomy and regional nodal radiation after neoadjuvant chemotherapy for clinically lymph node-positive breast cancer: a National Cancer Database (NCDB) analysis. *Annals of Oncology,* **27**, 818-827.

Solla, F., Tran, A., Bertoncelli, D., Musoff, C., and Bertoncelli, C. M. (2018). Why a *p*-value is not enough. *Clinical Spine Surgery*, **31**, 385-388.

Sullivan, G. M. and Feinn, R. (2012). Using effect size—or why the *p*-value is not enough. *Journal of Graduate Medical Education*, **4**, 279-282.

Wald, A. (1947). *Sequential Analysis*. J. Wiley & Sons.

Weiser, R., Haque, W., Polychronopoulou, E., Hatch, S. S., Kuo, Y.-f., Gradishar, W. J., and Klimberg, V. S. (2021). The 21-gene recurrence score in node-positive, hormone receptor-positive, HER2-negative breast cancer: a cautionary tale from an NCDB analysis. *Breast Cancer Research and Treatment*, **185**, 667-676.

Wrubel, E., Natwick, R., and Wright, G. P. (2021). Breast-conserving therapy is associated with improved survival compared with mastectomy for early-stage breast cancer: a propensity score matched comparison using the national cancer database. *Annals of Surgical Oncology*, **28**, 914-919.

Zhang, Z. (2016). Naïve Bayes classification in R. *Annals of Translational Medicine*, **4,** 241.

Zhu, L. and Chen, K. (2019). Clinicopathological features, treatment patterns, and prognosis of squamous cell carcinoma of the breast: an NCDB analysis. *BMC Cancer*, **19**, 1-9.

Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 205-215

# Investigating Regional Income-Based Shared Frailty among Women with Breast Cancer

**Vishal Deo[1,2], S. Pyne[2,3], Meghana Ray[2,4,5] and M. Vishnu Vardhan Rao[1]**
[1] *National Institute of Medical Statistics, ICMR, New Delhi, India*
[2] *Health Analytics Network, Pittsburgh, PA 15237, USA*
[3] *Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA*
[4] *HEED Lab, LLC, Rockville, MD 20852, USA*
[5] *Pramana Analytics LLP, New Delhi, India*

Received: 04 August 2023; Revised: 18 August 2023; Accepted: 21 August 2023

## Abstract

Breast cancer is the most commonly occurring cancer among women globally and also the most common cancer overall, with around 2.3 million new cases of breast cancer and around 685,000 deaths related to breast cancer reported globally in 2020. However, there are significant variations in the incidence and mortality rates across and within countries. Some studies have argued the role of neighbourhood socioeconomic status, family support mechanisms (marital status) and access to health care in explaining survival disparities.

The objective of this study is to identify and evaluate possible role of regional socioeconomic status in determining disparities in survival of women with breast cancer, while adjusting for other demographic and biomedical factors. Individual-level data of women in the US with breast cancer diagnosis between 1975 and 2019 was retrieved from the well-known Surveillance, Epidemiology, and End Results (SEER) Program submission 2021.

We hypothesized that individuals from counties in the US with similar socioeconomic status share unmeasured vulnerabilities towards survival from breast cancer. To validate this hypothesis, we have fitted semiparametric survival models with shared frailty defined for different categories of the county-level median household income.

Our modeling results show significant unmeasured heterogeneity between the clusters of individuals based on economic status of their counties. Individuals residing in counties with lower annual median household income share a higher risk of death due to breast cancer as compared to those from counties with higher median household incomes.

*Key words:* Breast cancer; Gamma frailty; Survival; Socioeconomic Determinants of Health; SEER; Healthcare disparity.

