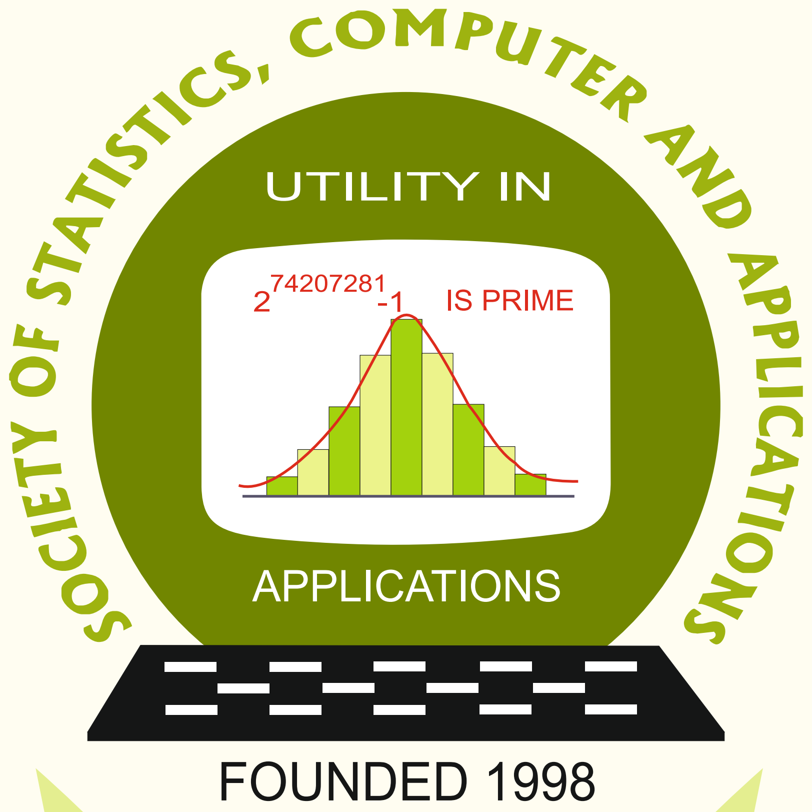
ISBN No.978-81-950383-1-2

Special Proceedings (24)
(Based upon the 24th Annual Conference of SSCA held at ICAR National Academy of Agricultural Research Management Hyderabad, Telangana)

23 - 27 February 2022


SOCIETY OF STATISTICS, COMPUTER AND APPLICATIONS
UTILITY IN
$2^{74207281}-1$ IS PRIME
APPLICATIONS
FOUNDED 1998

Society of Statistics, Computer and Applications
https://ssca.org.in/
2022

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.978-81-950383-1-2

Special Proceedings (24)
(Based upon the 24th Annual Conference of SSCA held at ICAR National Academy of Agricultural Research Management Hyderabad, Telangana)

23 - 27 February 2022

Editors

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

Date of Publication: September 2022

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
2022

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 #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022

Preface

The Society of Statistics, Computer and Applications (SSCA) was founded in 1998 with a goal to provide a platform for promotion and dissemination of research in Statistics, blended with information technology, among both theoretical and applied statisticians, who have keen interest in the applications of Statistics to varied fields like agriculture, biological sciences, medical sciences, financial statistics, and industrial statistics. Since then, the Society has been performing several activities and promoting development of theoretical and applied research work in Statistics and Informatics.

One of the major activities of SSCA has been to organize national/international conferences annually across the length and breadth of the country. SSCA also brings out a journal called *Statistics and Applications.* This is an open access journal and is available at the website of the Society (www.ssca.org.in). The full-length papers can be viewed and downloaded free of cost. Besides bringing out regular volumes of the journal, special volumes on emerging thematic areas of global/national importance are also brought out.

The twenty-fourth Annual Conference of the SSCA was organized during 23-27 February 2022 at the ICAR National Academy of Agricultural Research Management, Hyderabad, Telangana. This was the second web conference organized online to observe appropriate COVID – 19 protocols. The theme of the conference was *Recent Advances in Statistical Theory and Applications* (*RASTA* 2022). The conference was academically enriching with important and significant presentations made by scientists of international repute and eminence. Many technical sessions organized during the conference were participated by renowned international and national statisticians and a session on Financial Statistics, in which renowned statisticians and leading practitioners from National Bank for Agriculture and Rural Development, Indian Statistical Institute, Savitri Phule Pune University, S P Jain School of Global Management; USA and Wipro Ltd., USA made presentations. Two special memorial sessions were organized in the memory of Late Hukum Chandra and Late Lalmohan Bhar. The speakers in this session were close associates, collaborators, friends and students of Late Hukum Chandra and Late Lalmohan Bhar. We are pleased to share that the special proceedings have got ISBN #: 978-81-950383-1-2.

The Executive Council of the SSCA decided to bring out "Special Proceedings" of the conference covering some important selected talks including those presented in the Financial Statistics session. The selection of authors was made based upon the contents as presented during the conference. The Executive Council of the Society nominated V.K. Gupta, Baidya Nath Mandal, R. Vishnu Vardhan, Ranjit Kumar Paul, Rajender Parsad, and Dipak Roy Choudhury as Guest Editors for bringing out these special proceedings. The Guest Editors finalized the names of authors to be invited to submit their full paper, based upon their presentation during the conference, for the Special Proceedings.

Distinguished speakers shortlisted for making contributions to the special proceedings were invited to submit their research papers for possible inclusion in the special proceedings. After the usual review process, 14 research papers were accepted for publication and are included in the special proceedings. We would like to express our sincere thanks to all the authors for responding to our request and submitting their research for publication in these special proceedings in time. The reviewers have also made a very big contribution by way of finishing the review process in a short span of time and it is a pleasure to thank each one of them individually. We would like to place on record our gratitude to all members and office bearers of the Executive Council of SSCA for their support. We would also like to express our sincerest thanks to Prof. Ch. V. Srinivas Rao, Director, ICAR NAARM and his dedicated team especially Dr. G. Venkateshwarlu, Dr. A. Dhandapani, and Dr. S. Ravichandran for organizing such an academically enriching second web conference of the SSCA.

Guest Editors

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

New Delhi
September 2022

Special Proceedings: ISBN#: 978-81-950383-0-5
24th Annual Conference, 23–27 February 2022; pp 1–30

Personal Photographic Glimpses of Professor Bikas Kumar Sinha: Bikas Kumar Sinha Endowment Lecture

Simo Puntanen
Faculty of Information Technology and Communication Sciences
Tampere University, Finland

Received: 16 May 2022; Revised: 01 June 2022; Accepted: 10 June 2022

Abstract

My goal in this Endowment Lecture is to provide some glimpses of Professor Bikas Sinha, based on 30 years of friendship and appreciation. In particular I'll emphasize Bikas' remarkable impact as a demanding but inspiring mentor and collaborator of several doctoral students of Tampere University; among them, Arto Luoma, Laura Koskela, Anne Puustelli, Klaus Nordhausen, Jarkko Isotalo and Sami Helle; and extremely active and productive co-operation with Erkki Liski and Tapio Nummi and their research group.

On my personal side I'll go through, with photographic illustrations, some highlights like visiting Bikas' residence and family in Kolkata on the 3rd of January 1995; visits to our home in Nokia; international scientific meetings like in Kolkata, Hyderabad and in Będlewo, Poland, and the guided tour in Kolkata Botanical Gardens on the 2nd of January 2013.

Key words: Design of experiments; Linear models; Indian Statistical Institute; Calcutta University Department of Statistics, Tampere University.

AMS Subject Classifications: 62K05.

1. Introduction: Spring 1993

An Afternoon Full of Experimental Design, by Professor Bikas Kumar Sinha, Indian Statistical Institute, Calcutta, Wednesday, 14 April 1993, 12:15–16:00, Lecture Room C VII, University of Tampere:

Optimality Aspects of Experimental Designs

"I plan now to give a very lucid account of the concept of optimality in experimental designs. I will make a survey of all the available results, explaining their historical perspectives and the necessity underlying the continued research that has taken place in the last forty years. Of course, Weighing Designs is one of my very favourite topics and I would like to spend some time on this fascinating area."

Corresponding Author: Simo Puntanen
Email: simo.puntanen@tuni.fi

That was the email I circulated around Southern Finland in early April 1993. What was going to happen—the first lecture that Professor Bikas Sinha delivered in Finland, so it was a premier for many talks to be given by him in Finland. The most recent one, if I'm not mistaken, was on 2 November 2017, with the title "F AM NOT LICKED —The Twelve Penny Problem with Applications". And, between those 24 years, there were several talks and short courses with varying titles.

Bikas: it's an Olympiad from your latest tour to Finland! Clearly time to come back.

On 6 April 1993 Bikas sent the following email:

> Dear Simo ... Thanks for your e-reply. I got Finland Visa. That means I am all set now. I will reach Helsinki on April 11 (Sunday) in the afternoon at about 3 pm. Any suggestion for convenient train for Tampere? Fare? Currency exchange rate? Where to get down in Tampere and how to go to the University Campus? Accommodation at ???? Let me know all this. Thanks ... Bikas.

Below is a part of my reply:

> I think that you probably will catch the 17:00 train. So I will come to the Tampere train station at 18:48. If I don't see you then, I will come again at 20:02. [*That's when we actually met.*] ... At the Tampere train station, please wait in the platform, just where you come out of the train, I'll meet you there. –Simo

Then in Tampere, during Bikas' stay in April 1993, I had to be away three days due to a conference trip. Later, in an interview by Tapio Nummi (Tapio is Bikas' long-standing co-researcher) in Tampere in November 2017, Bikas described his first visit to Tampere [incl. my absence] as follows:

> I thought within myself: I am not going to sit down in one corner office and prove some theorems for myself! I must know what others are doing here. So next morning with a cup of hot tea in my hand, I knocked at the door of one faculty ... without having any prior knowledge about him/her at all ... absolutely nothing ... leaving it to the "chance" !!!
>
> And it turned out that I had entered into Professor Erkki Liski's room. ... He was to be my co-author for the next 10–12 years with some doctoral students at his disposal. ... he was the first I started interacting with very successfully.

The above events are descriptive characters of Bikas' academic and social career: his energy and open-mindedness are his mental eigenvalues, so to say. My goal in this Endowment Lecture is to provide some glimpses of Bikas, based on 30 years of friendship and appreciation. In particular I'll emphasize Bikas' remarkable impact as a demanding but inspiring mentor and collaborator of several doctoral students of Tampere University; among them, Arto Luoma, Laura Koskela, Anne Puustelli, Klaus Nordhausen, Jarkko Isotalo and Sami Helle.

On my personal side I'll go through, with photographic illustrations, some highlights like visiting Bikas' home and family in Kolkata in early January 1995, meetings at our home in Nokia, international scientific meetings like in Hyderabad and in Będlewo/Poland, and the guided tour in Kolkata Botanical Gardens in early January 2013.

2. Bikas' Research Experience

Bikas' research experience can be broadly divided into three phases:

- As a Research Scholar (1968–1971) and the youngest Faculty Member in the Calcutta University Department of Statistics, CUDS, (1972–1975).
- As a Faculty at the ISI, Indian Statistical Institute, Kolkata (1979–2011). Bikas regards CUDS and ISI as his Home Grounds.
- As a regular visitor (for long and short terms) in some universities in the USA, Canada, Germany, Finland, Poland and Thailand since 1982 till date. Since 1979, close interactions with the Faculty at the Home Grounds.

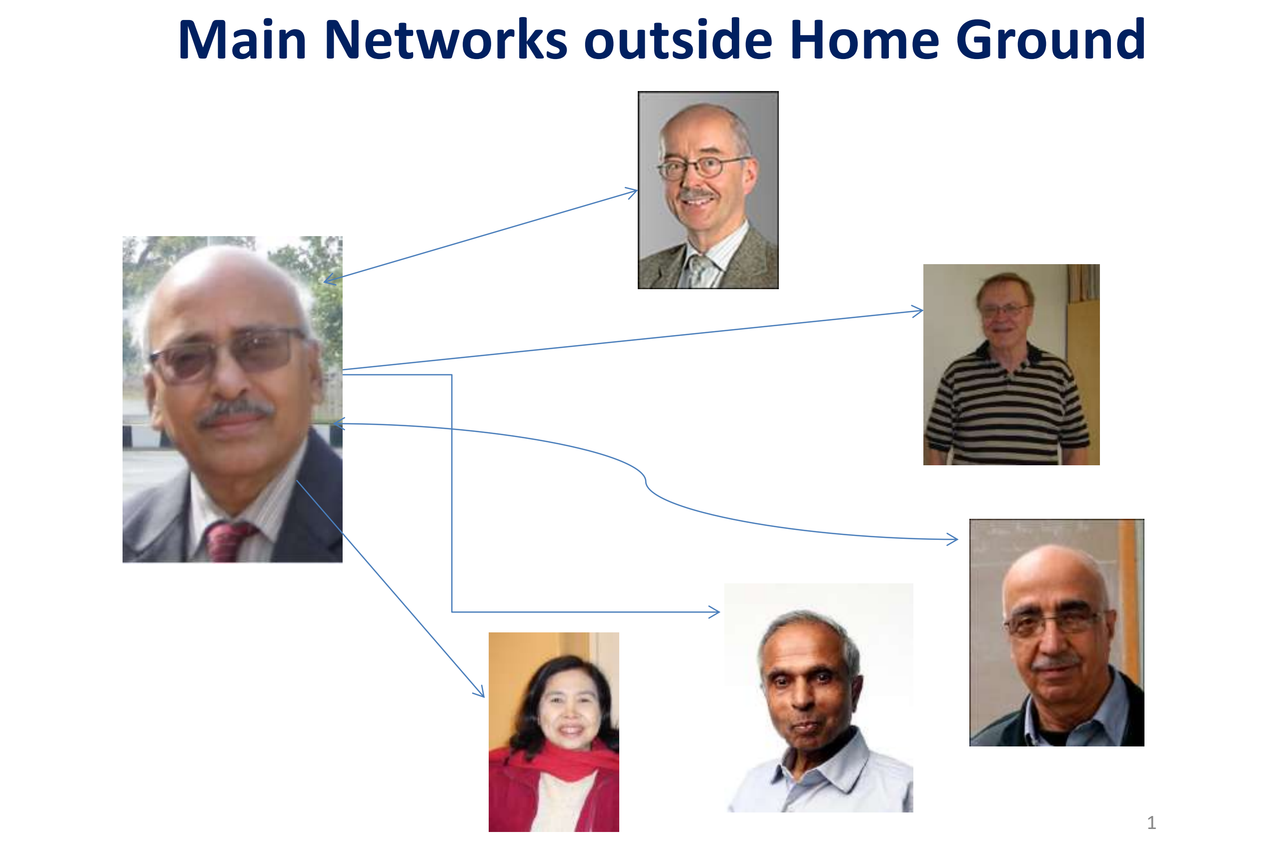
The Home Ground of Bikas' twin brother Bimal Kumar Sinha, has been the University of Maryland, Baltimore, USA, since 1985.

I don't go through in details the impressive scientific career of Bikas; there are excellent sources for that, see, e.g.,

- 2020: *Statistics and Applications*, Journal of the Society of Statistics, Computer and Applications, Vol. 18, No. 2. Special Issue "Challenges and Opportunities in Statistical Data Designing and Inference in the Emerging Global Scenario" in honour of twin statisticians (brothers) Bimal and Bikas Sinha on their 75th Birthday (16 March 2021). *Combined Issue pdf.*
- 2020: The Sinha Brothers (Bimal and Bikas)—The Twin Statisticians. Pages xi–xxx, in *Statistics and Applications*, Journal of the Society of Statistics, Computer and Applications, 2020, Vol. 18, No. 2. PDF.
- 2001: Bio-Sketch of Bikas K. Sinha. *American Journal of Mathematical and Management Sciences*, 21, 3–4, 222–225, DOI.

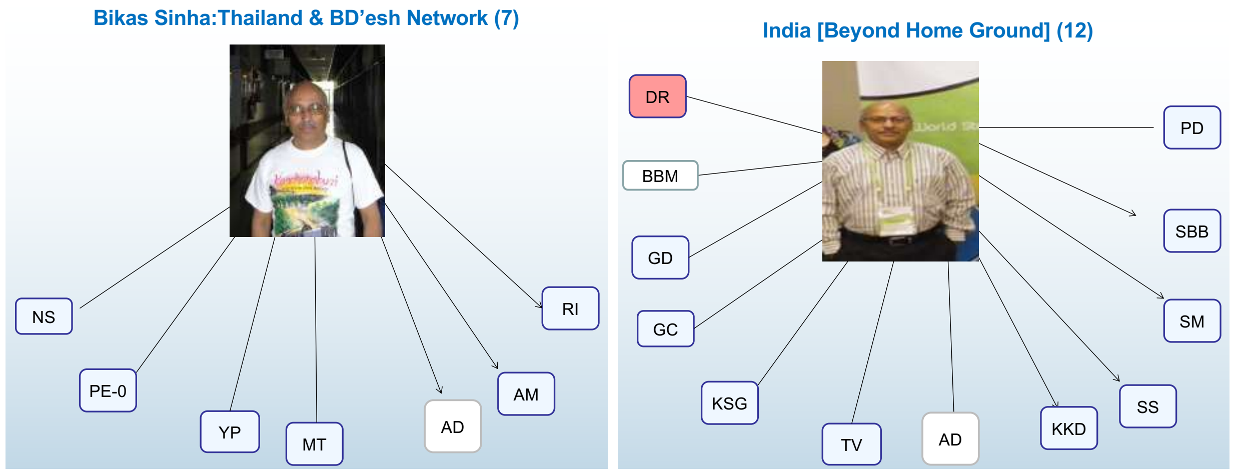
In particular, Bikas himself has prepared an outstanding summary of his career and collaborators: *Bikas Sinha Collaborators PPT October 2021.pptx.*

Believe or not but he has had about 111 collaborators by October 2021!



Main Networks outside Home Ground

Figure 1: Montip Tiensuwan, Kirti R. Shah, A. Samad Hedayat, Erkki Liski, Friedrich Pukelsheim. [Source: BKS's Collaborators PPT]



Bikas Sinha:Thailand & BD'esh Network (7)

India [Beyond Home Ground] (12)

Figure 2: "I have been really fortunate to have the ability to develop good personal relationship in no time . . . my collaborators have mostly been very reasonable and accepted me the way I have been." [Source: BKS's Collaborators PPT]

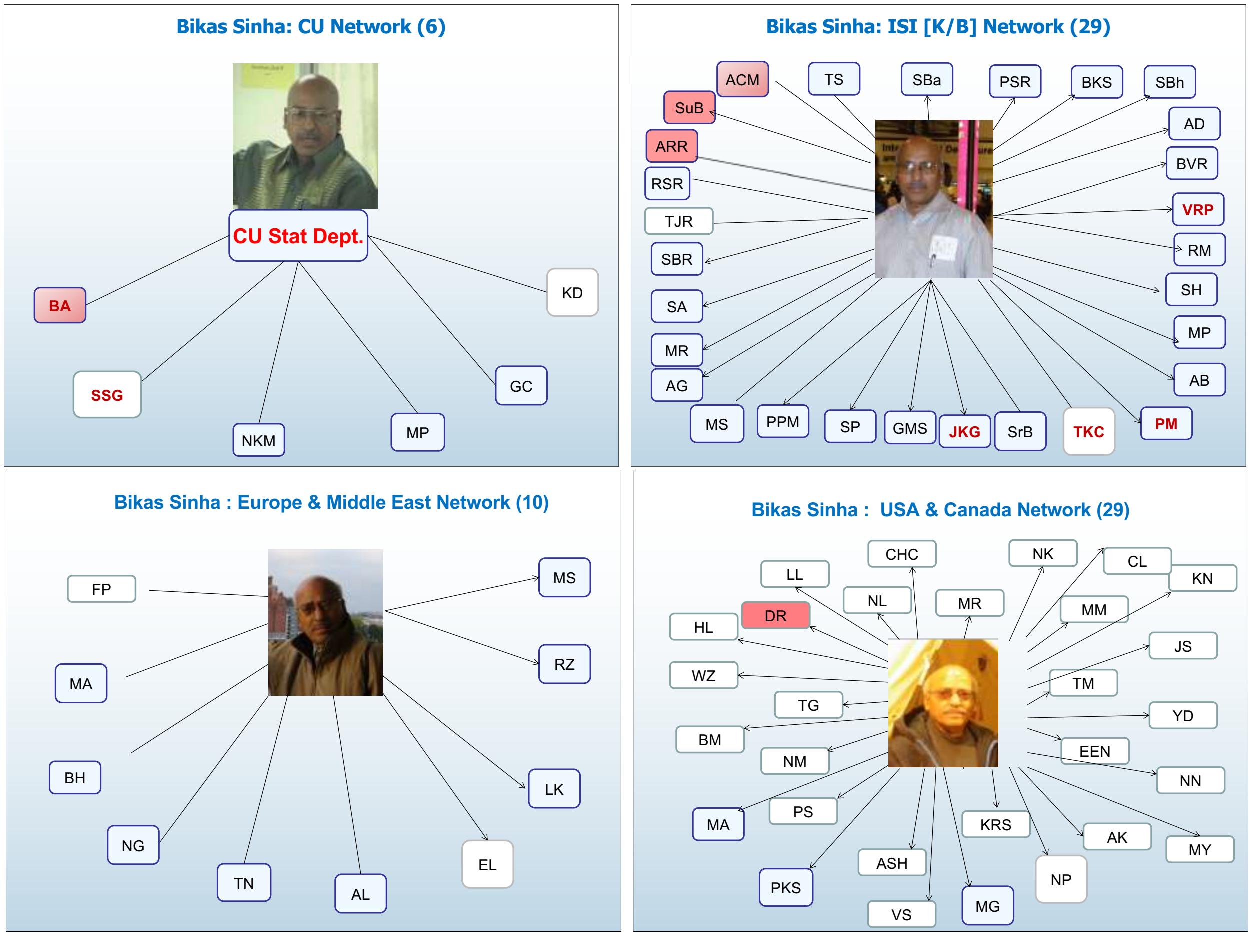


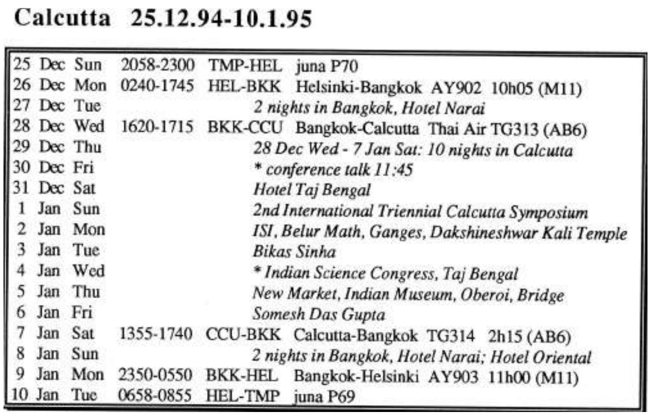
Figure 3: "You have contributed towards my understanding of the research problems, techniques to be developed and the solutions thus arrived at." [Source: BKS's Collaborators PPT]

Below are excerpts from P.K. Sen's Foreword in the special issue of the *Statistics and Applications* (2020), in honour of Bimal and Bikas Sinha on their 75th Birthday.

> It is indeed a pleasure to write this Foreword note in favor of two outstanding statisticians of our time: Bimal and Bikas Sinha, whom we affectionately address as the "Statistical Twins".
>
> They were born in 1946 in the village Atgharia in Pabna District of Bengal Province in undivided India, a year before the region was engulfed in the newly created East Pakistan (which later gave birth to Bangladesh in 1971 December).
>
> The Sinha family migrated to Kolkata, West Bengal, India, in 1958 when they were just 12 years old.
>
> For 12 years to follow, they had hard economic time, albeit both Bimal and Bikas excelled in their college education and earned their Ph.D. degrees in Statistics in 1972–1973 under the able guidance of [Late] Professor Hari Kinkar Nandi in Calcutta University. . . . During the past fifty years their research and organizational accomplishment may simply be categorized as outstanding.

3. 1995–2006: Bikas' Residence and Other Stories

As mentioned in Section 1, my first contact with Bikas took place in Kolkata in February 1993. Luckily I had opportunity to visit Kolkata soon again in December 1993 and over the new year's period 1994–1995.



Calcutta 25.12.94-10.1.95

25	Dec	Sun	2058-2300	TMP-HEL	juna P70
26	Dec	Mon	0240-1745	HEL-BKK	Helsinki-Bangkok AY902 10h05 (M11)
27	Dec	Tue			*2 nights in Bangkok, Hotel Narai*
28	Dec	Wed	1620-1715	BKK-CCU	Bangkok-Calcutta Thai Air TG313 (AB6)
29	Dec	Thu			*28 Dec Wed - 7 Jan Sat: 10 nights in Calcutta*
30	Dec	Fri			** conference talk 11:45*
31	Dec	Sat			*Hotel Taj Bengal*
1	Jan	Sun			*2nd International Triennial Calcutta Symposium*
2	Jan	Mon			*ISI, Belur Math, Ganges, Dakshineshwar Kali Temple*
3	Jan	Tue			*Bikas Sinha*
4	Jan	Wed			** Indian Science Congress, Taj Bengal*
5	Jan	Thu			*New Market, Indian Museum, Oberoi, Bridge*
6	Jan	Fri			*Somesh Das Gupta*
7	Jan	Sat	1355-1740	CCU-BKK	Calcutta-Bangkok TG314 2h15 (AB6)
8	Jan	Sun			*2 nights in Bangkok, Hotel Narai; Hotel Oriental*
9	Jan	Mon	2350-0550	BKK-HEL	Bangkok-Helsinki AY903 11h00 (M11)
10	Jan	Tue	0658-0855	HEL-TMP	juna P69

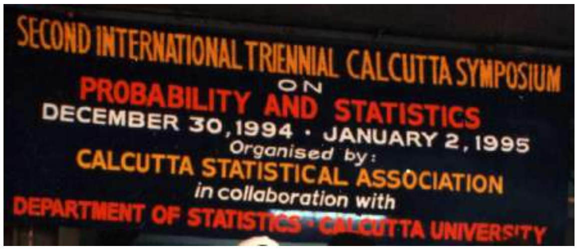

SECOND INTERNATIONAL TRIENNIAL CALCUTTA SYMPOSIUM
ON
PROBABILITY AND STATISTICS
DECEMBER 30, 1994 · JANUARY 2, 1995
Organised by:
CALCUTTA STATISTICAL ASSOCIATION
in collaboration with
DEPARTMENT OF STATISTICS · CALCUTTA UNIVERSITY

Figure 4: Timetable for Kolkata, 25 Dec 1994–10 Jan 1995. *Second International Triennial Calcutta Symposium on Probability and Statistics.* Accompanied by a good group of colleagues like Friedrich Pukelsheim and Götz Trenkler.

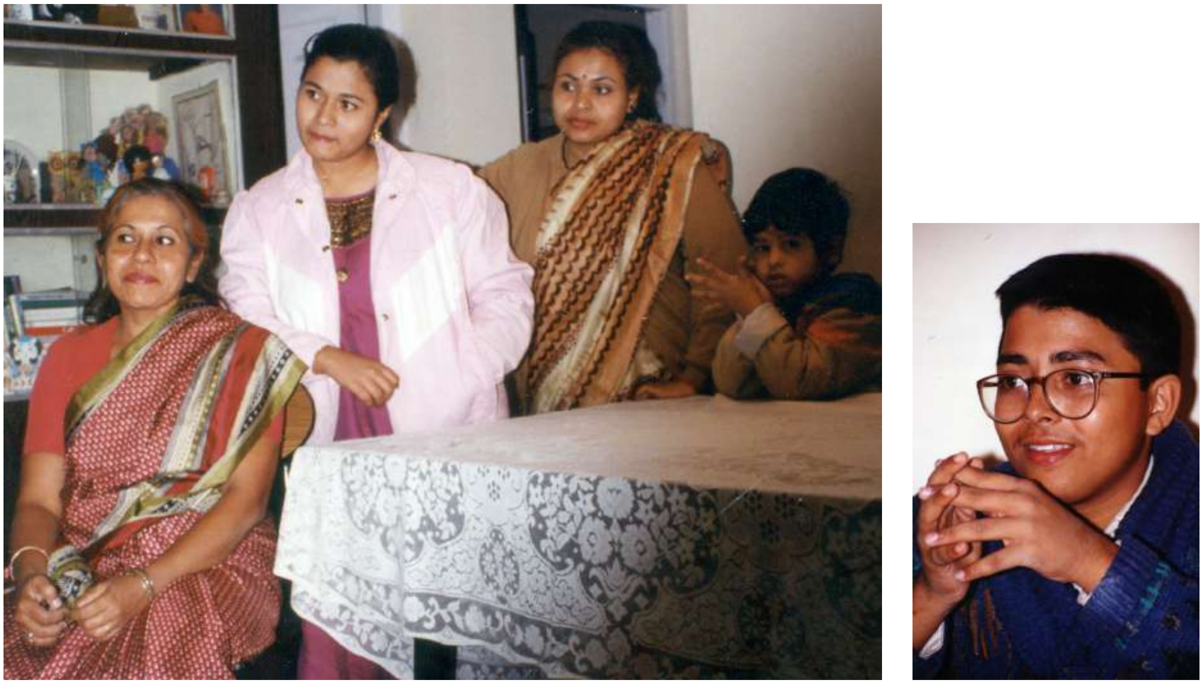


Figure 5: Pritha (= Mrs BKS), Karabi (= daughter of BKS), right: Kuver (= son of BKS); Bikas' residence, 3 January 1995.

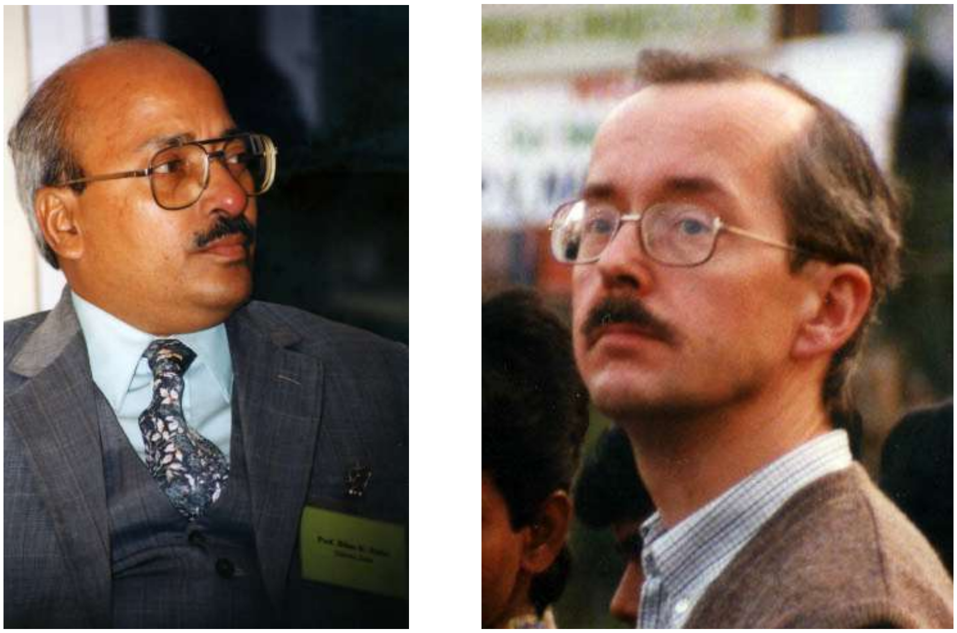


Figure 6: On the 3rd of January 1995, in Bikas' residence, Bikas himself was performing, singing to us (incl. Friedrich Pukelsheim). I took a video but could not find it … excellent singing, Bikas!

Figure 7: Pritha, Karabi and Kuver Sinha: train from Delhi to Joypur, 1989. — Pritha Sinha has been a housewife with devotion to nurturing the children, Karabi and Kuver. She holds Master's Degree in Education from Calcutta University. Only recently she completed a Certificate Course in Psychiatric Counseling and got absorbed as a Psychiatric Counsellor in one NGO, called Sanlaap. [Ph: BKS]

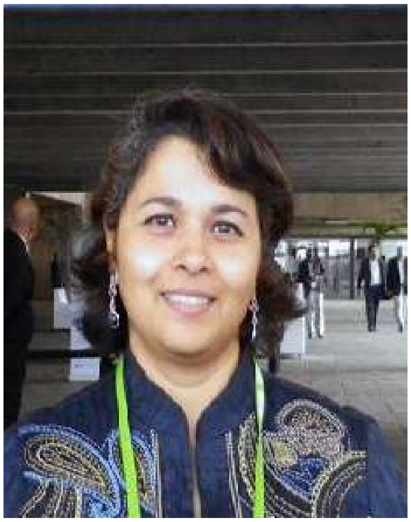


Figure 8: Karabi Nandy, née Sinha, is currently an Associate Professor at the University of Texas Southwestern Medical School (Clinical Sciences and Biostatistics).

PhD in 2004, from the University of Florida, Gainesville, supervisor Professor Malay Ghosh. The thesis was entitled *Some Contributions to Small Area Estimation.* [Ph: BKS]

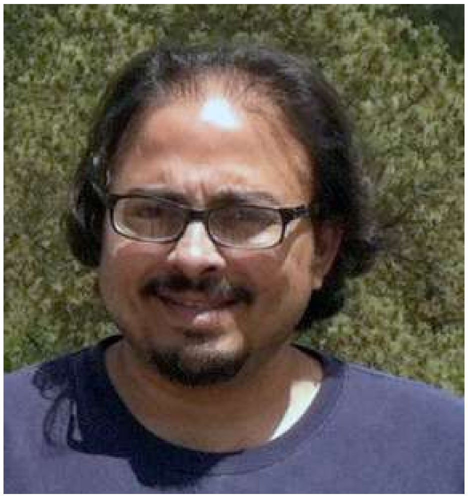


Figure 9: Kuver Sinha has been in the USA since 2001. PhD in 2008 from Rutgers University, USA; thesis: *Supersymmetry Breaking in Gauge Theories and String Theory.* Currently he is the Carl T. Bush Chair Professor of Astro-Physics at Oklahoma State University.

"I study the first three minutes of the history of the Universe (we've zeroed things down to the first three seconds)".

In late 1990s my daughter Anna made an interesting interview of Kuver to *Aamulehti*, a local newspaper in Tampere. Too bad, I could not find this interview anymore. [Ph: BKS]

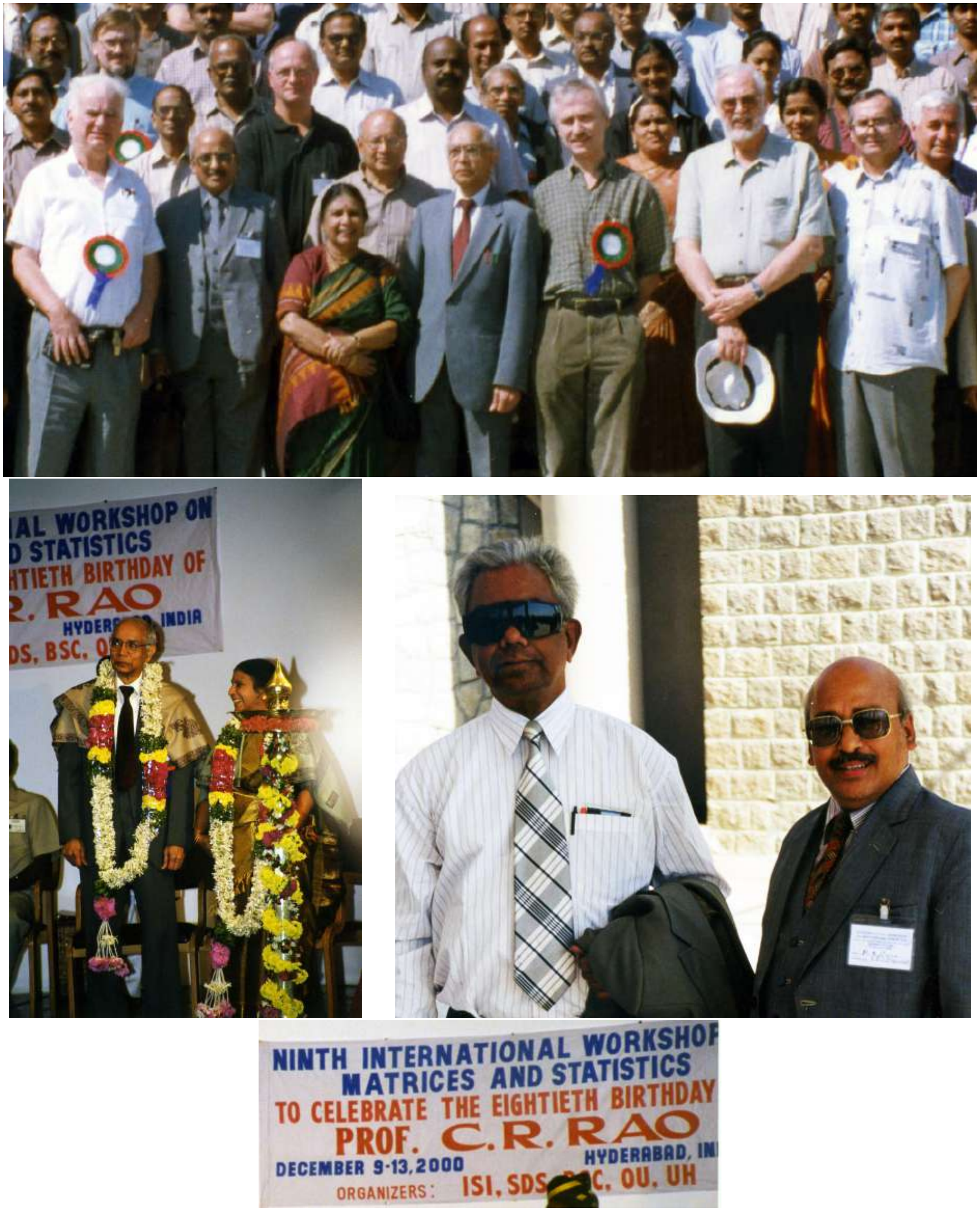


Figure 10: *The 9th International Workshop on Matrices and Statistics*, Hyderabad, 9–13 December 2000, celebrating C.R. Rao's 80th birthday. In the group photo you can catch George P.H. Styan, Bikas, Götz Trenkler, Mrs Rao, C.R. Rao, H.J. Werner, Gene Golub, Jerzy K. Baksalary, Fikri Akdeniz; Bikas with Somesh Dasgupta. Program.

Figure 11: Roman “Robin” Zmyślony and Bimal K. Sinha, Będlewo, Poland, 21–27 August 2003, *LinStat/StatLin Conference.* Program.

Figure 12: Tampere, 2 September 1983. You may catch Johan Fellman, Erkki Liski, Gustav Elfving, Tarmo Pukkila, Bill Farebrother, C.R. Rao, and Bimal K. Sinha (only 50 % visible). Conference in Linear Models in Tampere, 30 August – 2 September 1983. Bimal’s talk: “Robustness in linear models” (joint with Hilmar Drygas). Program. Proceedings.

Figure 13: Bikas and Bimal. Bimal or Bikas? [Ph: BKS]

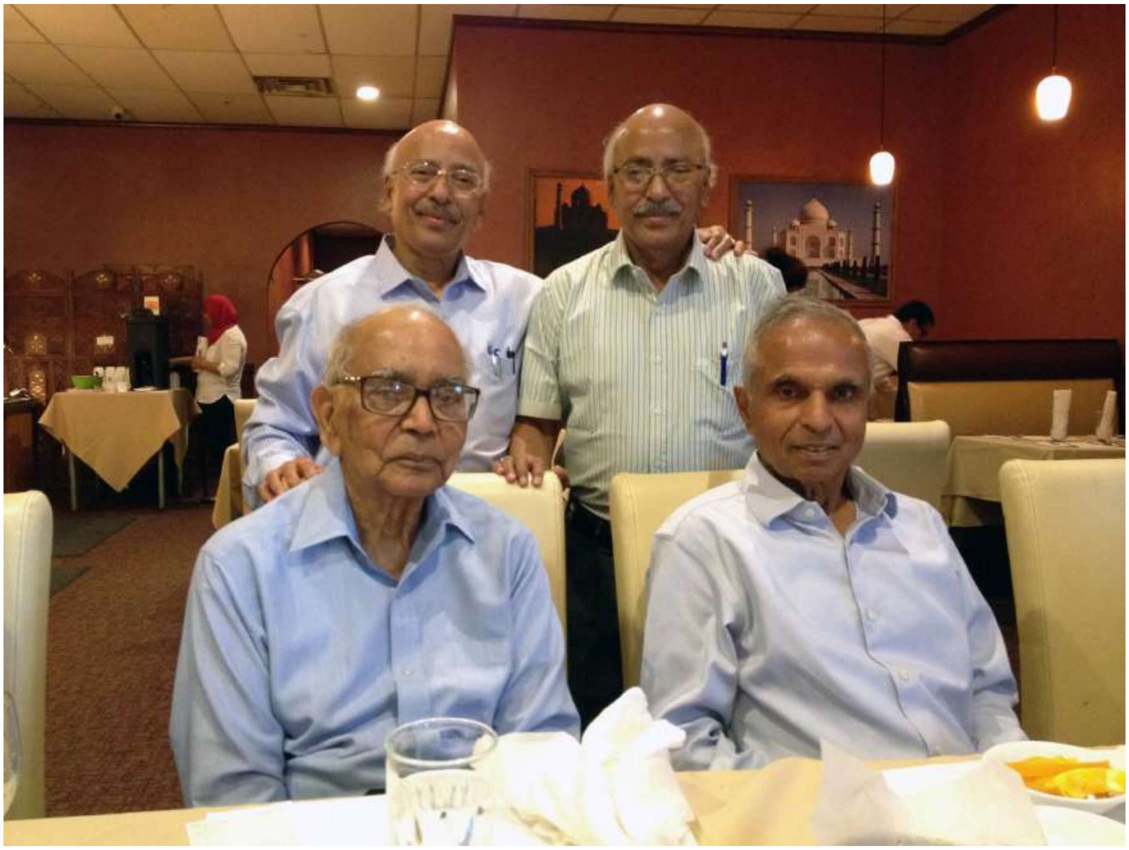
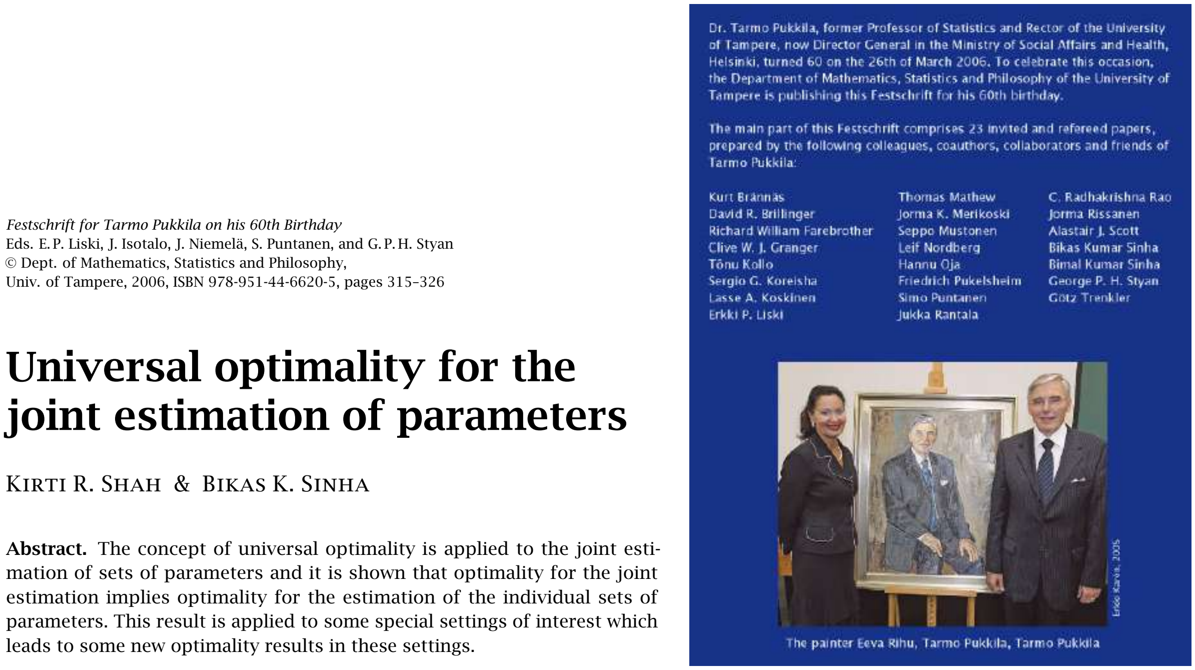


Figure 14: Bikas and Bimal with C.R. Rao and Kirti R. Shah, 2018. [Ph: BKS]



Festschrift for Tarmo Pukkila on his 60th Birthday
Eds. E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. P. H. Styan
© Dept. of Mathematics, Statistics and Philosophy,
Univ. of Tampere, 2006, ISBN 978-951-44-6620-5, pages 315–326

Universal optimality for the joint estimation of parameters

Kirti R. Shah & Bikas K. Sinha

Abstract. The concept of universal optimality is applied to the joint estimation of sets of parameters and it is shown that optimality for the joint estimation implies optimality for the estimation of the individual sets of parameters. This result is applied to some special settings of interest which leads to some new optimality results in these settings.

Dr. Tarmo Pukkila, former Professor of Statistics and Rector of the University of Tampere, now Director General in the Ministry of Social Affairs and Health, Helsinki, turned 60 on the 26th of March 2006. To celebrate this occasion, the Department of Mathematics, Statistics and Philosophy of the University of Tampere is publishing this Festschrift for his 60th birthday.