Corresponding Author: Vishal Deo
Email: vishaaldeo@gmail.com

## 1. Introduction

Breast cancer is the most commonly occurring cancer among women globally and also the most common cancer overall (WHO (2023a), WCRF (2023)). In the year 2020, approximately 2.3 million new cases of breast cancer were diagnosed, and around 685,000 deaths related to breast cancer were reported globally. There are significant variations in breast cancer incidence and mortality rates across different regions. As per the 2020 global data, breast cancer incidences were higher in developed countries whereas breast cancer deaths were higher in developing countries (WCRF (2023)). This may be due to systematic early detection programmes in developed countries, thereby underscoring the importance of economic and biomedical resources in ensuring adequate health programmes to facilitate awareness, early detection and treatment. Late diagnoses, inadequate health services and low universal health coverage are some of the important factors leading to global disparities in the outcomes of breast cancer (WHO (2023a)). Consider that breast cancer five-year survival rates are above 90% in high-income countries, as compared to 66% in India and 40% in South Africa (WHO (2023a)).

Centers of Disease Control and Prevention (CDC) have reported breast cancer as the second most common cancer among women (CDC (2023)). Breast cancer incidence is attributed to physical, hormonal, environmental, and genetic factors, including obesity, immunity, and the tumor environment. Interestingly, race, socioeconomic status and geography have also been found to determine patterns of breast cancer incidence. Incidence rates are the highest among non-Hispanic (NH) whites (130.8 per 100,000), followed closely by NH blacks (126.7 per 100,000). Yet, NH black women have the highest breast cancer death rate among all races in the US (28.4 deaths per 100,000). In fact, for every stage at diagnosis, NH black women have the lowest 5-year rate of survival (ACS (2019)).

Social Determinants of Health (SDH) have emerged in the recent years as a key issue alongside the traditional roles of genetic and demographic factors affecting survival of individuals with breast cancer. Disparities in access to and quality of healthcare could lead to disparities in health outcomes. WHO defines SDH as the "conditions in which people are born, grow, work, live and age, and the wider set of forces and systems shaping the conditions of daily life" (WHO (2023b)). The 2017 NIMHD Research Framework identified the domains of influence (Biological, Behavioral, Physical/Built Environment, Sociocultural Environment, Healthcare System) as well as different levels of influence (Individual, Interpersonal, Community, Societal) within those domains (NIMHD-NIH (2022)).

Among different SDH, geographical or physical/built environment health disparities are thought to be due – not only to limited physical access to health care – but also to differences in demography, attitudes, lifestyle factors, and cultural practices in regional and rural settings. A report by the US National Academies of Medicine stipulated that reducing geographical disparities in quality of care will benefit all its citizens but is likely to yield greater benefits to minority individuals (NRC (2004)). In the past few years, multilevel research on the local social context known as 'neighbourhood effects' and health led to findings about large racial/ethnic differences on mortality and morbidity (Chandra and Skinner (2004)). Neighbourhood as a SDH can be viewed through its components of the built environment, services, and the people within the neighbourhood (Bharmal *et al.* (2015)). Higher rates of obesity in neighbourhoods with poor walkability, access to healthy foods, health care facilities, are some of the mechanisms by which built environments can influence breast cancer outcomes (Obeng-Gyasi *et al.* (2022)). Neighbourhoods are in turn impacted by the socioeconomic conditions, *i.e.*, poorer neighbourhoods may have lower access to green spaces, healthy food markets, healthcare services, *etc.*, compared to richer neighbourhoods. Over time, such disparities may accrue and lead to unhealthy behaviours such as sedentary lifestyles, use of addictive substances, *etc.*, thereby perpetuating a cycle of poor health outcomes and worsening neighbourhood and socioeconomic conditions.

Disparities in breast cancer outcomes is currently an active area of research in epidemiology. A population-based cancer-specific survival study of patients diagnosed with breast, prostate, colorectal, or lung cancer between 2000 and 2013 in California, USA, by Ellis *et al.* (2018) estimated that the stage of diagnosis of cancer accounted for merely 24% of disparities in the survival of breast cancer patients. They found significant racial/ethnic survival disparities, with an overall reduction in survivability of black patients as compared to white patients. Their findings also suggested a significant role of neighbourhood socioeconomic status, family support mechanisms (marital status) and access to health care in explaining survival disparities. Hastert *et al.* (2021) used data from a regional cohort of African-American survivors of breast, colorectal, lung, and prostate cancer, to study the association between social needs of survivors and their health-related quality of life (HRQoL). They found a significant reduction in Functional Assessment of Cancer Therapy–General score, a measurement of HRQoL, for not getting care due to lack of transportation, for housing instability, for food insecurity, and for feeling unsafe in their neighbourhood.

In this direction, the present study aims to identify and evaluate the possible role(s) of regional socioeconomic status towards characterizing disparities in survival of women with breast cancer, while adjusting for other demographic and biomedical factors. Individual-level data of women in the US with breast cancer diagnosis between 1975 and 2019 was retrieved from the well-known Surveillance, Epidemiology, and End Results (SEER) Program submission 2021. In particular, we consider the location information of these individuals at the level of counties within the US states at which they reside. A county in the US provides a reasonably consistent environment for the local historic, geographic, and socioeconomic conditions that are commonly shared by the residents therein.

We hypothesized that individuals from counties with similar socioeconomic status share unmeasured vulnerabilities towards survival from breast cancer. To validate this hypothesis, we fitted semiparametric survival models with shared frailty defined for different categories of the county-level median household income. That is, individuals were clustered into different groups based on the median household income level of their county. Each of these clusters of individuals with breast cancer was expected to share some common unmeasured (or unaccounted for) risk of death due to breast cancer that was different from other clusters. This shared unmeasured risk could be due to various factors not directly included in the model, but those related to the local economic conditions of the individuals. For example, a county with lower median household income may imply less accessibility to healthcare services, higher risk of job loss, inadequate transportation, *etc.* Other relevant demographic and clinical factors available in the dataset, like age, cancer stage at diagnosis, breast cancer sub-type, and race are also included in the analysis. Posterior estimates of the random effects (frailty coefficients) corresponding to all clusters have been obtained. These estimates are a measure of unaccounted disparity between the clusters in the mortality risks from breast cancer.

The rest of the paper is organized in four sections. Description of the data and the shared frailty survival model has been provided in the next section on methodology. Section 3 contains descriptive summary of the data and the results of the fitted frailty survival models. A thorough discussion on the results is presented in section 4, and concluding remarks from the findings are provided in section 5.

## 2. Methodology

### 2.1. Data

SEER program of the National Cancer Institute (NCI) of the National Institutes of Health (NIH), USA, provides information on cancer statistics (SRP) [https://seer.cancer.gov/]. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 48 percent of the US population.

In this study, we have used the SEER research data on individuals who are women with breast cancer diagnosed between 1975 and 2019, based on the November 2021 submission of the SEER (SEER (2022)). The individual-level data is compiled from 8 cancer registries which are linked to county-specific attributes such as median household income, rurality, *etc.* The dataset covers a total of 465,908 individuals from eight US geographic areas, *viz.*, San Francisco-Oakland SMSA, Connecticut, Hawaii, Iowa, New Mexico, Seattle (Puget Sound), Utah, and Atlanta (Metropolitan). A list of variables used in the study and their description are provided in Table 1. Cases with complete information on these variables were included in the modeling exercise. A descriptive summary of the resulting dataset of 83,344 individuals is presented in Table 2.

### 2.2. Shared frailty survival model

The term *frailty* was used by Clayton (1978) to refer to any unobservable random effect shared by individuals with similar (unmeasured) risks in the analysis of mortality rates. Heterogeneity in unaccounted risks can be either defined for individuals experiencing recurrent events, or for clusters of individuals sharing common risks of an event (shared frailty). Shared frailty reflects excess risk for distinct groups of individuals sharing certain characteristics, over and above the risk explained by the measured covariates. In survival models, frailty is introduced as a random effect that acts multiplicatively on the hazard. The variance of the frailty measures the degree of heterogeneity in the hazard of different clusters of individuals- the case of shared frailty (Balan and Putter (2019)).