The main part of this Festschrift comprises 23 invited and refereed papers, prepared by the following colleagues, coauthors, collaborators and friends of Tarmo Pukkila:

Kurt Brännäs
David R. Brillinger
Richard William Farebrother
Clive W. J. Granger
Tõnu Kollo
Sergio G. Koreisha
Lasse A. Koskinen
Erkki P. Liski
Thomas Mathew
Jorma K. Merikoski
Seppo Mustonen
Leif Nordberg
Hannu Oja
Friedrich Pukelsheim
Simo Puntanen
Jukka Rantala
C. Radhakrishna Rao
Jorma Rissanen
Alastair J. Scott
Bikas Kumar Sinha
Bimal Kumar Sinha
George P. H. Styan
Götz Trenkler

Erkki Karvia, 2005

The painter Eeva Rihu, Tarmo Pukkila, Tarmo Pukkila

Figure 15: Contribution to the *Festschrift* for Tarmo Pukkila, Tampere, June 2006.

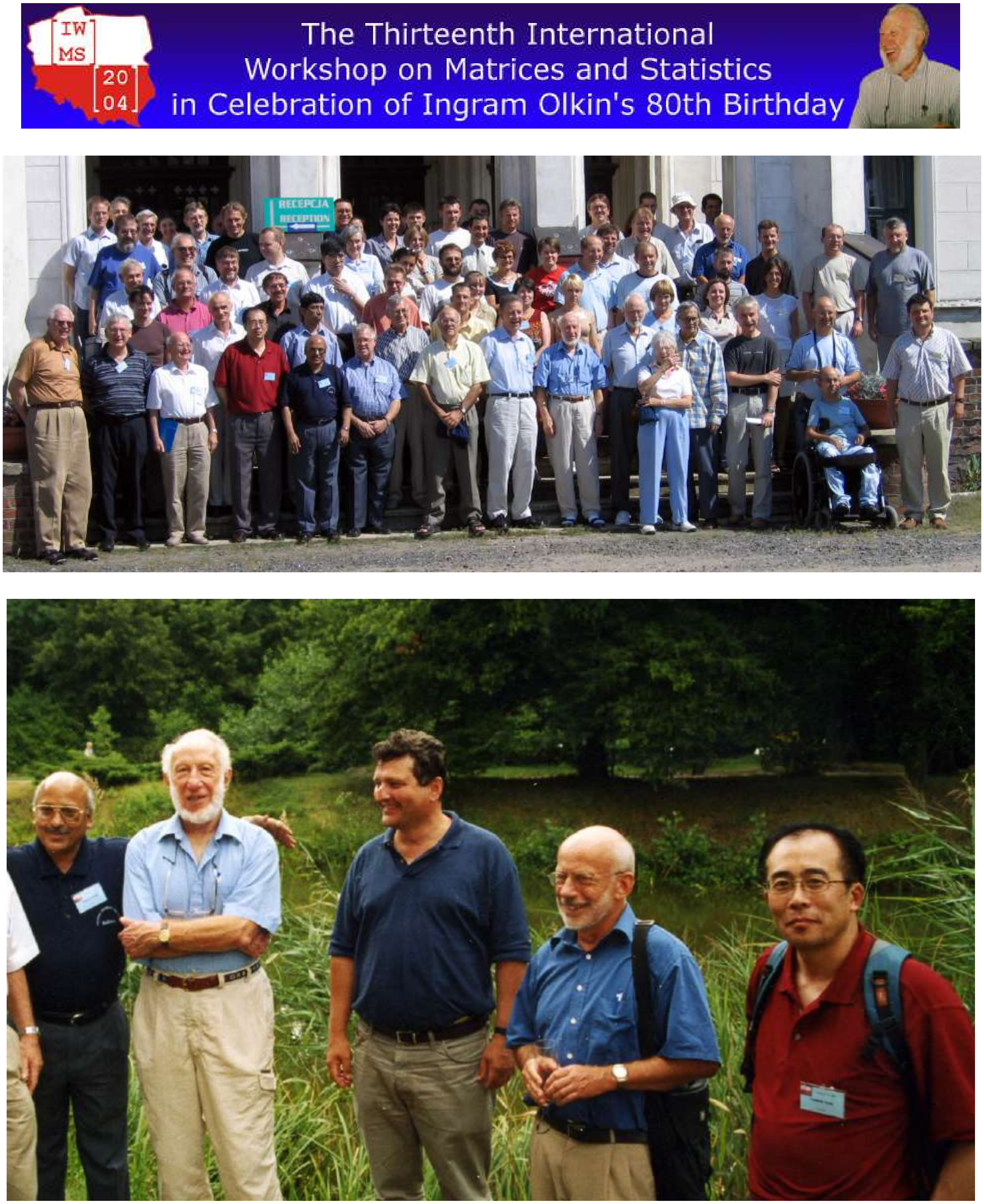

IW
MS
20
04
The Thirteenth International
Workshop on Matrices and Statistics
in Celebration of Ingram Olkin's 80th Birthday
RECEPCJA
RECEPTION

Figure 16: Bikas, Ingram Olkin, Augustyn Markiewicz, Ludvig Elsner, Yongge Tian. *The 13th International Workshop of Matrices and Statistics*, Będlewo, Poland, 19–21 August 2004. Program.

4. 2006–2007

Email from Bikas Sinha, 27 November 2006.

> Dear Erkki and Simo: Greetings from Kolkata !
>
> You will be glad to learn that my daughter Karabi is getting married next week Tuesday . . . Dec. 4th . . . here in Kolkata. The Groom is also from Kolkata . . . studied at ISI . . . PhD from Washington State University, Seattle, and currently a Faculty at UCLA.
>
> Bikas
>
> PS: Hope to see Simo in Kolkata and Hyderabad.

6th INTERNATIONAL TRIENNIAL
CALCUTTA SYMPOSIUM
On
PROBABILITY AND STATISTICS
"Kalah Kalayatam aham..."
"I am Time among all who measure and enumerate..."
...Gita, X. 30
WELCOME

- Linear Models & Generalized Inverses:
Organizer: **Dibyen Majumdar**
Speakers: **Simo Puntanen**
Thomas Mathew
Bhramar Mukherjee
Sumanta Adhdhya
Chair: **Dibyen Majumdar**

- Measurement Error,Biostatistics Dimension reduction:
Speakers: **Karabi Sinha**
Apratim Guha
Rahul Bhattacharyya
Abhijit Mandal
Chair: **Gudio Knapp**

"Finally, to the one who champions me through everything, to the one I look up to for everything, to my father, to Baba, I owe everything."
—*Thesis Preface / Karabi*

Figure 17: *The 6th Triennial CCU Symposium*, 29–31 December 2006. Program.

I recall that in 2004 or 2005 Bikas and Karabi were visiting Tampere and Bikas wanted to celebrate Karabi's thesis by taking our whole dept group for dinner to Restaurant Näsinneula: providing spectacular views!

Notice the excerpt from the Preface of Karabi's Thesis.

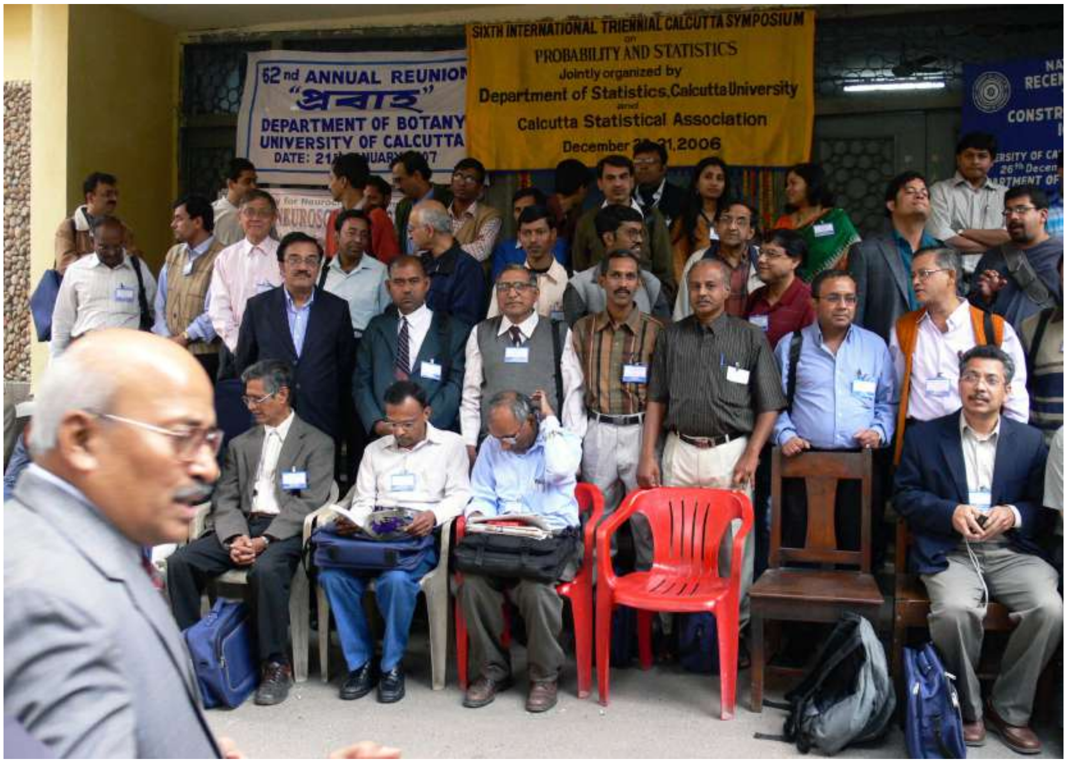
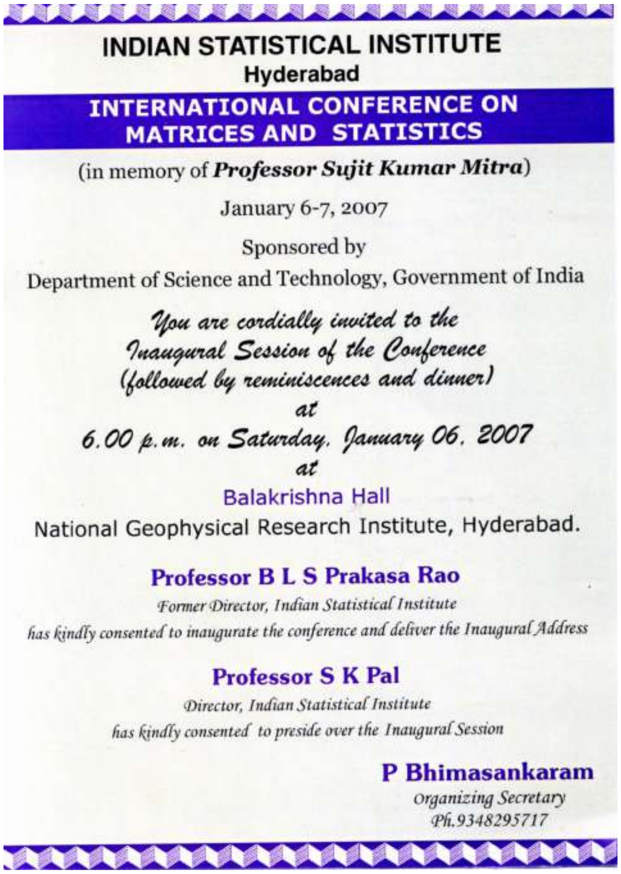

62nd ANNUAL REUNION
"প্রবাহ"
DEPARTMENT OF BOTANY
UNIVERSITY OF CALCUTTA
DATE: 21
SIXTH INTERNATIONAL TRIENNIAL CALCUTTA SYMPOSIUM
on
PROBABILITY AND STATISTICS
Jointly organized by
Department of Statistics, Calcutta University
and
Calcutta Statistical Association
December 29-31, 2006

Figure 18: Bikas rushing into the group photo. (He occupied the red chair and I took the next one, behind my backback.) *The 6th Triennial CCU Symposium*, 29–31 December 2006.

Figure 19: Hooghly River, an arm of Ganges. December 2006.



INDIAN STATISTICAL INSTITUTE
Hyderabad

INTERNATIONAL CONFERENCE ON
MATRICES AND STATISTICS

(in memory of ***Professor Sujit Kumar Mitra***)

January 6-7, 2007

Sponsored by
Department of Science and Technology, Government of India

You are cordially invited to the
Inaugural Session of the Conference
(followed by reminiscences and dinner)
at
6.00 p.m. on Saturday, January 06, 2007
at
Balakrishna Hall
National Geophysical Research Institute, Hyderabad.

Professor B L S Prakasa Rao
Former Director, Indian Statistical Institute
has kindly consented to inaugurate the conference and deliver the Inaugural Address

Professor S K Pal
Director, Indian Statistical Institute
has kindly consented to preside over the Inaugural Session

P Bhimasankaram
Organizing Secretary
Ph.9348295717

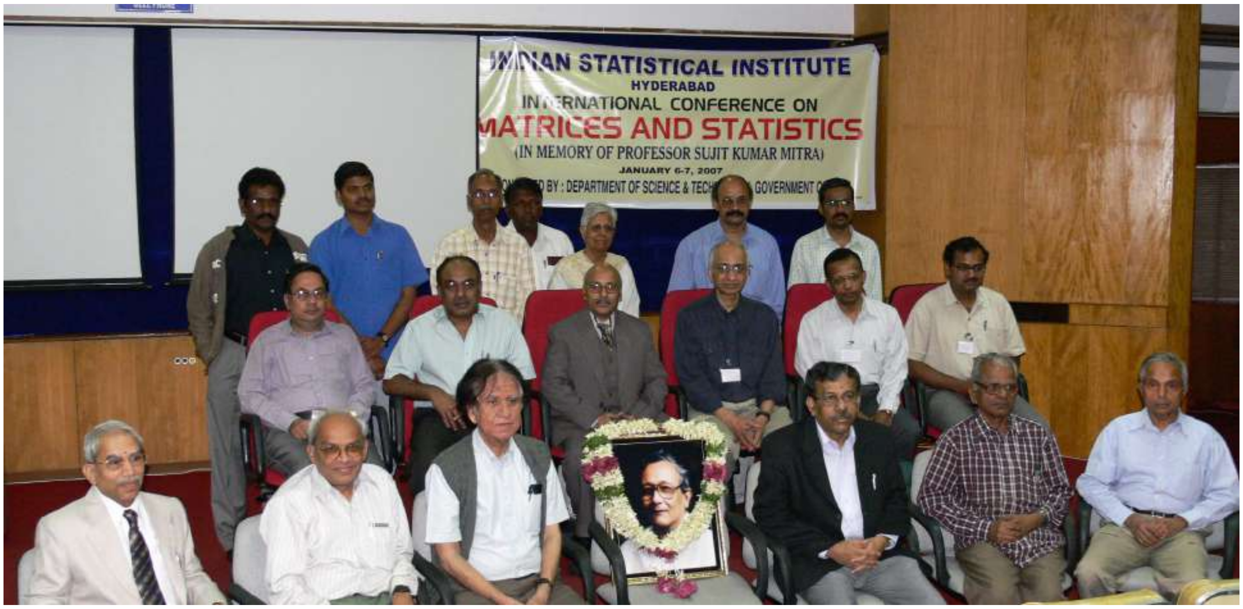

INDIAN STATISTICAL INSTITUTE
HYDERABAD
INTERNATIONAL CONFERENCE ON
MATRICES AND STATISTICS
(IN MEMORY OF PROFESSOR SUJIT KUMAR MITRA)
JANUARY 6-7, 2007
D BY : DEPARTMENT OF SCIENCE & TECH GOVERNMENT O

Figure 20: RB Bapat, P Bhimasankaram, JK Ghosh, K Krishnamoorthy, D Majumdar, S Malik, T Mathew, AR Meenakshi, SJP, BLSP Rao, P.S.S.N.V.P. Rao, SB Rao, D Sengupta, KR Shah, Bikas Kumar Sinha. *S.K. Mitra (1932–2004) Memorial Meeting,* Hyderabad, 6–7 January 2007. Report in *Image.*

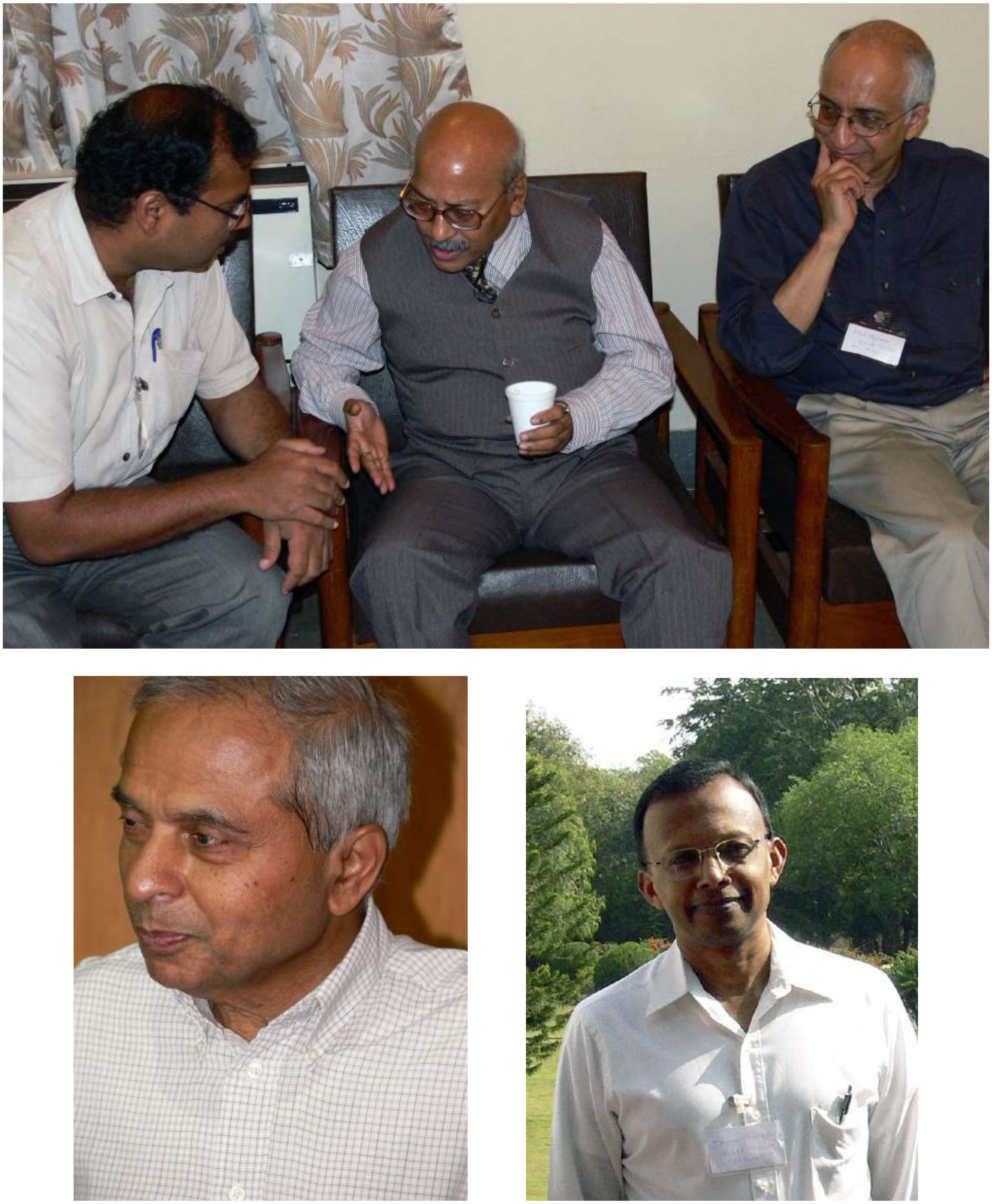


Figure 21: Above: Debasis Sengupta, Bikas, Dibyen Majumdar. Below: Kirti R. Shah, Thomas Mathew. Hyderabad, 6 January 2007.

5. Mentor in Tampere

Let's have a glance at Bikas' remarkable impact as a demanding but inspiring mentor and collaborator of several doctoral students of Tampere University; in particular, Arto Luoma and Laura Koskela.

ARTO LUOMA

Optimal Designs in Linear Regression Models

List of original papers

1 Liski, E.P., Luoma, A. and **Sinha**, Bikas K., (1996). Optimal Designs in Random Coefficient Linear Regression Models. *Calcutta Statistical Association Bulletin*, 46, 211–230.

2 Liski, E.P., Luoma, A., **Mandal**, N.K. and **Sinha**, Bikas K. (1998). Optimal Designs for Prediction in Random Coefficient Linear Regression Models. *Journal of Combinatorics, Informatics and Systems Sciences*, 23(1–4), 1–16.

3 Liski, E.P., Luoma, A., **Mandal**, N.K. and **Sinha**, Bikas K. (1997). Optimal Design for an Inverse Prediction Problem under Random Coefficient Regression Models. *Journal of the Indian Society of Agricultural Statistics*, XLIX, 277–288.

4 Liski, E.P., Luoma, A., **Mandal**, N.K. and **Sinha**, Bikas K. (1998). Pitman nearness, distance criterion and optimal regression designs. *Calcutta Statistical Association Bulletin*, 48(191–192), 179–194.

5 Liski, E.P., Luoma, A. and Zaigraev, A. (1999). Distance optimality design criterion in linear models. *Metrika*, 49(3), 193–211.

Below is an excerpt from the Preface of Arto Luoma's Thesis, 25 May 2000.

> I had my first contact with the subject of my thesis in summer 1994 when Professor **Bikas Sinha** from the Indian Statistical Institute visited our department. He had already cooperated with my supervisor **Erkki Liski** in experimental design and I had a joy to read one of their joint articles. They also welcomed me in their research group.
>
> In summer 1994 the preliminary drafts of the three first papers of the thesis were written. They provided a basis on which some future work could be done. It was a credit to Professor Sinha that he was able to formulate the problems in such a form that they could be dealt with. However, it was not easy to find complete solutions to some problems

which seemed simple at the outset. There was also a third person to join our research group, Dr. **Nripesh Mandal** from Calcutta University, who visited our department twice.

I am thankful to Professor Sinha who has been an example of creative scientific work and whose enthusiasm as a teacher I have admired.

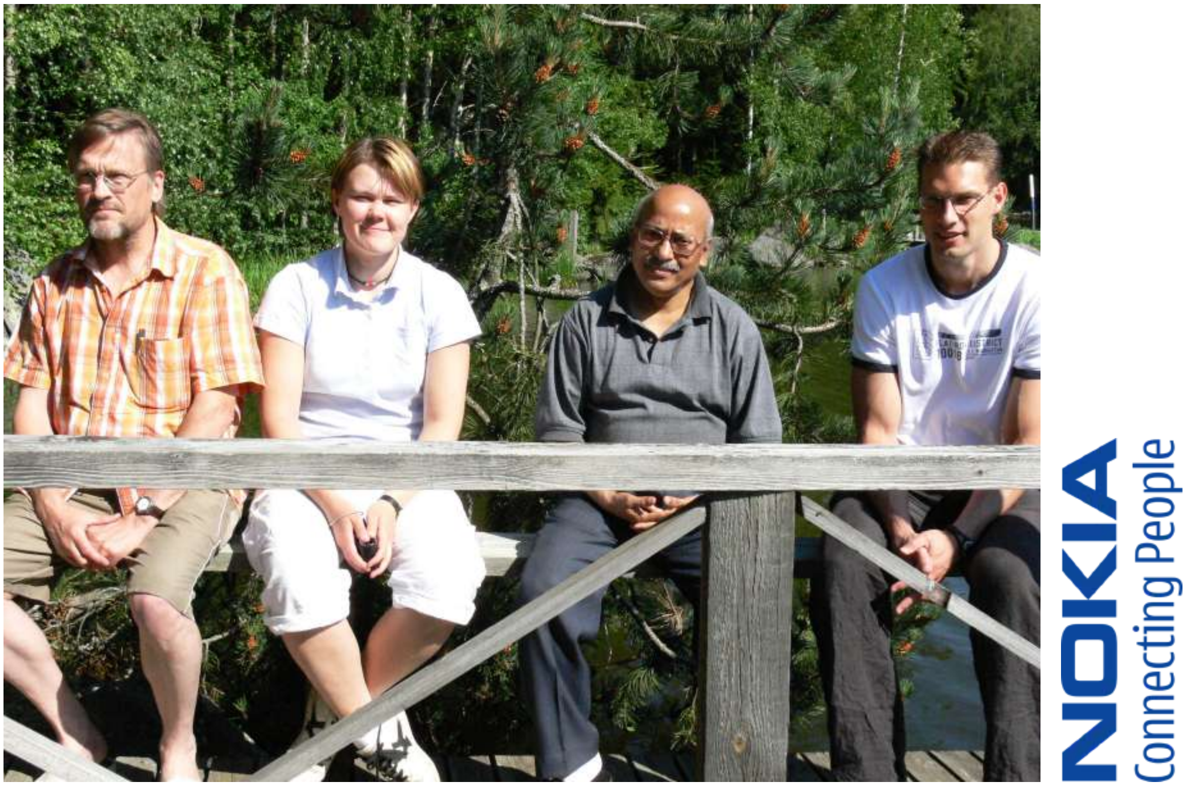
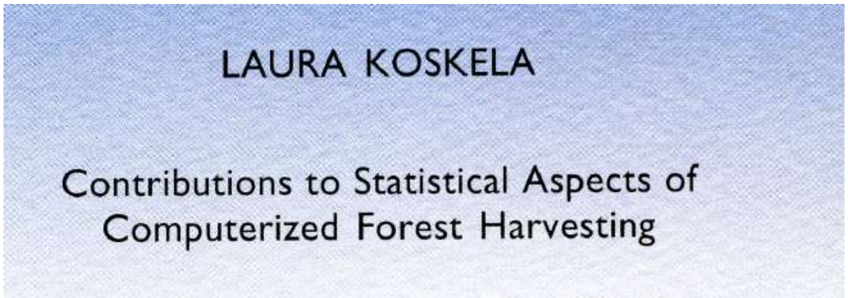


Figure 22: SJP, Laura Koskela, Bikas, Jarkko Isotalo, Nokia, 5 July 2006.

Laura Koskela had her thesis defence on 29 June 2007.



LAURA KOSKELA

Contributions to Statistical Aspects of
Computerized Forest Harvesting

List of Papers, Laura Koskela: 29 June 2007. [Photo below taken by Laura in July 2005.]

1. Koskela, L., Nummi, T., Wenzel, S., and Kivinen, V.-P. (2006). On the Analysis of Cubic Smoothing Spline-Based Stem Curve Prediction for Forest Harvesters. *Canadian Journal of Forest Research*, Vol. 36, 2909–2920.

2. Nummi, T., and Koskela, L. (2006). Analysis of Growth Curve Data Using Cubic Smoothing Splines. Submitted to *Journal of Applied Statistics*.

3. Nummi, T., **Sinha**, B.K., and Koskela, L. (2005). Statistical Properties of the Apportionment Degree and Alternative Measures in Bucking Outcome. *Revista Investigation Operacional*, 26, 3, 259–267.

4. Koskela, L., **Sinha**, B.K., and Nummi, T. (2007). Some Aspects of the Sampling Distribution of the Apportionment Index and Related Inference. Submitted to *Silva Fennica*.

5. **Sinha**, B.K., Koskela, L., and Nummi, T. (2005). On a Family of Apportionment Indices and Its Limiting Properties. *IAPQR Transactions*, 30, 2, 65–87.

6. **Sinha**, B.K., Koskela, L., and Nummi, T. (2005). On Some Statistical Properties of the Apportionment Index. *Revista Investigation Operacional*, 26, 2, 169–179.

Preface

I would also like to express my deepest gratitude to my assistant supervisor Professor Bikas K. Sinha, who has contributed immensely to my work during his regular visits to the University of Tampere.

I warmly thank him for his willingness to share his wide expertise and a number of ideas throughout the preparation of the thesis.

I am also grateful to him for arranging the opportunity to make an unforgettable and scientifically fruitful one month's visit to the Indian Statistical Institute (lSI), Kolkata, in July 2005.

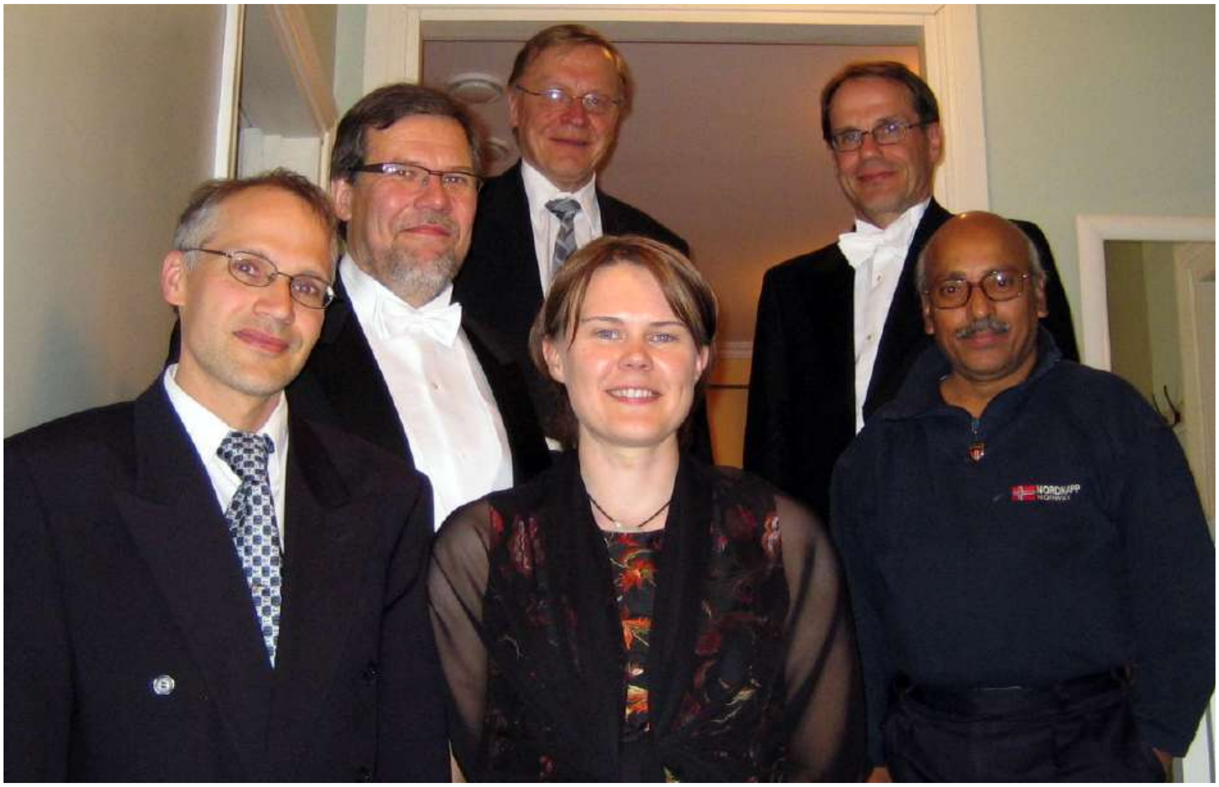


Figure 23: Laura's Thesis Defence Party, 29 June 2007. Laura surrounded by Arto Luoma, Tapio Nummi, Erkki Liski, Seppo Pynnönen and Bikas Kumar Sinha.

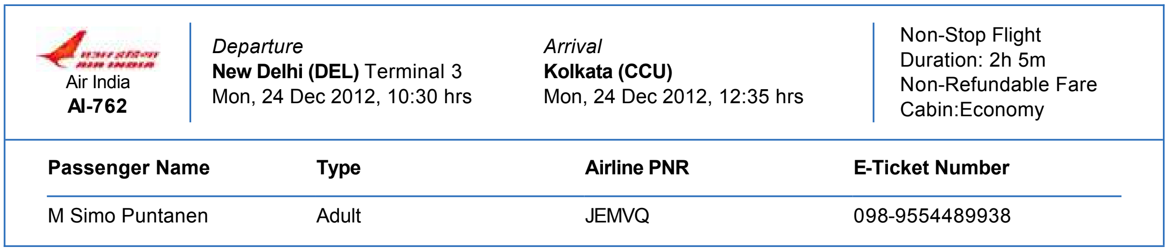
Figure 24: It was not full-time work every day . . . cruising with Erkki in Tapio's motorboat in Näsijärvi. Tampere is between two lakes, Näsijärvi and Pyhäjärvi. [Ph: Tapio Nummi]

6. 2012–13

Here is email from Karabi Nandy, 21 August 2012:

> Dear Friends in Far Finland,
>
> We are thrilled to announce the birth of our precious baby girl, **Rukmini**, who came to this world on Tuesday, August 14th at 9:32pm. We are all doing well and having a great time together, discovering new things every minute!
>
> Attached is a picture of our baby girl.
>
> Warmly, Karabi and Rajesh.

Figure 25: Sooo photogenic!! [Ph: BKS]



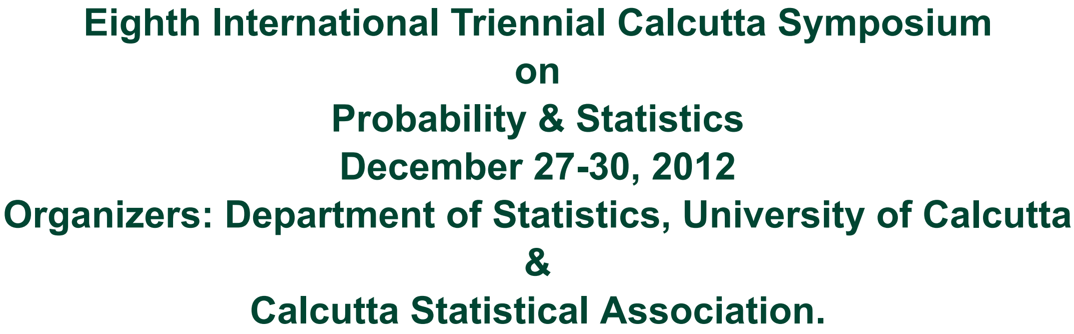
Air India
AI-762

Departure
New Delhi (DEL) Terminal 3
Mon, 24 Dec 2012, 10:30 hrs

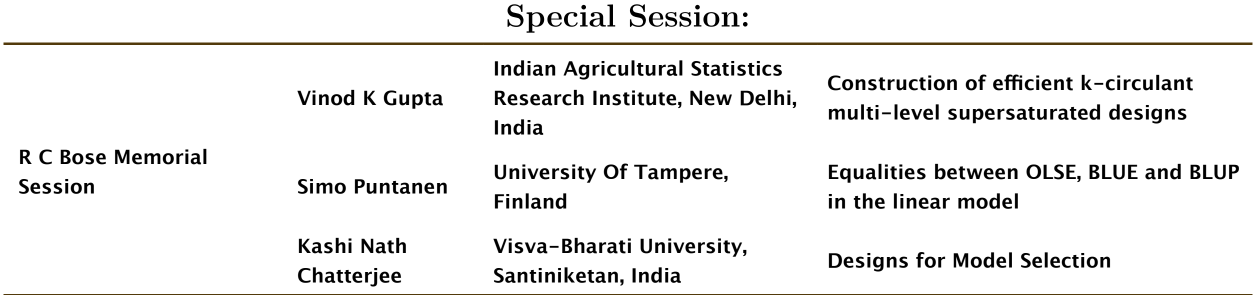
Arrival
Kolkata (CCU)
Mon, 24 Dec 2012, 12:35 hrs

Non-Stop Flight
Duration: 2h 5m
Non-Refundable Fare
Cabin:Economy

Passenger Name	**Type**	**Airline PNR**	**E-Ticket Number**
M Simo Puntanen	Adult	JEMVQ	098-9554489938



Eighth International Triennial Calcutta Symposium
on
Probability & Statistics
December 27-30, 2012
Organizers: Department of Statistics, University of Calcutta
&
Calcutta Statistical Association.



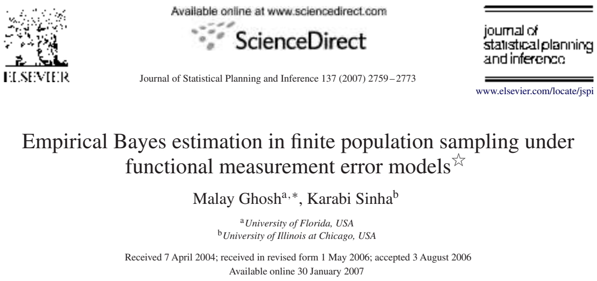
Special Session:

R C Bose Memorial Session	Vinod K Gupta	Indian Agricultural Statistics Research Institute, New Delhi, India	Construction of efficient k-circulant multi-level supersaturated designs
	Simo Puntanen	University Of Tampere, Finland	Equalities between OLSE, BLUE and BLUP in the linear model
	Kashi Nath Chatterjee	Visva-Bharati University, Santiniketan, India	Designs for Model Selection

Figure 26: 24 Dec 2012 ⟶ CCU, to attend the *Eighth International Triennial CCU Symposium*, 27–30 December 2012. Notice the speakers of the above Special Session! Program.

UNIVERSITY OF CALCUTTA
COLLEGE OF SCIENCE & AGRICULTURE
TARAKNATH PALIT SIKSHA PRANGAN
35, Ballygunge Circular Road, Kolkata-700019
INDIAN
STATISTICAL
INSTITUTE
203 B.T. ROAD KOLKATA 70010
228

Figure 27: Homegrounds of Bikas Kumar Sinha.



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 137 (2007) 2759–2773

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Empirical Bayes estimation in finite population sampling under functional measurement error models☆

Malay Ghosh[a,*], Karabi Sinha[b]

[a] University of Florida, USA
[b] University of Illinois at Chicago, USA

Received 7 April 2004; received in revised form 1 May 2006; accepted 3 August 2006
Available online 30 January 2007

Figure 28: Malay Ghosh, supervisor of Karabi Sinha, using his iPhone to book us a table in *Hotel Taj Bengal* for 31 December 2012 (with Kimmo Vehkalahti).

Here is email from Bikas Sinha, Wednesday, 2nd of January 2013, Kolkata:

> Simo,
>
> . . . we are meeting around 10 *this morning* . . .
>
> I will be accompanied by Prof. **Nripes Mandal** and Prof. Mrs **Manisha Pal** . . . both of CU Statistics Department.
>
> We will contact you at the Hotel Reception.
>
> Meanwhile my family joins me in extending a very happy New Year to you and your family and friends !
>
> Bikas

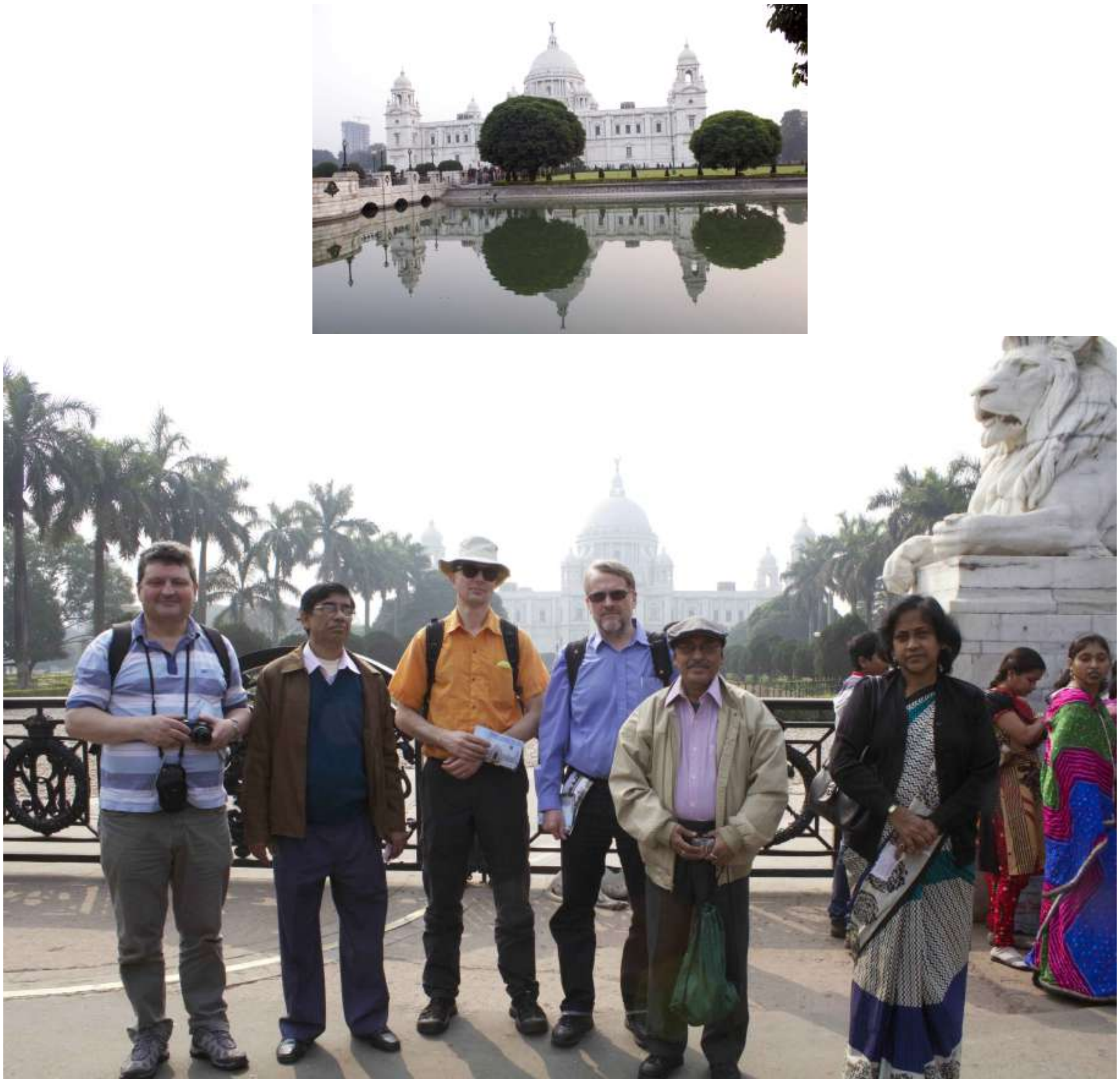


Figure 29: Augustyn Markiewicz, Nripes K. Mandal, Kimmo Vehkalahti, SJP, BKS, Manisha Pal. Victoria Memorial in the background. 2 January 2013.

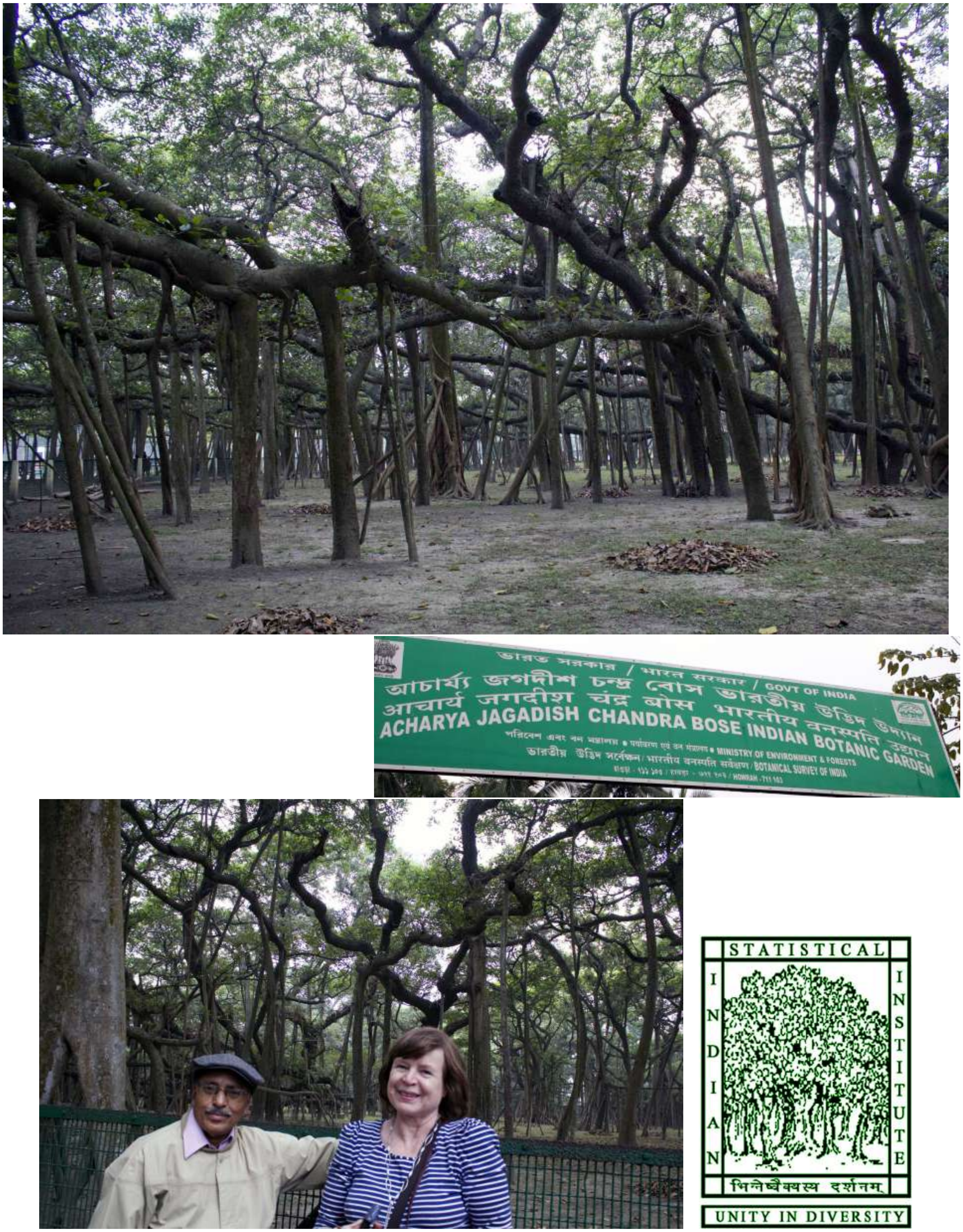

GOVT OF INDIA
ACHARYA JAGADISH CHANDRA BOSE INDIAN BOTANIC GARDEN
MINISTRY OF ENVIRONMENT & FORESTS
BOTANICAL SURVEY OF INDIA
HOWRAH
STATISTICAL
INDIAN
INSTITUTE
UNITY IN DIVERSITY

Figure 30: Bikas and Soile P. admiring the Great Banyan Tree (and the ISI logo), Kolkata Botanical Gardens, also known as the Acharya Jagadish Chandra Bose Indian Botanic Garden. 2 January 2013.