As a general structure for the shared frailty survival model, the conditional hazard function given the shared frailty can be written as

$$h_{ij}(t|Z_i) = h_0(t) Z_i exp(\beta^T x_{ij}(t)) \tag{1}$$

where, $Z_i$ is an unobserved random effect common to all observations from cluster $i$, *i.e.*, the shared frailty of cluster $i$, $\beta$ is a vector of unknown regression coefficients, $x_{ij}(t)$ are the observed covariates (which could also be time-dependent), and $h_0(t)$ is the baseline hazard function (Balan and Putter (2019), Hanagal (2019)). $Z_i$, being latent random terms having multiplicative effect on the hazard, are assumed to be iid random variables with a non-negative distribution, referred to as $Z$. We have used the R package "frailtyEM" by Balan and Putter (2019) to fit the model defined in (1). They have assumed that the distribution of the unobservable random variable $Z$ is defined by the Laplace transform

$$L_Z(c;\alpha,\gamma) \equiv E[exp(-Zc)] = exp(-\alpha\psi(c,\gamma)) \tag{2}$$

With $\alpha > 0$ and $\gamma > 0$, this formulation for the frailty distribution includes several distributions such as gamma, positive stable, inverse Gaussian, and compound Poisson, which belong to the Power Variance Function (PVF) family. Since these distributions are a part of the same family, likelihood values of models fitted with different frailty distributions are comparable.

## 3. Results

Semiparametric shared frailty survival models were fitted using the *emfrail* function of the R package frailtyEM. Shared frailty was defined with respect to clusters based on median household income of the counties. The frailty parameter was assumed to follow Gamma distribution. Models were fitted with various baseline hazard distributions such as Breslow, Exponential, Weibull, Lognormal; and the best model was chosen based on the highest log-likelihood value. The resulting model, adjusted for the factors – age category, race, breast cancer stage, and breast cancer subtype – had a Gamma frailty with Weibull baseline hazard. A summary of results of the selected model is shown in Table 3.

From Table 3, we observed that both the Commenges-Andersen test and the likelihood-ratio test conclude that the random effect (shared frailty) is highly significant. This is further validated by the non-zero variance of the frailty parameter, Var(Z), the confidence interval of which does not contain zero. This variation in frailty can be visualized from the histogram of the frailty estimates (Figure 1). Values of the posterior frailty estimates corresponding to each category of county-level median household income is provided in Table 4.

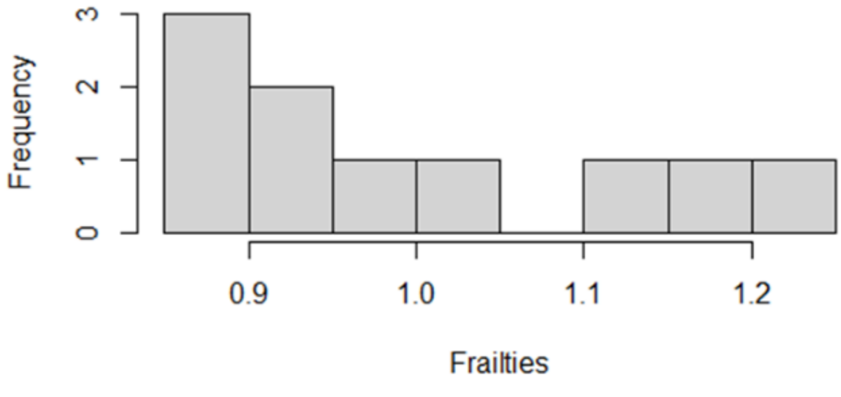


**Figure 1: Histogram of posterior frailty estimates of the selected model**

We can see from Table 4 that, in general, the frailty estimates decrease as the median household income level increases. That is, the additional unaccounted risk of an individual of dying due to breast cancer increases as the median household income level of their county decreases. The lower risk for the \$35,000 - \$39,999 income category could be a result of inappropriate representation of the cluster due to the relatively low number of individuals that belong to it (see Table 2).

The coefficient estimates of the selected survival model given in Table 3 provide a comprehensive insight into the impact of the factors on survival of women with breast cancer after adjusting for the unmeasured random effects (shared frailty) in the clusters. All the estimates were highly significant. Among the individuals with breast cancer, those belonging to white or other races are at significantly lower risk (around 30% lower) of dying due to breast cancer as compared to their black counterparts. As expected, risk of death increases by almost 10 times for individuals getting diagnosed in a late stage (stage III and higher) as compared to those diagnosed early. Among the four-breast cancer sub-types, risk of death is significantly higher for the HR−/HER2− category, also known as Triple-negative breast cancer (TNBC), as compared to each of the other three categories, namely Luminal A, Luminal B, and HER2. In other words, the risk of death associated with the other three sub-types is lower by around 54% – 62% than that for the TNBC sub-type. Compared to women with breast cancer who are lower than 50 years of age, the hazard of dying increases by 45% among those over 50 years of age.