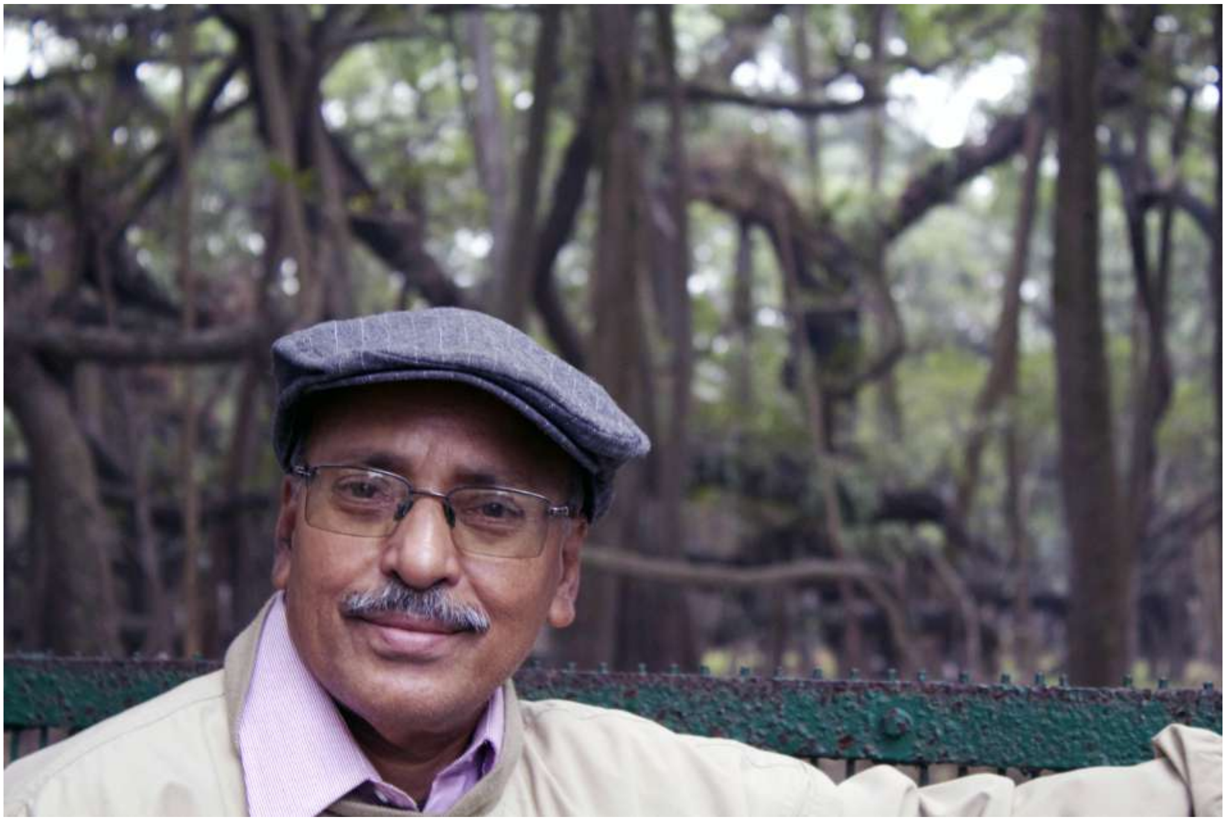


Figure 31: Bikas Kumar Sinha, 2 January 2013.

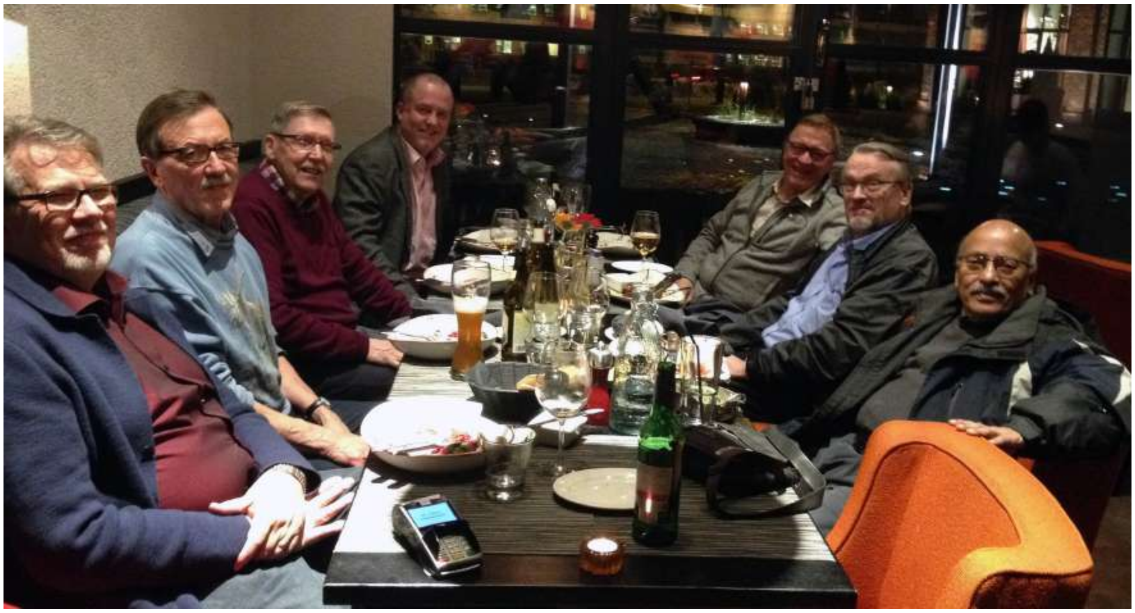
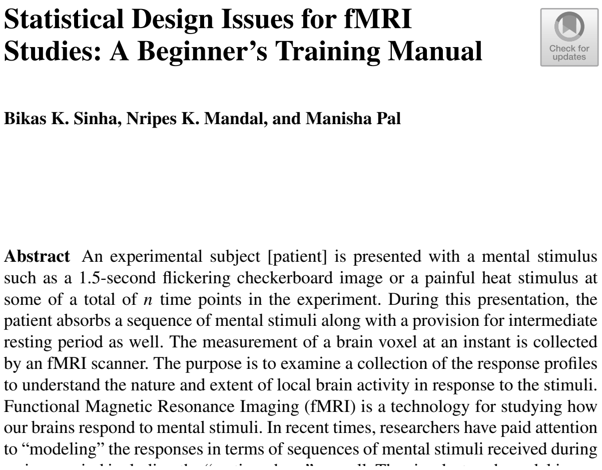


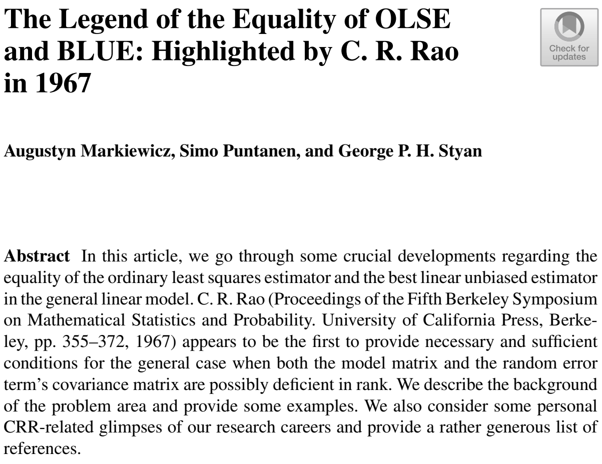
Figure 32: Bikas and the Statisticians Tamperensis, 30 October 2017. Tapio Nummi, Hannu Oja, Pentti Huuhtanen, Lasse Koskinen, Erkki Liski, SJP, BKS.



Statistical Design Issues for fMRI Studies: A Beginner's Training Manual

Bikas K. Sinha, Nripes K. Mandal, and Manisha Pal

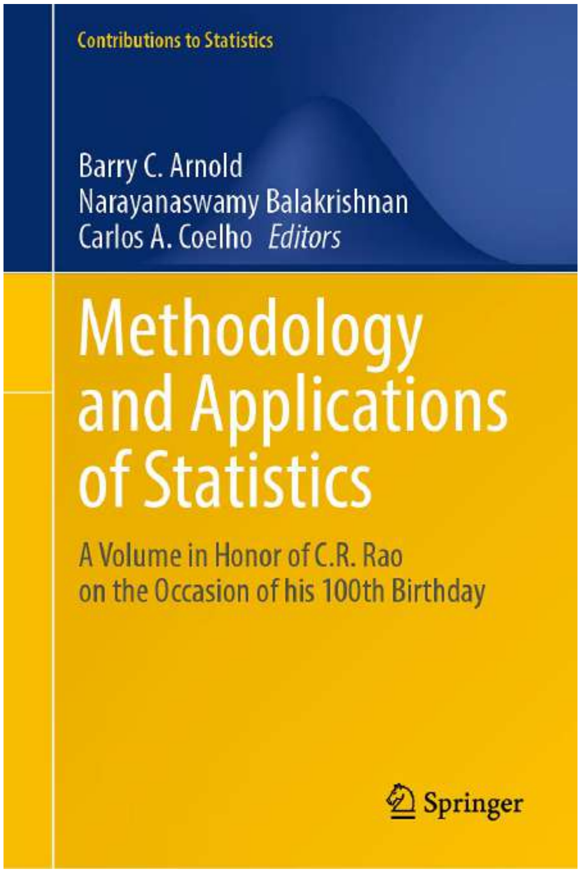
Abstract An experimental subject [patient] is presented with a mental stimulus such as a 1.5-second flickering checkerboard image or a painful heat stimulus at some of a total of n time points in the experiment. During this presentation, the patient absorbs a sequence of mental stimuli along with a provision for intermediate resting period as well. The measurement of a brain voxel at an instant is collected by an fMRI scanner. The purpose is to examine a collection of the response profiles to understand the nature and extent of local brain activity in response to the stimuli. Functional Magnetic Resonance Imaging (fMRI) is a technology for studying how our brains respond to mental stimuli. In recent times, researchers have paid attention to "modeling" the responses in terms of sequences of mental stimuli received during



The Legend of the Equality of OLSE and BLUE: Highlighted by C. R. Rao in 1967

Augustyn Markiewicz, Simo Puntanen, and George P. H. Styan

Abstract In this article, we go through some crucial developments regarding the equality of the ordinary least squares estimator and the best linear unbiased estimator in the general linear model. C. R. Rao (Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, pp. 355–372, 1967) appears to be the first to provide necessary and sufficient conditions for the general case when both the model matrix and the random error term's covariance matrix are possibly deficient in rank. We describe the background of the problem area and provide some examples. We also consider some personal CRR-related glimpses of our research careers and provide a rather generous list of references.



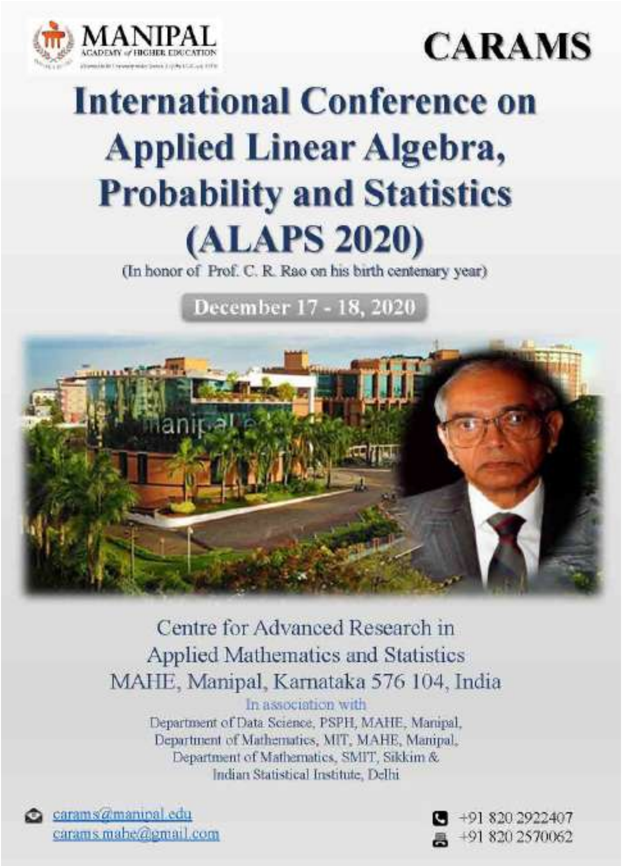
Contributions to Statistics

Barry C. Arnold
Narayanaswamy Balakrishnan
Carlos A. Coelho *Editors*

Methodology and Applications of Statistics

A Volume in Honor of C.R. Rao on the Occasion of his 100th Birthday

Springer



MANIPAL
ACADEMY of HIGHER EDUCATION

CARAMS

International Conference on Applied Linear Algebra, Probability and Statistics (ALAPS 2020)

(In honor of Prof. C. R. Rao on his birth centenary year)

December 17 - 18, 2020

Centre for Advanced Research in
Applied Mathematics and Statistics
MAHE, Manipal, Karnataka 576 104, India

In association with
Department of Data Science, PSPH, MAHE, Manipal,
Department of Mathematics, MIT, MAHE, Manipal,
Department of Mathematics, SMIT, Sikkim &
Indian Statistical Institute, Delhi

carams@manipal.edu
carams.mahe@gmail.com

☎ +91 820 2922407
+91 820 2570062

Figure 33: *Methodology and Applications of Statistics: A Volume in Honor of C.R. Rao on the Occasion of his 100th Birthday.* (B.C. Arnold, N. Balakrishnan and C.A. Coelho, Eds.), Springer, Cham. DOI of SiMaPa. DOI of MaPuSt.

In the ALAPS-2020 Conference, Bikas and I had an opportunity to perform online.

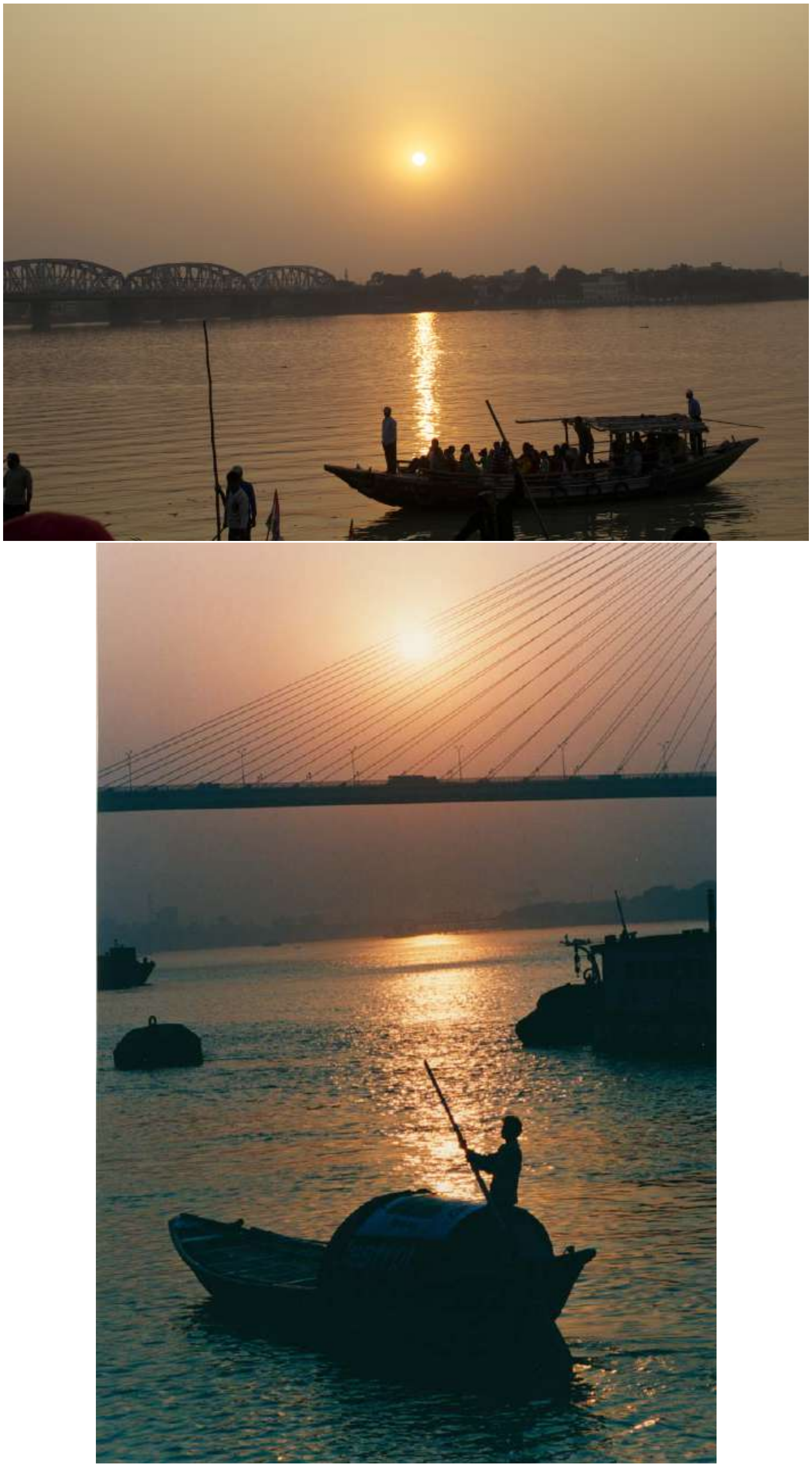


Figure 34: Sunset in Hooghly River and Howrah Bridge, Kolkata, December 1994.

Acknowledgements

This article is based on my Bikas Kumar Sinha Endowment Lecture (online) in the 24th Annual Conference of the Society of Statistics, Computers and Applications: Recent Advances in Statistical Theory and Applications, RASTA-2022, on 24 February 2022.

Sincere thanks go to Bikas Kumar Sinha and Family for most helpful assistance, in particular for providing the access to several photographs and the document Bikas-Sinha-Collaborators-PPT-Oct-2021.pptx

I'm also grateful to Laura Koskela, Erkki Liski and Tapio Nummi for their comments. The photographs from the archives of Bikas Sinha Family are denoted as [Ph: BKS]. One picture of Bikas Sinha is taken by Laura Koskela in Kolkata, July 2005, and one of Bikas Sinha and Erkki Liski by Tapio Nummi in Näsijärvi Lake in early 2000. The other photographs are taken by me.

Special thanks for the invitation for this article and the Endowment Lecture go to Professor Vinod K. Gupta and Bikas Kumar Sinha and Family and the RASTA-2022 Organisers.

References and Conference Links

Statistics and Applications (2020). Journal of the Society of Statistics, Computer and Applications, Vol. 18, No. 2. Special Issue "Challenges and Opportunities in Statistical Data Designing and Inference in the Emerging Global Scenario" in honour of twin statisticians (brothers) Bimal and Bikas Sinha on their 75th Birthday (16 March 2021). PDF

The Sinha Brothers (Bimal and Bikas) —The Twin Statisticians. Pages xi–xxx, in *Statistics and Applications*, Journal of the Society of Statistics, Computer and Applications, 2020, Vol. 18, No. 2. PDF

Bio-Sketch of Bikas K. Sinha (2001). *American Journal of Mathematical and Management Sciences*, 21, 3–4, 222–225, DOI

Bikas Sinha Collaborators PPT October 2021.pptx.

1983: The First International Tampere Seminar on Linear Statistical Models and their Applications, 30 August – 2 September 1983. Program. Proceedings.

2000: The 9th International Workshop on Matrices and Statistics, Hyderabad, 9–13 December 2000, celebrating C.R. Rao's 80th birthday Program.

2003: StatLin/LinStat Conference, Będlewo, Poland, 21–27 August 2003. Program.

2004: The 13th International Workshop of Matrices and Statistics, Będlewo, Poland, 19–21 August 2004, celebrating Ingram Olkin's 80th birthday. Program.

2006: The 6th International Triennial CCU Symposium, 29–31 December 2006. Program.

2007: Sujit Kumar Mitra (1932–2004) Memorial Meeting, Hyderabad, 6–7 January 2007. Report in *Image.*

2012: The 8th International Triennial CCU Symposium, 27–30 December 2012. Program.

2020: ALAPS-2020 Conference International Conference on Applied Linear Algebra, Probability and Statistics. Manipal Academy of Higher Education, MAHE, Karnataka State, South India, 17–18 December 2020. A virtual conference in honor of Professor C.R. Rao on his birth centenary year.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 31- 34

Alternatives to Global Hunger Index

A. K. Nigam

Consultant Advisor, Institute of Applied Statistics and Development Studies, (IASDS), Lucknow

Received: 25 May 2022; Revised: 06 June 2022; Accepted: 15 June 2022

Abstract

Recent research has established that Global Hunger Index (GHI) is flawed and is, therefore, not suitable for measuring the status of hunger. The work by Nigam and Co-workers (2016, 2018, 2019) showed that estimates of GHI have high upward bias. The extent of bias is substantial as hunger is a very small part of undernutrition and mortality.

Following the importance of the work, ICMR constituted an Expert Committee of eminent Statisticians, Pediatricians, and Public Health Experts to review the suitability of indicators used in the GHI. The Expert Committee in its unanimous Report observed that GHI does not measure hunger per se, and ranking countries using GHI is not appropriate. A White Paper by Padam Singh *et al*. (2021) also appeared in ICMR Journal highlighting these results. The findings were presented on request in different organizations like DWCD, NAAS, NITI Aayog and PMO and were approved.

There is, therefore, a need to examine alternative ways to measure hunger. Survey based USAID's Food and Nutrition Technical Assistance (FANTA, 2001) related Food Access Survey Tools (FAST, 2003) and Modified FAST (MFAST) by Institute of Applied Statistics and Development Studies (IASDS) have shown lot of promise. They allow us to assess hunger through lack of access and anxiety. This was validated by USAID in Bangladesh using FAST and in Banda district by IASDS using MFAST. Thus, survey-based approach needs to be further probed. To this, we may add data through dietary and consumption surveys for evaluating quality, quantity and frequency of consumption.

Key Words: Food insecurity; Hunger; Global Hunger Index; MFAST; Access; Anxiety; Dietary intakes.

1. Introduction

The end of hunger is one of the United Nations Sustainable Development Goals (SDGs). Measuring hunger itself is a complex issue. Hunger is defined in different ways by international organizations like Food and Agriculture Organization (FAO), World Health Organization (WHO), and World Food Program (WFP). These are: lack of food, food and nutrition insecurity, reduced food intakes with physical sensation caused by lack of food, constant worries where and when their next food will come from and chronic undernourishment. In long time hunger leads to undernutrition/mortality. While hunger leads to undernutrition, absence of hunger does not necessarily imply absence of undernutrition.

Corresponding Author: Arun K. Nigam
Email: dr_aknigam@yahoo.com

Global Hunger Index (GHI) was developed by International Food Policy Research Institute (IFPRI) to measure and compare hunger in different countries. It is the arithmetic mean of

- % undernourished population,
- % stunted children of under five years
- % wasted children of under five years, and
- % mortality rate of under five children.

All the indicators are standardized and assigned equal weights.