## 4. Discussion

Factors considered in our study including age, breast cancer sub-type, race, and cancer stage, have been reported as significant risk factors by various previous studies (Ellis *et al.* (2018), Narod *et al.* (2018), Wadsten *et al.* (2017), CDC (2022)). Results from our modelling exercise reiterated the significant role of these risk factors in survival of women with breast cancer. In addition, as our results also show that black women with breast cancer are at significantly higher risk of death which has been previously reported by Ellis *et al.* (2018). This racial disparity can be partly because of genetic factors, and partly because of the socioeconomic disparities as well as disparities in accessibility to screening, treatment, and relevant resources across races.

The primary objective of this study was to investigate the possible role of unmeasured impact of regional socioeconomic status on survival of individuals with breast cancer in the US, while adjusting for other factors that have been previously shown to be associated with survival risk. Our findings validate the hypothesis that there is a significant unmeasured heterogeneity among clusters of individuals based on economic status of their respective counties. Individuals residing in counties with lower annual median household income share a higher risk of death due to breast cancer as compared to those from counties with higher median household incomes. The higher unmeasured risk for such individuals can be attributed to lack of adequate healthcare facilities, insecurity of jobs, less adherence to clinical follow-up due to potential loss of wages, lack of resources (which could include transportation, time and/or paid leave), poor diet and exercise, among other factors interdependent with the economic status of a county. This inference is concurrent with the findings of some previous studies that have explored the socioeconomic determinants of disparities in survival of individuals with cancer (Hastert *et al.* (2021), Merletti *et al.* (2011)).

## 5. Conclusion

Our findings indicate that economic vulnerabilities of women with breast cancer present added risks of mortality beyond the identified risks due to other common covariates. Socioeconomic and external environmental factors can play a significant role in cancer survival. Apart from efforts to improve healthcare practices such as early clinical interventions, detection programmes and screening, community education, *etc.*, there is a need to address the socioeconomic disparities to improve cancer survivability. Efforts in this direction begin with development of new policies aimed to reduce disparities and mitigate risks arising from the various SDH for individuals with breast cancer.

Our study has certain limitations. Due to constraints of data availability, we could not include additional factors such as treatment type or appropriateness, education level, *etc.* in the analysis. Additionally, since the data obtained pertains to only 8 registries associated with SEER, future studies could benefit from including additional registries to gain further insights into the effects of regional SDH on breast cancer outcomes.

## Acknowledgements

We are grateful to the editors and the reviewers for their guidance in improving this paper. Dr. Vishal Deo thanks the organizers of the Silver Jubilee Annual Conference of the SSCA, held in February 2023 at Jammu and Kashmir, India, for the opportunity to present this study.

## References

ACS (2019). Breast cancer facts and figures 2019-20. Atlanta: American Cancer Society.

Balan, T. A. and Putter, H. (2019). frailtyEM: An R package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, **90**, 1–29.

Bharmal, N., Derose, K. P., Felician, M., and Weden, M. M. (2015). Understanding the upstream social determinants of health. RAND Corporation.

CDC (2022). Disparities in breast cancer deaths. https://www.cdc.gov/cancer/dcpc/research/articles/disparities-breast-cancer-deaths.htm.

CDC (2023). Breast cancer statistics. https://www.cdc.gov/cancer/breast/statistics/index.htm.

Chandra, A. and Skinner, J. S. (2004). *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life.* Chapter Geography and Racial Health Disparities. The National Academies Press.

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

Ellis, L., Canchola, A. J., Spiegel, D., Ladabaum, U., Haile, R., and Gomez, S. L. (2018). Racial and ethnic disparities in cancer survival: The contribution of tumor, sociodemographic, institutional, and neighborhood characteristics. *Journal of Clinical Oncology*, **36**, 25–33.