The work by Nigam and Co-workers (2016,2018,2019) showed that estimates of GHI have high upward bias and have many limitations. The extent of bias is substantial as hunger is a very small part of undernutrition and mortality.

Following the importance of the work, ICMR constituted an Expert Committee of eminent Statisticians, Pediatricians, and Public Health Experts to review the suitability of indicators used in the GHI. The Expert Committee in its unanimous Report observed that GHI does not measure hunger per se, and therefore, ranking countries using GHI is not appropriate. A White Paper by Padam Singh et al (2021) also appeared in ICMR Journal highlighting these results. The findings were presented on request in different organizations like DWCD, NAAS, NITI Aayog and PMO, and were approved. It was therefore necessary to look for alternative ways to measure hunger.

Before describing alternative ways, it may be desirable to discuss limitations of GHI.

- Estimates of GHI have an upward bias: hunger implies undernutrition though undernutrition does not imply hunger. Similarly, under-5 mortality may have reasons other than hunger.
- Upward bias in GHI has serious implications. It pushes up the hunger estimate. The extent of bias is likely to be substantial as hunger is most likely to be a small part of under-nourishment, under-nutrition and mortality.
- It is not possible to theoretically evaluate the bias because of the confounding between indicators and hunger. It may only be possible to evaluate the bias empirically through large data sets.
- It has the problem of multiple counts; Masset (2011) showed earlier that it has the problem of double counts.
- It ignores lack of access and anxiety though this is how hunger is defined.
- Arithmetic mean (with equal weightage) of indicators having extreme values.

2. Alternatives to GHI

‘Zero’ hunger is one of the very important goals of Sustainable Development Goals. We look for good quality survey-based data which capture access and anxiety through indicators having a direct bearing on hunger and reinforce it by dietary intakes of important food stuffs and other variables. For the first two types, we may have the following routes:

- survey based behavioral responses-based indicators on access and anxiety.

- dietary intakes of important food stuffs like cereals, pulses and fats & oils as the major sources of energy, protein and micronutrients.

Presently, for survey based behavioral responses available options are:

- FAO's Food Insecurity Experience Scale (FIES)
- USAID's Food and Nutrition Technical Assistance (FANTA) based Food Access Survey Tools (FAST) and its modified version (MFAST) by IASDS**.**

Recently, Ministry of Statistics and Programme Implementation (MoSPI) and FAO, jointly organized a virtual workshop on FIES to discuss various methodological issues at facilitating the design of pilot surveys that will be undertaken in selected States and districts to ascertain the applicability of this tool in the Indian context. This indicates that FIES is still in the pre-validation stage and therefore has to be ruled out for further consideration.

Among the other options, FANTA was validated using FAST in Bangladesh, 2003. MFAST was also validated by IASDS in Banda district, India, 2010. Both FAST and MFAST had 9 questions seeking behavioral responses for individuals/households experiencing food insecurity.

A study of 600 FAST and 8953 comparable households by Nigam (2018, 2019) indicated that access and anxiety are measurable. Results of the two are not much different.

The questions used in MFAST were:

1. The family ate few meals per day on a regular basis;
2. Obliged to eat non-preferred instead of preferred food;
3. Sometimes food stored in the house ran out and no cash to buy;
4. Worried frequently about where the next meal would come from;
5. Needed to purchase food frequently (because own production or purchased stores ran out);
6. Took food on credit from a local store;
7. Needed to borrow food from relatives/neighbors to make a meal;
8. Needed to borrow food in order to meet social obligations
9. Members of the household who had to skip the meal due to lack of food: (i) working adult, (ii) house-wife, (iii) both, (iv) elderly persons, and (v) children

Questions 3-9 together provide food insecure households with hunger. Question 9 gives individual level hunger. It reflects severe form of hunger. Questions 3-4 give anxiety.

A close look into 9 questions reveals that these can be remodeled in to the following 3-question module:

1. The family ate meals per day on a regular basis for the last 15 days.
2. Worried frequently (at least once in the last 15 days) about where the next meal would come from as the food stored in the house ran out and no cash to buy more
3. Had to take food on credit from a local store/relatives or neighbors (at least once in the last 15 days) to make a meal for the family or to serve a meal to guests or relatives

For value addition, we may ask about Members of household who had to skip the meal due to lack of food: (i) working adult, (ii) house-wife, (iii) both, (iv) elderly persons, and (v) children.

This type of MFAST can easily be regularly canvassed by agencies like National Statistical Office (NSO) in India and by similar survey agencies across the countries. It has the potential of being part of the consumption surveys of NSO.

More information can be added from secondary data on Public Distribution System (PDS) supply and related variables to assess the access component. Such data can be available from studies like food insecurity and food and nutrition insecurity Atlases which provide data on availability, access and absorption.

Dietary and consumption surveys should be conducted every 3 years. This would allow evaluation of hunger status at a low cost without much difficulty.

Acknowledgement

The author is grateful to Dr. Sheila Vir for making useful suggestions on the manuscript.

References

Food Access Survey Tools (FAST) (2003). Food and Nutrition Technical Assistance (FANTA) sponsored validation study in Bangladesh.

Food and Nutrition Technical Assistance Project (FANTA) (2001). Academy for Educational Development, USAID.

Masset, E. (2011). A review of hunger indices and methods to monitor country commitment to fighting hunger. For FAO Food Policy, **36**, S102-8.

Micro Level Hunger Mapping (2010) Project Report of Institute of Applied Statistics and Development Studies for Indian Council of Medical Research.

Nigam, A. K., Srivastava, R., Tiwari, P. P., Saxena, Reeta and Shukla, Shruti (2016). Hunger in gram panchayats of Banda district (U.P.) ~ A micro-level study. *Journal of the Indian Society of Agricultural Statistics*, **70**(**1**), 41-50.

Nigam, A. K. (2018). Global Hunger Index revisited. *Journal of the Indian Society of Agricultural Statistics*, **72**, 225–230.

Nigam, A. K. (2019). Improving Global Hunger Index. *Agricultural Research*, **8**, 132–139.

Singh, Padam, Kurpad, A. V., Verma, D., Nigam, A. K., *et al.* (2021). Global Hunger Index does not really measure hunger - An Indian perspective. White Paper in *Indian Journal of Medical Research*, published by Wolters Kluwer - Medknow for Director-General, Indian Council of Medical Research ijmr_2057_21_WP.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 35-43

Jayanta Kumar Ghosh: A Short Description of the Evolution of His Research Work

Minerva Mukhopadhyay
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Received: 05 June 2022; Revised: 20 June 2022; Accepted: 24 June 2022

1. Introduction

Professor Jayanta Kumar Ghosh was a legendary statistician. In 60 years of his research career, he worked in various areas of statistics including both the frequentist as well as the Bayesian paradigms. Prof. Ghosh's areas of interest included (but, not restricted to) classical statistical inference, sequential analysis, higher order asymptotics, rates of convergence, reliability theory, stochastic modelling, Bayesian asymptotics in parametric, non-parametric and semi-parametric inference, multiple decision theory, survey sampling, model selection, *etc.* He had applied his vast statistical knowledge to solve many real-life problems in geology, ecology, public health and environment studies. In his own words, *"many of the papers in the above areas solve long standing open problems or pioneer new areas, and have been cited often."* [1]

In his research career, Prof. Ghosh was honored with several awards including the Shanti Swarup Bhatnagar Award for Science and Technology in 1981, the Mahalanobis Gold Medal from Indian Science Congress Association in 1998, the P. V. Sukhatme Prize for Statistics in 2000, the Doctor of Science (D.Sc.) award by B. C. Roy Agricultural University, India in 2006, a Lifetime Achievement Award from the International Indian Statistical Association in 2010, and Padma Shri by the Government of India in 2014. He was selected as the 'President' of Statistics Section of the Indian Science Congress Association in 1991, and of the International Statistical Institute in 1993.

2. Early Career

Prof. Ghosh's research career began in early 1960s, while pursuing his Ph.D. dissertation under the supervision of Prof. Hari Kinkar Nandi. In his first two papers, he investigated the properties of Wald's sequential probability ratio test (SPRT) (Ghosh, 1960a,b). Prof. Ghosh continued to work on SPRT during 1960s and later in phases of his research life as well. Some pioneering contributions include formalization of the result of Stein establishing relationship between sufficiency and invariance (Hall *et al.* (1965)), the Ghosh-Pratt identity (Ghosh, 1961), *etc.*

[1] The quotations used in this manuscript are obtained from Prof. Ghosh's personal webpage: https://www.stat.purdue.edu/∼ghosh/.

Correponding Author: Minerva Mukhopadhyay
Email: minervam@iitk.ac.in

3. A Brief Description of Prof. Ghosh's Work in Classical Asymptotics

In Prof. Ghosh's words, "*my main work after this (SPRT) and till 1982 had been on higher order asymptotics*". His entire volume of work on classical asymptotics has been classified into seven broad categories in Clarke and Ghosal (2008). Here, we highlight only a few among them.

(i) *Bahadur representation of quantiles:* Prof. Ghosh relaxed the assumptions of Bahadur representation, which is a representation of the sample quantile as a sum of independent and identically distributed (i.i.d.) random variables, at the cost of a different ordered (in probability) remainder. The generalization to place in two aspects: first, from the existence of second order bounded derivative to the existence of first order (positive) derivative of the cumulative distribution function (c.d.f.), and second, from fixed dimension p to variable dimension $p_n = p + O\left(n^{-1/2}\right)$ (see Ghosh (1971)).

(ii) *Higher order asymptotics:* We split this topic into three sub-parts.

Edgeworth expansions. In his seminal work, Prof. Ghosh (with Prof. Rabi Bhattacharya) formalized the validity of r-th order Edgeworth expansion for smooth functionals of sample averages under appropriate moment conditions and some conditions on the associated characteristic functions (see Bhattacharya and Ghosh (1978, 1980)). Prof. Ghosh, along with his collaborators, also derived Edgeworth expansions for likelihood ratio type test statistics, which have non-normal limiting distributions (see, e.g., Chandra and Ghosh (1979)).

Second order efficiency and admissibility. Related to Edgeworth expansion are the second order efficiency and the second order admissibility type properties of estimators. The performance of asymptotically efficient estimators can be compared in the light of second order efficiency. Similarly, second order admissibility compares statistics using the second order risk. Along with his collaborators, Prof. Ghosh established pioneering results in comparing maximum likelihood estimator (MLE) and related estimators in terms of second order efficiency and admissibility (see, e.g., Ghosh and Subramanyam (1974); DasGupta and Ghosh (1983)).

Comparison of likelihood ratio, Wald and Rao's Tests. Comparison of the likelihood ratio test (LRT), Rao's and Wald's tests has been an important problem, which has drawn attention of many researchers in 1970s. Prof. Ghosh (along with Prof. Tapas Kumar Chandra) rigorously derived the asymptotic expansions of the distribution functions of LRT, Rao's and Wald's tests (see Chandra and Ghosh (1980)).

(iii) *Bartlett's correction.* A series of Prof. Ghosh's work hinges on Bartlett's correction. Asymptotic normality of LRT statistics with Bartlett's corrected error was proved for multidimensional settings in the seminal paper of Bickel and Ghosh (1990). This result was proved through a Bayesian argument, where Bartlett's approximation was applied to the limiting posterior distribution. Further studies in this route related to frequentist and Bayesian Bartlett's approximation can be obtained from Ghosh and Mukerjee (1991, 1992a).

(iv) *Neyman-Scott problem:* In the Neyman-Scott problem, one is interested in estimating the common parameter θ based on the sequence $\{X_n\}_{n\geq 1}$ of independent random variables with X_i having the density $f(\cdot, \theta, \xi_i)$. Prof. Ghosh (along with Prof. Bhanja) proposed asymptotically efficient estimators of θ under parametric and semiparametric setup depending on the nature of the sequence $\{\xi_n\}_{n\geq 1}$, considering it to be fixed or random following a common distribution. This work and related extensions can be found in Bhanja and Ghosh (1992a,b,c).

4. A Brief Description of Prof. Ghosh's Work on Bayesian Inference

In the later part of his research career, Prof. Ghosh contributed equally, if not more to Bayesian asymptotics. In his own views, his work in Bayesian asymptotics can be classified into four broad categories:

(i) *"The first related to theorems like posterior normality, necessary and sufficient conditions for a normalized posterior to converge to a non-degenerate distribution, etc."* Prof. Ghosh extended the Bernstein von-Mises (BvM) theorem, which shows asymptotic normality of the posterior distribution after proper centering and scaling, in various aspects. Extensions of the BvM theorem and related work include providing precise conditions for higher order expansion of the posterior distribution under the marginal distribution of the data (Ghosh *et al.*, 1982); providing a version of the BvM theorem where the posterior is calculated given the sample mean instead of the full sample (Clarke and Ghosh, 1995); proving posterior strong consistency under weak assumptions on the normalizing factors and for non-regular densities (Ghosh *et al.*, 1994; Ghosal *et al.*, 1995); obtaining first-order asymptotic approximation to the posterior distribution of the unknown change point θ in a change point problem (Ghosal *et al.*, 1999c), *etc.*

(ii) *"In the second set of problems I* (Prof. Ghosh) *try to derive priors for which posterior probabilities are close to frequentist probabilities in various senses. This is relevant for validating non-informative priors or constructing confidence intervals."* In this context, Prof. Ghosh introduced the idea of probability matching prior, where a prior is so chosen that the posterior joint cumulative distribution function (c.d.f.), or the highest posterior density regions of a standardized version of the parametric vector matches with the corresponding frequentist c.d.f. up to an order of $o(n^{-1/2})$ (Ghosh and Mukerjee, 1993b), or $o(n^{-1})$ (Ghosh and Mukerjee, 1993a; Datta and Ghosh, 1995). Other related objective priors like reference priors have been studied in Ghosh and Mukerjee (1992b), and a comparison of several objective priors has been done in Datta and Ghosh (1995).

(iii) *"In the third set of problems I* (Prof. Ghosh) *deal with Bayesian analysis of infinite dimensional problems like Bayesian survival analysis, Bayesian density estimation, Bayesian semiparametric,* etc. *A major concern here is the consistency of posterior and rate of convergence."* Prof. Ghosh made significant contribution in the area of Bayesian nonparametric, and has been one of the main architects in designing the modern asymptotic theory of Bayesian nonparametric. His work in Bayesian nonparametric started with the work Ghosh and Ramamoorthi (1995), co-authored with Prof. R. V. Ramamoorthi, where they studied the convergence of the posterior distribution towards the true underlying distribution in the context of survival data both in censored and uncensored cases (see also Ghosh *et al.* (1999)). Generally, the choice of an appropriate prior is crucial towards establishing posterior consistency for infinite-dimensional models. Prof. Ghosh has shown posterior consistency for the parameter of symmetry for any unknown symmetric distributions under a semiparametric setup using symmetrized Pòlya tree prior (Ghosal *et al.*, 1999a). Similarly, Pòlya tree priors and Dirichlet mixtures of a normal kernel were used in a regression context in Amewou-Atisso *et al.* (2003).

The above mentioned paper as well as the modern Bayesian asymptotic theory for infinite-dimensional models uses Schwartz's theorem (Schwartz, 1965), which is *the right tool for studying consistency* in infinite-dimensional problems. Implementation of the Schwartz's idea using a sieve based approach was used for density estimation by Dirichlet mixture of normal densities in Ghosal *et al.* (1999b), and with the logistic Gaussian process prior in Tokdar and Ghosh (2007).

Consistency is just the first step. Given consistency, one would be interested in results on rates of convergence, which has also been investigated in (Ghosal *et al.*, 2000).

(iv) Finally, Prof. Ghosh has *"made major progress in understanding Bayesian and Empirical Bayes model selection rules in high dimensional problems. I* (Prof. Ghosh) *have also been working on Bayes Testing, Model Selection in low dimensional problems and have thrown light on some recent as well as long standing asymptotic problems."* The early work on Prof. Ghosh on model selection dates back to 1975, when Ghosh and Subramanyam (1975) had formalized the idea of separated hypothesis in terms of the L_1 distance. Among the more recent works, generalizations of BIC for high-dimensional data are considered in Berger *et al.* (2003); Chakrabarti and Ghosh (2006). A pertinent review of application of model selection procedures like AIC, BIC and their comparison is provided in Chakrabarti and Ghosh (2011). In Mukhopadhyay and Ghosh (2003), the relative performance of several parametric empirical Bayes methods was compared with AIC and BIC using asymptotic results and simulation studies both under the $0-1$ as well as the prediction loss.

In Bayesian model selection (or, hypothesis testing), difficulties arise when improper non-informative priors are used to calculate the Bayes factors. Several methods have been proposed to remove these difficulties. Ghosh and Samanta (2002) discusses a unified derivation of some of these methods. Towards the implementation of Bayes factor in factor model, Dutta and Ghosh (2013) provides justification of the regularity conditions needed for path sampling.

In the context of multiple testing, Bogdan *et al.* (2007) compared the empirical Bayes approach with Benjamini-Hochberg method, focusing mainly on the 'sparse mixture' case (see also Bogdan *et al.* (2008)). Within a Bayesian decision theoretic framework, some asymptotic optimality properties of a large class of multiple testing rules were investigated in Bogdan *et al.* (2011). A comprehensive review of model selection and multiple testing can be obtained from Dutta *et al.* (2012).

5. Conclusion

Prof. Ghosh was equally interested in meaningful applications of statistics in scientific and industrial research (Sinha and Sinha (2017)). For example, Maiti *et al.* (2016) discusses discrepancies in data sources for estimating the household consumption expenditure to derive Gross Domestic Product (GDP) of India and provides simple implementable strategies to improve the estimation of GDP.

Apart from this, there are multiple research areas in which Prof. Ghosh has contributed significantly and yielded numerous elegant and insightful research articles which could not be covered in this short write up. An account of his personal and research careers can be obtained from Ramamoorthi (2018), and also from Dasgupta (2017).

6. Jayanta Kumar Ghosh Endowment Lecture

In the J. K. Ghosh endowment lecture, I presented my recent research work on category learning (see Mukhopadhyay *et al.* (2021)). An abstract of this work is given below:

Understanding how adult humans learn to categorize can shed novel insights into the mechanisms underlying experience-dependent brain plasticity. Drift-diffusion processes are popular in such contexts for their ability to mimic underlying neural mechanisms but require data on both category responses and associated response times for inference. Category response accuracies are, however, often the only reliable measure recorded by behavioral scientists to describe human learning. Building carefully on drift-diffusion models with latent response times, a novel biologically interpretable class of 'inverse-probit' categorical probability models is derived for such data. The model, however, presents significant identifiability and inference challenges. These challenges are addressed via a novel projection-based approach with a symmetry preserving identifiability constraint that allows working with conjugate priors in an unconstrained space. This model is adapted for group and individual level inference in longitudinal settings. Building again on the model's latent variable representation, an efficient Markov chain Monte Carlo algorithm is designed for posterior computation. The method's empirical performances are evaluated through simulation experiments. The method's practical efficacy is illustrated in applications to longitudinal tone learning studies.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, **9(2)**,291–312.

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. volume 112, pgs. 241–258. Model selection, model diagnostics, empirical Bayes and hierarchical Bayes, Special Issue II (Lincoln, NE, 1999).

Bhanja, J. and Ghosh, J. K. (1992a). Efficient estimation with many nuisance parameters. I. *Sankhyā Ser. A*, **54(1)**, 1–39.

Bhanja, J. and Ghosh, J. K. (1992b). Efficient estimation with many nuisance parameters. II. *Sankhyā Ser. A*, **54(2)**, 135–156.

Bhanja, J. and Ghosh, J. K. (1992c). Efficient estimation with many nuisance parameters. III. *Sankhyā Ser. A*, **54(3)**, 297–308.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Annals of Statistics*, **6(2)**, 434–451.

Bhattacharya, R. N. and Ghosh, J. K. (1980). Correction to: "On the validity of the formal Edgeworth expansion" [Annals of Statistics **6** (1978), no. 2, 434–451; MR **57** #10880]. *Annals of Statistics*, **8(6)**, 1399.

Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. *Annals of Statistics*, **18(3)**, 1070–1090.

Bogdan, M., Ghosh, J. K., Ochman, A., and Tokdar, S. T. (2007). On the empirical bayes approach to the problem of multiple testing. *Quality and Reliability Engineering International*, **23(6)**, 727–739.

Bogdan, M. g., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Annals of Statistics*, **39(3)**, 1551–1579.

Bogdan, M. g., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. Em *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 de *Inst. Math. Stat. (IMS) Collect.*, pgs. 211–230. Inst. Math. Statist., Beachwood, OH.

Chakrabarti, A. and Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *Journal of Statistical Planning and Inference*, **136(9)**, 2847–2872.

Chakrabarti, A. and Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of Statistics*, pgs. 583–605.

Chandra, T. K. and Ghosh, J. K. (1979). Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-square variables. *Sankhyā Ser. A*, **41(1-2)**, 22–47.

Chandra, T. K. and Ghosh, J. K. (1980). Valid asymptotic expansions for the likelihood ratio and other statistics under contiguous alternatives. *Sankhyā Ser. A*, **42(3-4)**, 170–184.

Clarke, B. and Ghosal, S. (2008). J. K. Ghosh's contribution to statistics: a brief outline. Em *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 de *Inst. Math. Stat. (IMS) Collect.*, pgs. 1–18. Inst. Math. Statist., Beachwood, OH.

Clarke, B. and Ghosh, J. K. (1995). Posterior convergence given the mean. *Annals of Statistics*, **23(6)**, 2116–2144.

DasGupta, A. and Ghosh, J. K. (1983). Some remarks on second-order admissibility in the multiparameter case. *Sankhyā Ser. A*, **45(2)**, 181–190.

Datta, G. S. and Ghosh, J. K. (1995). Noninformative priors for maximal invariant parameter in group models. *Test*, **4(1)**, 95–114.

Dutta, R., Bogdan, M., and Ghosh, J. K. (2012). Model selection and multiple testing—a Bayes and empirical Bayes overview and some new results. *Journal of Indian Statistical Association*, **50(1-2)**, 105–142.

Dutta, R. and Ghosh, J. K. (2013). Bayes model selection with path sampling: Factor models and other examples. *Statistical Science*, **28(1)**, 95–115.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999a). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference*, **77(2)**, 181–193.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999b). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **27(1)**, 143–158.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). On convergence of posterior distributions. *Annals of Statistics*, **23(6)**, 2145–2152.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1999c). Approximation of the posterior distribution in a change-point problem. *Annals of Institute of Statistical Mathematics*, **51(3)**, 479–497.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, **28(2)**, 500–531.

Ghosh, J. (1960a). On some properties of sequential t-test. *Calcutta Statistical Association Bulletin*, **9**, 77–86.

Ghosh, J. K. (1960b). On the monotonicity of the *OC* of a class of sequential probability ratio tests. *Calcutta Statistical Association Bulletin*, **9**, 139–144.

Ghosh, J. K. (1961). On the optimality of probability ratio tests in sequential and multiple sampling. *Calcutta Statistical Association Bulletin*, **10**, 73–92.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics*, **42**, 1957–1961.

Ghosh, J. K., Ghosal, S., and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. Em *Statistical decision theory and related topics, V (West Lafayette, IN, 1992)*, pgs. 183–199. Springer, New York.

Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *Journal of Multivariate Analysis*, **38(2)**, 385–393.