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

Hastert, T. A., McDougall, J. A., Strayhorn, S. M., Nair, M., Beebe-Dimmer, J. L., and Schwartz, A. G. (2021). Social needs and health-related quality of life among african american cancer survivors: Results from the detroit research on cancer survivors study. *Cancer*, **127**, 467–475.

Merletti, F., Galassi, C., and Spadea, T. (2011). The socioeconomic determinants of cancer. *Environmental Health*, **10**, S7.

Narod, S. A., Giannakeas, V., and Sopik, V. (2018). Time to death in breast cancer patients as an indicator of treatment response. *Breast Cancer Research and Treatment*, **172**, 659–669.

NIMHD-NIH (2022). National institutes of minority health and disparities research framework. https://www.nimhd.nih.gov/about/overview/research-framework/nimhd-framework.html.

NRC (2004). *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life.* National Research Council. The National Academies Press. https://doi.org/10.17226/11086.

Obeng-Gyasi, S., Obeng-Gyasi, B., and Tarver, W. (2022). Breast cancer disparities and the impact of geography. *Surgical Oncology Clinics of North America*, **31**, 81–90.

SEER (2022). Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database: Incidence - seer research data, 8 registries, nov 2021 sub (1975-2019) - linked to county attributes – time dependent (1990-2019) income/rurality, 1969-2020 counties, national cancer institute, dccps, surveillance research program, released april 2022, based on the november 2021 submission.

Wadsten, C., Garmo, H., Fredriksson, I., Sund, M., and Wärnberg, F. (2017). Risk of death from breast cancer after treatment for ductal carcinoma in situ. *British Journal of Surgery*, **104**, 1506–1513.

WCRF (2023). Breast cancer statistics. https://www.wcrf.org/cancer-trends/breast-cancer-statistics/.

WHO (2023a). Breast cancer. https://www.who.int/news-room/fact-sheets.

WHO (2023b). Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1.

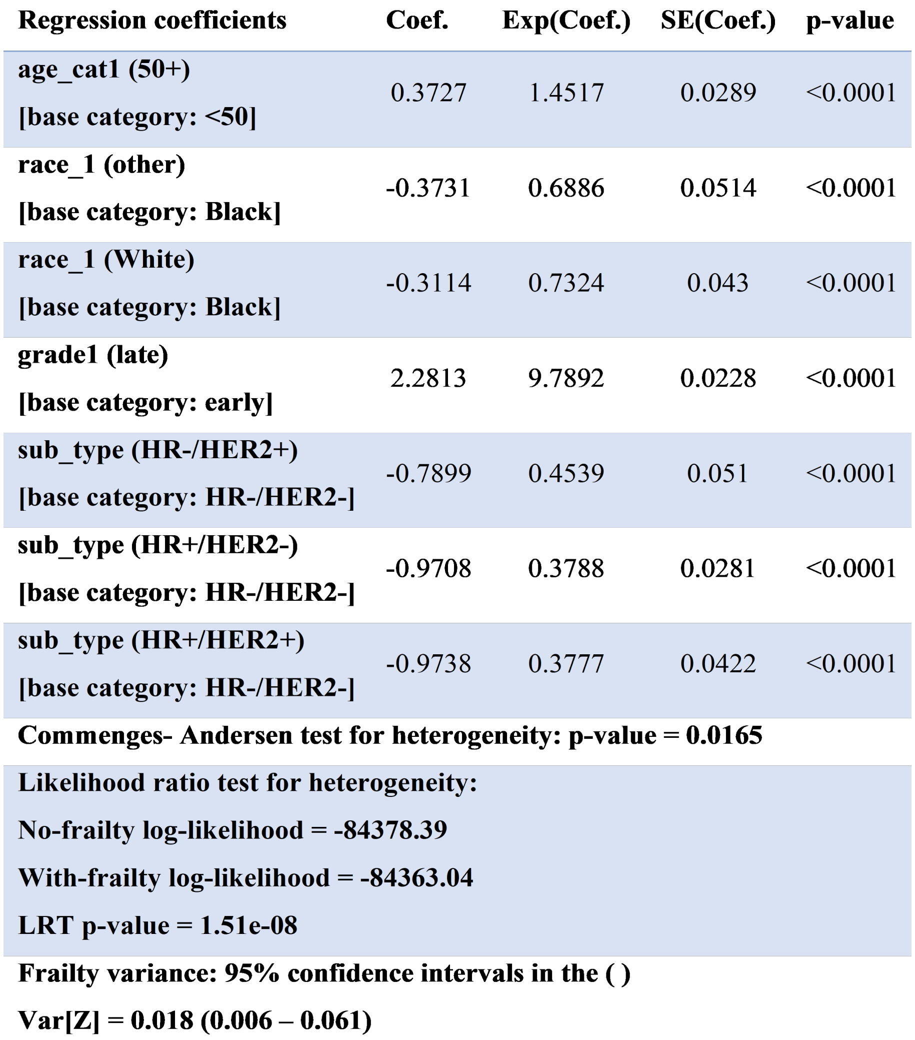
**Table 1: List of variables and their description**