Ghosh, J. K. and Mukerjee, R. (1992a). Bayesian and frequentist Bartlett corrections for likelihood ratio and conditional likelihood ratio tests. *Journal of Royal Statistical Society Ser. B*, **54(3)**, 867–875.

Ghosh, J. K. and Mukerjee, R. (1992b). Non-informative priors. *Bayesian Statistics, 4 (Peñíscola, 1991)*, pgs. 195–210. Oxford Univ. Press, New York.

Ghosh, J. K. and Mukerjee, R. (1993a). Frequentist validity of highest posterior density regions in the multiparameter case. *Annals of Institute of Statistical Mathematics*, **45(2)**, 293–302.

Ghosh, J. K. and Mukerjee, R. (1993b). On priors that match posterior and frequentist distribution functions. *Canadian Journal of Statistics*, **21(1)**, 89–96.

Ghosh, J. K. and Ramamoorthi, R. V. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. Em *Analysis of censored data (Pune, 1994/1995)*, volume 27 de *IMS Lecture Notes Monogr. Ser.*, pgs. 95–103. Inst. Math. Statist., Hayward, CA.

Ghosh, J. K., Ramamoorthi, R. V., and Srikanth, K. R. (1999). Bayesian analysis of censored data. volume 41, pgs. 255–265. Special issue in memory of V. Susarla.

Ghosh, J. K. and Samanta, T. (2002). Nonsubjective Bayes testing—an overview. volume 103, pgs. 205–223. C. R. Rao 80th birthday felicitation volume, Part I.

Ghosh, J. K., Sinha, B. K., Joshi, S. N., and and (1982). Expansions for posterior probability and integrated Bayes risk. Em *Statistical decision theory and related topics, III, Vol. 1 (West Lafayette, Ind., 1981)*, pgs. 403–456. Academic Press, New York-London.

Ghosh, J. K. and Subramanyam, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhyā Ser. A*, **36(4)**, 325–358.

Ghosh, J. K. and Subramanyam, K. (1975). Inference about separated families in large samples. *Sankhyā Ser. A*, 37(4), 502–513.

Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Annals of Mathematical Statistics*, **36**, 575–614.

Maiti, P., Rao, T. J., and Ghosh, J. K. (2016). The Indian official statistical system revisited. *Sankhya B*, **78(2)**, 215–237.

Mukhopadhyay, M., McHaney, J. R., Chandrasekaran, B., and Sarkar, A. (2021). Bayesian semiparametric longitudinal inverse-probit mixed models for category learning. *arXiv preprint arXiv:2112.04626*.

Mukhopadhyay, N. and Ghosh, J. (2003). Parametric empirical Bayes model selection—some theory, methods and simulation. Em *Probability, statistics and their applications: papers in honor of Rabi Bhattacharya*, volume 41 de *IMS Lecture Notes Monogr. Ser.*, pgs. 229–245. Inst. Math. Statist., Beachwood, OH.

Ramamoorthi, R. (2018). Jayanta Kumar Ghosh (1937–2018).

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **4**, 10–26.

Sinha, B. K. and Sinha, B. K. (2017). Jayanta Kumar Ghosh. *Calcutta Statistical Association Bulletin*, **69(2)**, 129–131.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, **137(1)**, 34–42.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 45-55

Statistical Analysis of AEAD Ciphers for Transport Layer Security: An Experimental Approach

Jitendra Kurmi , Suresh Prasad Kannojia
Department of Computer Science, Lucknow University, Lucknow, UP, India -226007

Received: 12 May 2022; Revised: 22 June 2022; Accepted: 29 June 2022

Abstract

Nowadays, data security or information over the network is more critical as technology grows. To secure the data on network transport layer security SSL/TLS is used. TLS uses different ciphers of Authenticated Encryption and Associated Data (AEAD) to encrypt data during the transmission. In this paper, we perform statistical analysis of AEAD ciphers as AES_GCM (Galois Counter Mode), AES_SIV (Synthetic Initialization Vector), and AES_GCM_SIV with key size 128/256 bit based on encryption/decryption time for a message of block size 128/1024/2048/4096/8192 bit on Windows, Linux and Mac operating systems. We also measure the central tendency to determine where the most values fall in distribution, which will help us to identify the best suitable AEAD encryption algorithm. On the other hand, we use analysis of variance (ANOVA), the purpose of ANOVA is to measure the differences in strength of AEAD algorithms with different key sizes.

Key words: Authenticated encryption; SSL/TLS; AEAD; Security; Network.

1. Introduction

Cryptography algorithms are the building blocks of security and are widely used by many people and organizations worldwide. Encryption is used to protect the data from unauthorized access before transmitting the data by the sender, and decryption at receiver end. Cryptography already provides a lot of encryption/decryption algorithms. But most cryptographic algorithms do not provide confidentiality, integrity, and authenticity of data over the network. Authenticated Encryption (AE) is came into existence to deal with this problem. The AE and AEAD algorithms guarantee the confidentiality and integrity of data transmitted over the network. Various AEAD algorithms already exist, such as AES_GCM, AES_SIV, AES_GCM_SIV, Counter with CBC-MAC (CCM), Chacha20Poly1305, Deoxys, EAX, MGM, and Xsalsa20Poly1305. But in this paper, we are only considering three widely used AEAD algorithms for statistical analysis purpose.

This paper consist six sections as follows. Section 2 provides the related work that various research enthusiasts have done. The flow diagram represents the proposed methodology that is given in Section 3. Section 4 presents the experimental setup, and experimental result analysis has been done in Section 5. Finally, conclusion has given in Section 6.

2. Literature Review

In communication, message security by the asymmetric encryption technique guarantees both privacy and integrity, and it is considered authenticated encryption. Informally, such a

Corresponding Author: Jitendra Kurmi
Email: jitendrakurmi458@gmail.com

strategy ensures that no adversary can generate a ciphertext that decrypts to a valid plaintext, encryptions are indistinguishable from one another. Authenticated encryption was formerly accomplished using the "encrypt-then-authenticate" paradigm, which specifies that the resultant ciphertext should first be encrypted before applying a message authentication code. This strategy is sound, although it is inefficient most of the time. A more comprehensive analysis of composition approaches was carried out, taking into account a variety of alternatives and security goals. However, dedicated encryption modes optimized for high performance have been proposed in several circumstances.

Consider an encryption method that accepts a short secret key as input and creates a lengthy key stream as output, and used to encrypt the message bits by adding that is (modulo 2). The generated key stream must be 'pseudo-random, ' an essential security requirement for such a system. The distinction between true randomness and pseudo-randomness is a complicated one. On the other hand, the key stream should passes numerous well-known statistical tests for pseudo-randomness at the most fundamental level. Passing these tests is an essential but not sufficient requirement. The runs and autocorrelation tests are two well-known tests contains a more thorough list Knuth (2014). In an intriguing test has been developed that can be considered as universal bit generator by Maurer (1992).The chi-squared test is commonly used for determining a given sequence that follows a particular distribution or not. A brief discussion of this strategy may be found in Hell *et al*. (2009). Hypothesis testing is a practical statistical framework for analyzing cryptanalytic attacks. Often, an assault may be modeled as a hypothesis test to see if a parameter equals one of two possible values. Estimating the effort required for a successful attack becomes much more accessible when seen in this light. This approach provides a formal treatment Vaudenay (1996); Junod and Vaudenay (2003); Junod (2003). Various Statistical techniques have grown significantly to analyzing data from various power measurements is presented by Prouff *et al.* (2009). The AEAD algorithms are vulnerable to forgery and salamander attacks. An attack detection framework was proposed for authenticated ciphers to deal with this problem Kannojia and Kurmi (2021). A comparative analysis of various TLS libraries has been done, including authenticated encryption cipher, hashing, and public-key cryptography Kannojia and Kurmi (2021). Various TLS libraries have been analyzed based on supported languages, cryptographic token interface, thread safety, and CPU-assisted cryptography with Kannojia and Kurmi (2021).

3. Proposed Methodology

AEAD algorithms are used in TLS web servers for secure data transmission. The statistical randomness in encryption/decryption time of authenticated encryption algorithms with key sizes 128 bit and 256 bit are used to measure the central tendency. The purpose of measuring the central tendency is to determine where the most values fall in a distribution or not. In this paper, we proposed a statistical architecture to measure the central tendency of three AEAD algorithms with three operating systems such as Windows, Linux, and Mac shown in Figure 1.

To measure central tendency, we first implement these algorithms in python and measure the time taken by authenticated encryption ciphers to encrypt/decrypt a message of block size 128/1024/2048/4096/8192 bit with key sizes 128 bit and 256 bit. The time taken by AEAD ciphers is measured in milliseconds (ms), and further central tendency is calculated.

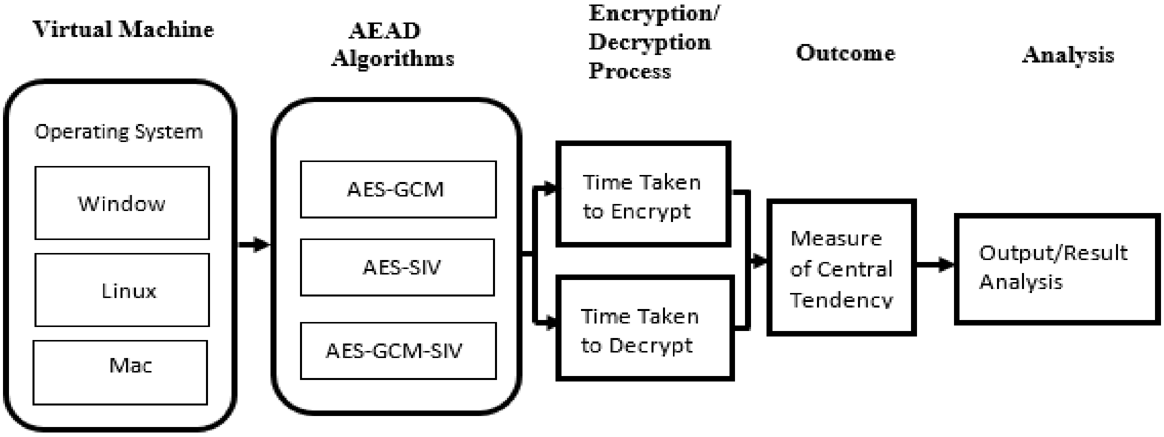


Figure 1: Proposed Architecture for Statistical analysis of AEAD ciphers

4. Experimental Setup

We performed the test on an Intel Core i3-3217U 1.80GHz CPU (4 cores) with 8GB of RAM, running Windows 10 professional machine, in a virtual environment with three operating systems Windows, Linux, and Mac.

5. Experimental Result and Analysis

5.1. Measurement of Central Tendency

The findings of the experimental results have been compiled and given in this sections, rounded upto two decimal points. The mean, standard deviation (σ_x), and standard error of the mean ($\overline{\sigma x}$) were calculated using equations (1), (2) to estimate the statistical error of the observed data (Xi) across the number of runs (N) using equation (3).

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} Xi \quad (1)$$

$$\sigma\text{x} = \sqrt{\frac{1}{N}} \sum_{i=1}^{N} (Xi - \overline{X})^2 \quad (2)$$

$$\overline{\sigma x} = \frac{1}{\sqrt{N}}\,\sigma_x \quad (3)$$

The measure of central tendency on windows, Linux, and Mac operating systems are shown in Tables 1, 2, and 3 and bar chart visualization method is used and presented in Figures 1 and 2.

The experimental result shown in Table 1 show that the time is taken to encrypt the message block by AEAD algorithm AES_GCM_SIV with the key size 128/256 bit is small AES_GCM_SIV 128/256 bit encryption algorithm is faster than AES_GCM and AES_SIV. The time to decrypt the message block by AEAD algorithm AES_GCM with key size 128 bit is faster than AES_SIV and AES_GCM_SIV. However, with a 256 bit key size, the AES_GCM_SIV is faster than AES_GCM and AES_SIV on the windows operating system.

Table 1: The measure of Central Tendency on Windows Operating System

	Ciphers	Time taken to Encrypt/Decrypt Message size (ms)					Mean	Std.Dev (%)	Error (%)
		128 bit	1024 bit	2048 bit	4096 bit	8192 bit			
Windows	AES_GCM (128 bit) Encryption	255	272	314	336	354	306.20	0.4191	0.1874
	AES_SIV (128 bit) Encryption	250	267	298	313	348	295.20	0.3857	0.1725
	AES_GCM_SIV (128 bit) Encryption	252	261	282	305	339	287.80	0.3518	0.1573
	AES_GCM (128 bit) Decryption	245	261	334	356	363	311.80	0.5502	0.2461
	AES_SIV (128 bit) Decryption	251	266	345	364	379	321.00	0.5855	0.2619
	AES_GCM_SIV (128 bit) Decryption	257	272	357	373	389	329.60	0.6073	0.2716
	AES_GCM (256 bit) Encryption	391	562	687	806	960	681.20	2.1895	0.9792
	AES_SIV (256 bit) Encryption	386	554	657	799	959	671.00	2.2048	0.9860
	AES_GCM_SIV (256 bit) Encryption	385	542	638	793	957	663.00	2.2130	0.9897
	AES_GCM (256 bit) Decryption	388	552	686	807	969	680.40	2.2435	1.0033
	AES_SIV (256 bit) Decryption	381	548	677	801	958	673.00	2.2277	0.9963
	AES_GCM_SIV (256 bit) Decryption	375	541	667	791	947	664.20	2.2072	0.9871

The experimental result shown in table 2 shows that the time is taken to encrypt the massage block by AEAD algorithm AES_GCM_SIV with the key size 128/256 bit is small AES_GCM_SIV 128/256 bit encryption is faster as compared to AES_GCM and AES_SIV. The time to decrypt the message block by AEAD algorithm AES_GCM with key size 128 bit is faster than AES_SIV and AES_GCM_SIV. However, with a 256 bit key size, the AES_GCM_SIV is faster than AES_GCM and AES_SIV on Linux operating system.

Table 2: The measure of Central Tendency on Linux Operating System

	Ciphers	Time taken to Encrypt/Decrypt Message size (ms)					Mean	Std.Dev (%)	Error (%)
		128 bit	1024 bit	2048 bit	4096 bit	8192 bit			
Linux	AES_GCM (128 bit) Encryption	249	269	317	339	358	306.40	0.4618	0.2065
	AES_SIV (128 bit) Encryption	246	266	302	312	347	294.60	0.3963	0.1772
	AES_GCM_SIV (128 bit) Encryption	247	267	279	299	332	284.80	0.3244	0.1451
	AES_GCM (128 bit) Decryption	241	268	327	365	354	311.00	0.5424	0.2426
	AES_SIV (128 bit) Decryption	247	272	336	366	369	318.00	0.5565	0.2489
	AES_GCM_SIV (128 bit) Decryption	252	277	349	368	392	327.60	0.6024	0.2694
	AES_GCM (256 bit) Encryption	385	556	682	816	951	678.00	2.2041	0.9857
	AES_SIV (256 bit) Encryption	381	552	659	809	953	670.80	2.2191	0.9924
	AES_GCM_SIV (256 bit) Encryption	379	549	642	803	948	664.20	2.2074	0.9872
	AES_GCM (256 bit) Decryption	383	546	680	811	972	678.40	2.2835	1.0212
	AES_SIV (256 bit) Decryption	379	542	676	805	975	675.40	2.3033	1.0301
	AES_GCM_SIV (256 bit) Decryption	370	538	671	797	967	668.60	2.3008	1.0290

The experimental result shown in table 3 shows that the time is taken to encrypt the massage block by AEAD algorithm AES_GCM_SIV with the key size 128/256 bit is small AES_GCM_SIV 128/256 bit encryption is faster as compared to AES_GCM and AES_SIV. The time to decrypt the message block by AEAD algorithm AES_GCM with key size 128 bit is faster than AES_SIV and AES_GCM_SIV. However, with a 256 bit key size, AES_GCM_SIV is faster than AES_GCM and AES_SIV on Mac operating system.

Table 3: The measure of Central Tendency on Mac Operating System

	Ciphers	Time taken to Encrypt/Decrypt Message size (ms)					Mean	Std. Dev(%)	Error (%)
		128 bit	1024 bit	2048 bit	4096 bit	8192 bit			
Mac	AES_GCM (128 bit) Encryption	253	267	317	332	355	304.80	0.4336	0.1939
	AES_SIV (128 bit) Encryption	252	270	302	317	349	298.00	0.3833	0.1714
	AES_GCM_SIV (128 bit) Encryption	256	266	285	309	342	291.60	0.3467	0.1551
	AES_GCM (128 bit) Decryption	241	263	337	359	366	313.20	0.5741	0.2568
	AES_SIV (128 bit) Decryption	253	264	345	363	373	319.60	0.5680	0.2540
	AES_GCM_SIV (128 bit) Decryption	261	269	352	371	383	327.20	0.5792	0.2590
	AES_GCM (256 bit) Encryption	387	565	684	810	962	681.60	2.2108	0.9887
	AES_SIV (256 bit) Encryption	389	556	668	803	960	675.20	2.2001	0.9839
	AES_GCM_SIV (256 bit) Encryption	382	547	641	798	952	664.00	2.2059	0.9865
	AES_GCM (256 bit) Decryption	384	556	683	802	966	678.20	2.2342	0.9992
	AES_SIV (256 bit) Decryption	382	552	677	801	962	674.80	2.2317	0.9981
	AES_GCM_SIV (256 bit) Decryption	379	548	669	795	951	668.40	2.2031	0.9852

The error percentage Encryption/Decryption of the AEAD algorithm on the different operating systems are measured and presented in Figure 2. The error percentage of encryption of AES_GCM_SIV 128 bit and encryption of AES_GCM 256 bit is less on Windows and Linux, while the error percentage of decryption AES_GCM 128 bit and AES_GCM_SIV 256 bit is less on windows as well as Linux. The error percentage of encryption of AES_GCM_SIV 128 bit and AES_GCM_SIV 256 bit is less on Mac, while the error percentage of decryption of AES_GCM 128 bit and AES_GCM_SIV 256 bit is less on Mac.

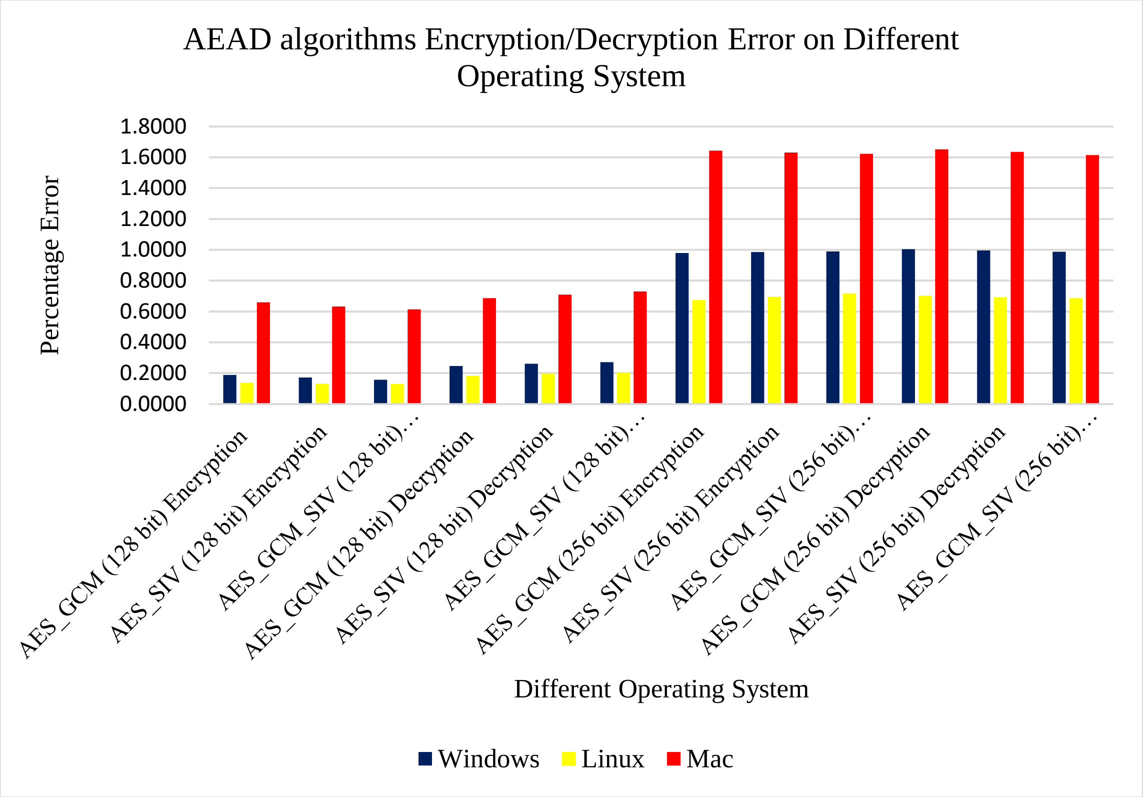


Figure 2: Comparison of AEAD algorithms Encryption/Decryption Error with three Operating System

The measure of Standard Deviation and central tendency of the AEAD algorithm on the different operating systems is measured and presented in Figure 3. The Standard Deviation and central tendency of encryption of AES_GCM_SIV 128 bit and encryption of AES_GCM 256 bit is slight on Windows and Linux, while the Standard Deviation and central tendency of decryption AES_GCM 128 bit and AES_GCM_SIV 256 bit is slight on windows as well as Linux. The Standard Deviation and central tendency of encryption of AES_GCM_SIV 128 bit and AES_GCM_SIV 256 bit is slight on Mac, while the Standard Deviation and central tendency of decryption AES_GCM 128 bit and AES_GCM_SIV 256 bit is slight on Mac.