| Name | Description | Categories |
|---|---|---|
| surv_months | Survival time in months (since diagnosis).<br>It is right censored for individuals who remained alive till the end of follow-up. | |
| event | Survival status of the individual | 1: Died<br>0: Alive (censored) |
| age | Baseline age of the individual<br>(15 years and above) | |
| age_cat1 | Factor variable derived from the baseline age. | less than 50<br>50 or more |
| sub_type | Breast cancer subtype.<br>For more information, see https://seer.cancer.gov/seerstat/databases/ssf/breast-subtype.html. | HR+/HER2− (Luminal A)<br>HR−/HER2− (TNBC)<br>HR+/HER2+ (Luminal B)<br>HR−/HER2+ (HER2) |
| race_1 | Race recode of individuals.<br>This recode is independent of Hispanic ethnicity of the individuals. | 1: White<br>2: Black<br>3: Other (American Indian, AK Native, Asian/Pacific Islander) |
| AJCC6_stage | AJCC 6TH STAGE classification of breast cancer. Refer http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th | 0, I, IIA, IIB, IIIA, IIIB, IIIC, IIINOS, IV |
| grade1 | A broader classification of stage of breast cancer of the individual at baseline, based on the AJCC6_stage variable. | early: 0, I, IIA, IIB, IIA<br>late: IIIA, IIIB, IIIC, IIINOS, IV |
| Median_HI | Median annual household income of the county where the individual resides. | less than $35,000<br>$35,000 - $39,999<br>$40,000 - $44,999<br>$45,000 - $49,999<br>$50,000 - $54,999<br>$55,000 - $59,999<br>$60,000 - $64,999<br>$65,000 - $69,999<br>$70,000 - $74,999<br>$75,000 and more |

Table 2: Factor-wise summary of the dataset from SEER

| Variable | Overall distribution/ summary [Total cases = 83344] | Survival status-wise distribution/ summary | |
|---|---|---|---|
| | | Died [8109] | Censored [75235] |
| **age_cat1** | <50: n = 16084 | <50: n = 1502 | <50: n = 14582 |
| | 50+: n =67260 | 50+: n = 6607 | 50+: n = 60653 |
| **race_1** | 1: White: n= 67123 | 1: White: n= 6450 | 1: White: n= 60673 |
| | 2: Black: n= 3972 | 2: Black: n= 616 | 2: Black: n= 3356 |
| | 3: Other: n =11901 | 3: Other: n = 1033 | 3: Other: n = 10868 |
| | 4: Unknown: n = 348 | 4: Unknown: n = 10 | 4: Unknown: n = 338 |
| **Median_HI** | < $35,000: n = 500 | < $35,000: n = 76 | < $35,000: n = 424 |
| | $35,000 - $39,999: n = 355 | $35,000 - $39,999: n = 38 | $35,000 - $39,999: n = 317 |
| | $40,000 - $44,999: n = 1524 | $40,000 - $44,999: n = 223 | $40,000 - $44,999: n = 1301 |
| | $45,000 - $49,999: n = 2266 | $45,000 - $49,999: n = 289 | $45,000 - $49,999: n = 1977 |
| | $50,000 - $54,999: n = 6703 | $50,000 - $54,999: n = 746 | $50,000 - $54,999: n = 5957 |
| | $55,000 - $59,999: n = 5590 | $55,000 - $59,999: n = 580 | $55,000 - $59,999: n = 5010 |
| | $60,000 - $64,999: n = 6943 | $60,000 - $64,999: n = 667 | $60,000 - $64,999: n = 6276 |
| | $65,000 - $69,999: n = 9919 | $65,000 - $69,999: n = 924 | $65,000 - $69,999: n = 8995 |
| | $70,000 - $74,999: n = 6949 | $70,000 - $74,999: n = 654 | $70,000 - $74,999: n = 6295 |
| | $75,000+: n = 42595 | $75,000+: n = 3912 | $75,000+: n = 38683 |
| **sub_type** | HR+/HER2-: n = 63379 | HR+/HER2-: n = 5030 | HR+/HER2-: n = 58349 |
| | HR-/HER2-: n = 8350 | HR-/HER2-: n = 1743 | HR-/HER2-: n = 6607 |
| | HR+/HER2+: n = 8123 | HR+/HER2+: n = 836 | HR+/HER2+: n = 7287 |
| | HR-/HER2+: n = 3492 | HR-/HER2+: n = 500 | HR-/HER2+: n = 2992 |
| **grade** | Early: n = 71433 | Early: n = 3580 | Early: n = 67853 |
| | Late: n = 11911 | Late: n = 4529 | Late: n = 7382 |

Table 3: Results of the fitted semiparametric shared frailty survival model with gamma frailty and Weibull baseline hazard



| Regression coefficients | Coef. | Exp(Coef.) | SE(Coef.) | p-value |
|---|---|---|---|---|
| **age_cat1 (50+)**<br>**[base category: <50]** | 0.3727 | 1.4517 | 0.0289 | <0.0001 |
| **race_1 (other)**<br>**[base category: Black]** | -0.3731 | 0.6886 | 0.0514 | <0.0001 |
| **race_1 (White)**<br>**[base category: Black]** | -0.3114 | 0.7324 | 0.043 | <0.0001 |
| **grade1 (late)**<br>**[base category: early]** | 2.2813 | 9.7892 | 0.0228 | <0.0001 |
| **sub_type (HR-/HER2+)**<br>**[base category: HR-/HER2-]** | -0.7899 | 0.4539 | 0.051 | <0.0001 |
| **sub_type (HR+/HER2-)**<br>**[base category: HR-/HER2-]** | -0.9708 | 0.3788 | 0.0281 | <0.0001 |
| **sub_type (HR+/HER2+)**<br>**[base category: HR-/HER2-]** | -0.9738 | 0.3777 | 0.0422 | <0.0001 |

**Commenges- Andersen test for heterogeneity: p-value = 0.0165**

**Likelihood ratio test for heterogeneity:**

**No-frailty log-likelihood = -84378.39**

**With-frailty log-likelihood = -84363.04**

**LRT p-value = 1.51e-08**

**Frailty variance: 95% confidence intervals in the ( )**

**Var[Z] = 0.018 (0.006 – 0.061)**

Table 4: Posterior frailty estimate for each category of county-level median household income

| less than $35,000 | $35,000 - $39,999 | $40,000 - $44,999 | $45,000 - $49,999 | $50,000 - $54,999 |
|---|---|---|---|---|
| 1.1899 | 0.9941 | 1.2126 | 1.137 | 1.004 |

| $55,000 - $59,999 | $60,000 - $64,999 | $65,000 - $69,999 | $70,000 - $74,999 | $75,000 and above |
|---|---|---|---|---|
| 0.9244 | 0.9133 | 0.8821 | 0.8740 | 0.8838 |

Publisher
**Society of Statistics, Computer and Applications**
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2023

**Printed by : Galaxy Studio & Graphics**
**Mob: +91 9818 35 2203, +91 9582 94 1203**