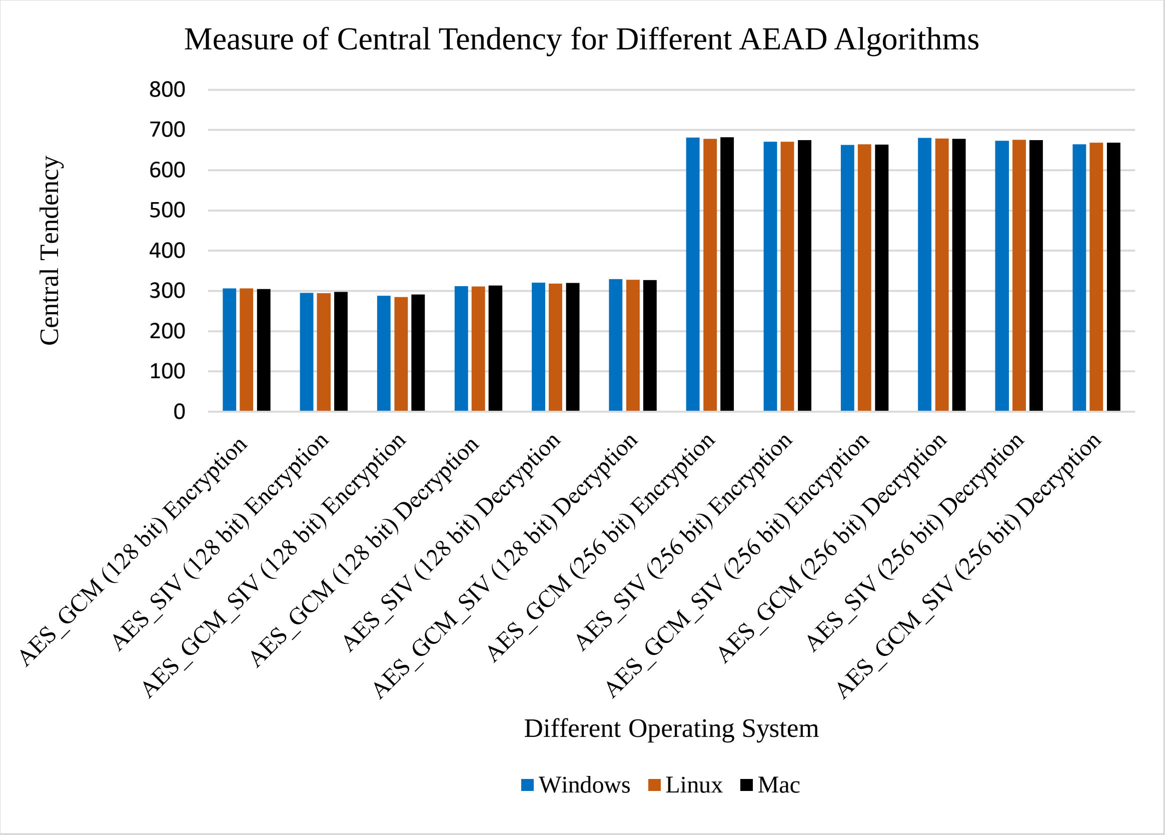


Figure 3: Comparison of Measure of central tendency of different AEAD algorithms with three operating system

5.2. Two Way ANOVA

The two way ANOVA (Analysis of Variance), also known as two-factor ANOVA used to determine if two or more samples have the same “mean” or average. Anova is a technique of understanding the variance of variables. The two way ANOVA is measured on the basis of AEAD algorithm encryption/decryption with two key size 128-bit as type 1 and 256-bit as type 2 from Table 1. It makes it possible to calculate how much a particular variable affects the final result. Anova technique does this by eliminating or confirming the null hypothesis. A null hypothesis means that there exists no relationship at all between the two entities under observation. The significance of a particular variable or entity is calculated by comparing the values with the overall impact on the target value. Anova requires a certain number through which it can analyze the null hypothesis that we pose at the start of the analysis. The three critical values for this calculation are F ratios and F-critical, with some significance values. For example, X’s significance will be more on A, if even a small change in X can affect in changing the value of A. The F ratios are calculated by the Mean sum of squares of an entity and the mean sum of residuals squares. The mean sum of squares is calculated by dividing the mean sum of squares by the degree of freedom. The degree of freedom is the number of possible cases of the nominal variable, minus one. F critical is based on the significance values. F ratios are calculated manually through the process explained above. The validity of the hypothesis is dependent on the values of F ratios and F critical. Here are the two cases:

- If the F-critical > F ratio, then the hypothesis holds, and there is no relation between the variables under observation

- If the F-critical < F ratio, then the hypothesis can be declared invalid, and in turn, supports the idea that the variables affect each other.

The purpose of ANOVA is to measure the differences in strength of AEAD algorithms with different key sizes.

Null Hypothesis: The null hypothesis states that there is no relationship between two population parameters, i.e., independent and dependent variables. If the hypothesis shows a relationship between the two parameters, the outcome could be due to an experimental or sampling error. However, if the null hypothesis returns false, there is a relationship in the measured phenomenon. The null hypothesis is helpful because it can be tested to conclude whether or not there is a relationship between two measured phenomena. It can inform the user whether the results obtained are due to chance or manipulating a phenomenon. Testing a hypothesis sets the stage for rejecting or accepting a hypothesis within a certain confidence level.

The overall average for the two different key sizes (128-bit and 256-bit) in the data is shown in Table 4. The difference is about 363.5333. The average for 5 different message sizes has been shown in a total average of two factors with replication. The averages for message sizes 128-bit, 1024-bit, 2048-bit, 4096-bit and 8192-bit are 318, 408.1667, 495.1667, 570.3333, and 660.1667.

Suppose the key size from the different message sizes all works the same on both key sizes (type 1 and type 2). In that case, the averages for the individual groups should follow the same patterns: The average for key size (type 1) should be higher, and the average for message size 8192-bit should be higher.

The group averages show a different pattern than the overall averages for the two factors. Message size 8192-bit‘s average is higher than the other four because there is a more significant difference between the message sizes for the second key size (type 2).

The comparison of averages should prepare us for what to expect about the null hypothesis for two-way ANOVA that the factors do not affect the response variable.

Table 4: The measure of Analysis of variance using ANOVA: two-factor with replication

Anova: Two-Factor With Replication						
SUMMARY	**128-bit**	**1024-bit**	**2048-bit**	**4096-bit**	**8192-bit**	**Total**
type 1 (128-bit)						
Count	6	6	6	6	6	30
Sum	1510	1599	1930	2047	2172	9258
Average	251.6667	266.5	321.6667	341.1667	362	308.6
Variance	17.46667	24.3	827.4667	776.5667	361.6	2228.179
type 2 (256-bit)						
Count	6	6	6	6	6	30
Sum	2306	3299	4012	4797	5750	20164
Average	384.3333	549.8333	668.6667	799.5	958.3333	672.1333
Variance	31.86667	62.56667	357.0667	43.1	49.46667	40631.22

Total						
Count	12	12	12	12	12	
Sum	3816	4898	5942	6844	7922	
Average	318	408.1667	495.1667	570.3333	660.1667	
Variance	4822.545	21933.42	33377.24	57664.24	97172.33	
ANOVA						
Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Sample	1982347	1	1982347	7769.442	1.57E-56	4.03431
Columns	860602.3	4	215150.6	843.2427	3.33E-45	2.557179
Interaction	369563.1	4	92390.77	362.1085	3.05E-36	2.557179
Within	12757.33	50	255.1467			
Total	3225270	59				

For the two-way ANOVA, our largest p-value is about 1.57 * 10^{-56}. That is much smaller than the traditional cutoff value for statistical significance of 0.05.

Because the p-value for the interaction is small, we cannot make a simple statement that one key size leads to a higher strength in terms of robustness.

The hypothesis test confirms, what we might have expected from the examination of averages: The effect of the different AEAD algorithms depends on the key size (128-bit and 256-bit).

6. Conclusion

Authenticated encryption algorithms are building blocks of secure communication over the network. In this paper, statistical analysis of three different algorithms with the key size 128/256 bit for a message of block size 128/1024/2048/4096/8192 bit with three different operating systems in a virtual environment has been observed. The overall performance of AES_GCM_SIV 128/256 bit encryption is faster than AES_GCM, AES_SIV and has a slight percentage error concerning central tendency. Here AES_GCM with key size 128 bit is faster than AES_SIV and AES_GCM_SIV and has a minor percentage error. However, decryption of AES_GCM_SIV 256 bit is faster than AES_GCM and AES_SIV and have minor percentage error on almost every operating system. So the AES_GCM_SIV 128/256 bit is the best encryption algorithm than AES_GCM and AES_SIV, while decryption of AES_GCM 128 bit and AES_GCM_SIV 256 bit are the better algorithms. From the result ANOVA two factor with replication, we conclude that the P-value for the interaction is small. So, we cannot make a simple statement that one key size leads to a higher strength in terms of robustness. The statistical analysis of other AEAD algorithms also can be compared and fine-tuned for better results in the future.

References

Hell, M., Johansson, T. and Brynielsson, L. (2009). An overview of distinguishing attacks on stream ciphers. *Cryptography and Communications*, **1(1)**, 71-94.

Junod, P. and Vaudenay, S. (2003). Optimal key ranking procedures in a statistical cryptanalysis. *In International Workshop on Fast Software Encryption, Springer, Berlin, Heidelberg*, 235-246.

Junod, P. (2003). On the optimality of linear, differential, and sequential distinguishers. *In International Conference on the Theory and Applications of Cryptographic Techniques, Springer, Berlin, Heidelberg*, 17-32.

Kannojia, S. P. and Kurmi, J. (2021). Attack Detection Framework for Authenticated Encryption cipher: An Experimental Approach. *Shodh Sarita*, **8(29)**, 210-215.

Kannojia, S. P. and Kurmi, J. (2021). Comparative Study of SSL/TLS Cryptographic Libraries. *International Journal of Innovative Research in Science, Engineering and Technology*, **10(8)**, 11658-11662.

Kannojia, S. P. and Kurmi, J. (2021). Analysis of Cryptographic Libraries(SSL/TLS). *International Journal of Computer Sciences and Engineering*, **9(9)**, 59-62.

Knuth, D. E. (2014). Art of computer programming. *Semi-Numerical Algorithms. Addison-Wesley Professional*, Volume **2**.

Maurer, U. M. (1992). A universal statistical test for random bit generators. *Journal of Cryptology*, **5(2)**, 89-105.

Prouff, E., Rivain, M. and Bevan, R. (2009). Statistical analysis of second order differential power analysis. *IEEE Transactions on computers*, **58(6)**, 799-811.

Vaudenay, S. (1996, January). An experiment on DES statistical cryptanalysis. *In Proceedings of the 3rd ACM Conference on Computer and Communications Security* (139-147).

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 57-68

Computer Simulation: Some Views on Model Development, Experimentation, and Robust Design

K. C. James
Department of Statistics, Cochin University of Science and Technology, Kerala, India

Received: 07 June 2022; Revised: 11 July 2022; Accepted: 13 July 2022

Abstract

A computer simulation is a computation that emulates the behaviour of some real or conceptual systems over time. We conduct experimentation with such models to understand the behaviour of a system. Simulation models are widely used in modern scientific research, education, industry and manufacturing, and public policy matters. These models tend to be extremely complex, often with many factors and sources of uncertainty. The complexity reflected in the system simulation models is characterized by the presence of entity elements that are dynamically created, asynchronous interactions between the entities, the use of shared resources, and connectivity between the entities. Conceptual modelling is a very relevant task in simulation modelling, but is often neglected by analysts. Simulation itself does not serve as an optimization technique. Computer experiment design principles differ from physical experiment design principles, and the three concepts of blocking, replication, and randomization are inessential or irrelevant to computer experiment design. In this paper, a few ideas on how to develop discrete-event simulation models and perform designed experiments are discussed, which helps in better solutions for the analysts. The focus is on some recent developments in the field of simulations which include ideas of visual analytics, data farming, knowledge discovery, and robust design.

Key words: Computer simulation; Conceptual modelling; Visual analytics; Knowledge discovery; Robust design.

1. Introduction

Simulation involves building a model that mimics the behavior of a system, experimenting with the model to create observations of these behaviours, and attempting to comprehend, summarise, and/or generalise these behaviours. Simulation also entails testing and comparing various designs, as well as validating, explaining, and supporting simulation outcomes and research recommendations in many cases. Simulations can also be classified based on how they are implemented. The implementation methodologies for continuous system simulation, Monte Carlo simulation, discrete-event simulation (DES), hybrid simulation, and agent-based simulation are all different.

Simulation has several advantages. Many integrated operations systems are subject to both variability and complexity (combinatorial and dynamic). Because it is difficult to anticipate the performance of systems that are subject to any one of variability, interconnectedness, or complexity, predicting the performance of operations systems that are potentially exposed to all three is extremely difficult, if not impossible.

Corresponding Author : K C James
Email: jamesmech@cusat.ac.in

Simulation models, on the other hand, can explicitly depict a system's unpredictability, interconnection, and complexity. As a result, a simulation can be used to anticipate system performance, evaluate various system designs, and assess the impact of different designs and policies on system performance.

1. Simulation allows researchers to investigate and experiment with the internal interactions of a complex system or a subsystem within one.
2. Informational, organisational, and environmental changes can be simulated, and the impact on the model's behavior can be determined.
3. Because simulation resembles what happens in an actual system or what is perceived for a system in the design stage, it appeals to clients instinctively.

A simulation's output data should be identical to the outputs that may be recorded from the real system. Furthermore, theoretically solvable models can be used to create a simulation model of a system that does not rely on dubious assumptions (such as the same statistical distribution for every random variable). Simulation is frequently the technique of choice in problem-solving for these and other reasons. Simulation models, unlike optimization models, are "run" rather than "solved." The model is run and the simulated behaviour is evaluated given a specific set of input and model variables.

Computer simulation is applied in a large number of industrial systems that include

- Manufacturing systems
- Public systems: health care, military, natural resources
- Transportation systems
- Construction systems
- Restaurant and entertainment systems
- Business process reengineering/management
- Food processing
- Computer system performance

In a recent attempt, Discrete event simulation (DES) is used even to help livestock farmers, by simulating potential growth strategies and observing the impact in relation to existing farm processes (Gittins *et al*., 2020). To know more about a wide variety of application areas of simulations, readers may refer to any Winter Simulation Conference proceedings of recent years.

In the following sections, some considerations required for effective simulation modelling are discussed. In section 2 few ideas of conceptual modelling, a largely forgotten area by many modellers are presented. Section 3 includes a few thoughts on simulation experimentations and some recent developments in this area. In section 4 few ideas of robust design relevant for simulations are discussed.

2. Conceptual Modelling, The Soft Operations Research Exercise

Simulations involve a number of steps as summarised in figure 1

1. A conceptual model: a description of the model that is to be developed

2. A computer model: the simulation model implemented on a computer
3. Improvements and/or understanding: derived from the results of the experimentation
4. An improvement in the real world: obtained from implementing the improvements and/or understanding gained

Although effective conceptual modeling is vital, it is also the most difficult and least understood stage in the modeling process (Law, 2015). Conceptual modelling is a very relevant task in simulation modelling, but is often neglected by analysts. The author believes that many of statistical modelling tasks also require good conceptual modelling exercises, but largely this step is ignored. It can be treated as a soft operations research exercise and a good conceptual model significantly enhances the accuracy and acceptability of the computer model. It minimizes the likelihood of incomplete, unclear, inconsistent, and wrong requirements. It helps build the credibility of the model and forms the basis for model verification and guides model validation. It helps experimentation by expressing the modeling objectives, and model inputs and outputs.

Conceptual modelling consists of the following sub-activities (Robinson, 2011):

- Develop an understanding of the problem situation
- Determine the modelling objectives
- Design the conceptual model: inputs, outputs, and model content

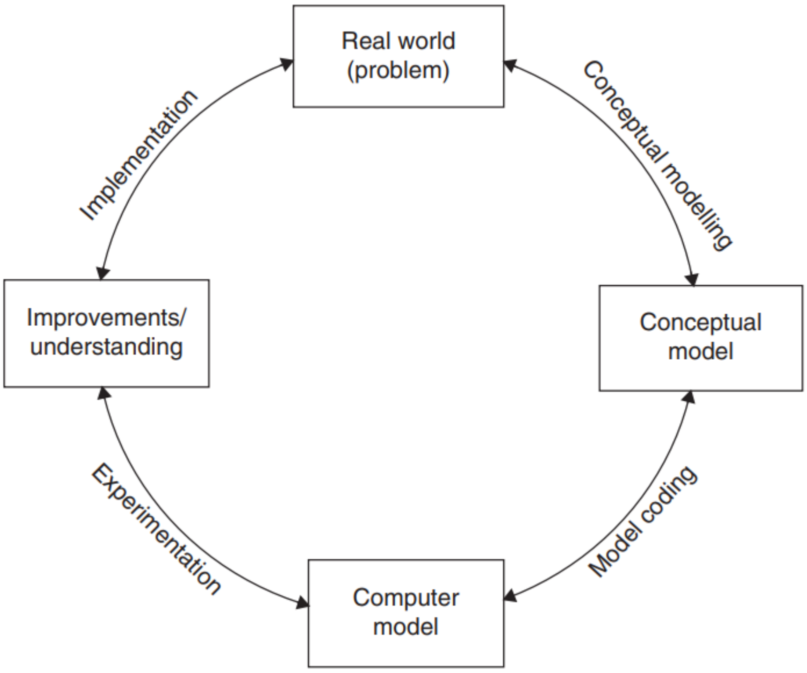


Figure 1: Simulation model development process (Source: Robinson, 2014)

For effective simulation, during conceptual modellers consideration should be given to

- Subject matter experts
- Organizing and structuring knowledge
- Adoption of “soft” OR approaches (Rosenhead and Mingers, 2004)
- Dimensions for determining the performance of a conceptual
- Identifying, adapting, and developing modeling frameworks

- Model simplification methods
- Model representation methods
- Use of software engineering techniques

A model should be created for a specific reason, and its validity should be determined in relation to that purpose. A constructed model should typically be a parsimonious model, which means it is as simple as feasible while yet accomplishing its goal. Furthermore, a model's accuracy (also known as model fidelity) should normally be limited to what is required to meet the model's function or purpose. If the goal of a model is to answer a range of questions, the model's validity must be assessed separately for each question. Soft operations research/problem structuring approaches have been used by OR practitioners for many years. When problem structuring approaches are used in combination with analytical approaches such as computer simulation, it is sensible to regard the two approaches as complementary(Pidd, 2007).

Model developers and users, decision-makers who use information derived from model results, and persons who are affected by model-based decisions are all interested in whether a model is valid. Strict verification and validations of conceptual models and computer models are essential to develop confidence in the customer's mind.

3. Experimentation and Knowledge Discovery

The focus of early experimental designs was mostly on physical experiments. Traditionally, simulation experts conduct experiments on the computer model for predetermined system specifications focusing on single model aspects and specific analysis questions. Modellers compare multiple system configurations and choose the one which presents the best system performance. Computer experiment design principles differ from physical experiment design principles, and the three concepts of blocking, replication, and randomization are inessential or irrelevant to computer experiment design. The "space-filling property" is commonly used in deterministic computer models based on partial differential equations to cover the experimental region with design points. Such analysis are nowadays finding value even in stochastic discrete event simulations.

Recent developments in big data analytics have also influenced experimentation and analysis of simulation. In the following subsections, some of these developments are presented. Subsection 3.2 discuss how big data concepts have a different flavor in simulations; 3.3 discusses data farming; 3.4 describes briefly various tools used in the knowledge discovery process.

3.1. Experimentation: traditional simulation vs. knowledge discovery

Traditionally, simulation studies make use of several runs on predetermined experimental scenarios. Recent developments in the "Knowledge Discovery" process as applied to simulations focus on the use of large sets of experimental data from simulations to find out hidden patterns for useful interpretations of the system. Table 1 below differentiates these approaches. Figure 2 shows a procedure for knowledge discovery (Feldkamp *et al.*, 2015a) which make use of concepts of big data, data farming, data mining and visualisations.

Table 1: Traditional Simulation Vs Knowledge discovery

Traditional Simulation	Knowledge discovery
• Project goals formulated beforehand • Simulation study is carried out by comparing predetermined scenarios that the user already had in mind before. • The target function has to be set up beforehand for optimization. • Analyst usually takes an educated guess which input parameters (factors) might be influential on the project scope.	• Use a combination of data mining and visual analysis • Find hidden and potentially interesting • Knowledge generated outside of prior defined project scopes

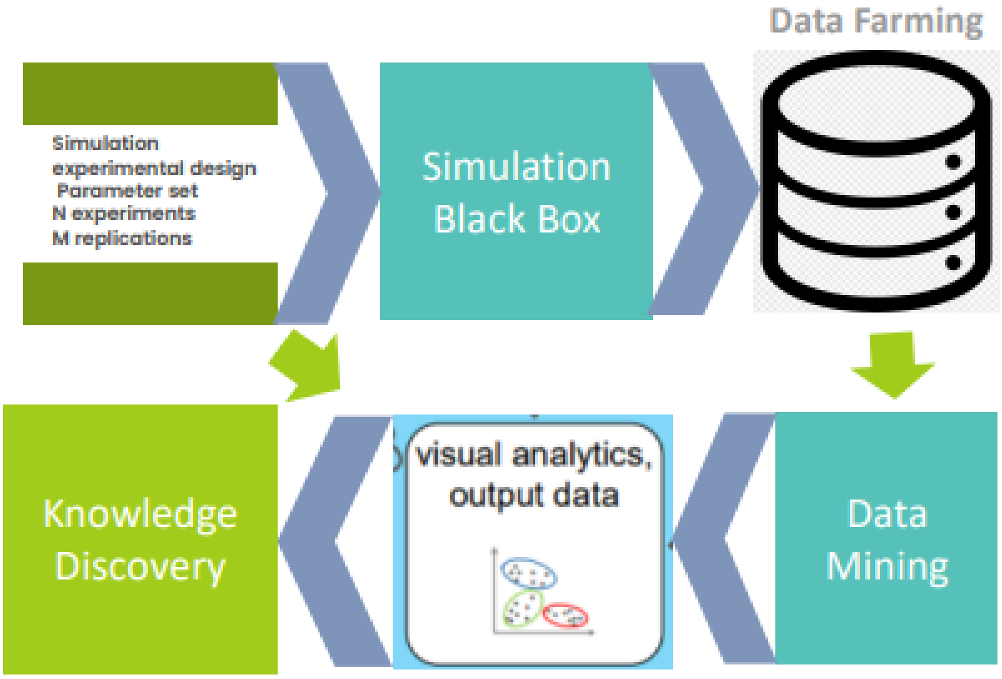


Figure 2: Knowledge discovery Process

3.2. Big data: the 3 (or more) V's have a different flavor in simulation

The term "Big Data" refers to a large amount of data that can't be stored or processed by conventional data storage or processing equipment. Big Data is generated on a massive scale, and it is being processed and analysed by many global corporations in order to unearth insights and enhance their businesses. Simulation experiments can generate a huge amount of data if the experiment considers a large number of factors and levels. This data can be considered as having the "V"(volume, velocity, veracity, etc) characteristics of conventional big data. However, we can find some subtle differences as shown in Table 2.

Table 2: Big data vs simulation data: The 3 (or more) V's have a different flavor in simulation

Big Data Characteristics	Big Data in Simulation
Analysts usually will not have any control over the data. Data may come from sources like information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, *etc*	Velocity and volume are partially controlled by the analyst. Analyst determines how to run the simulation (*e.g.,* on a single core or on a high-performance computing cluster), how much data to output (*e.g.*, aggregate statistics at the end-of-run, batch statistics, or full time-series output) for each performance measure, and the number of performance measures to study.
Data is not under the control of analysts and data may be Structured Data, Semi-Structured Data, or Unstructured Data. Most big data contain lots of missing data, errors, and incompatible data formats	The variety does not include many of the problems that we find with observational data (*e.g.*, incompatible data formats, inconsistent data semantics).
A large variety may be seen in big data	Simulations can have a variety of types of inputs and responses

3.3. Data farming

"A 'data farming' metaphor captures the notion of purposeful data generation from simulation models. Large-scale experiments let us grow the simulation output efficiently and effectively. We can use modern statistical and visual analytic methods to explore massive input spaces, uncover interesting features of complex simulation response surfaces, and explicitly identify cause-and-effect relationships"(Sanchez, 2018). Data can be grown in simulation experiments to extract many useful insights

- *Data farmers manipulate simulation models to advantage—but using large-scale designed experimentation.*
- *This allows them to learn more about the simulation model's behavior in a structured way.*
- *they "grow" data from their models, but in a manner that facilitates identifying useful information.*
- *The data sets are also better, in the sense they let us identify root cause-and-effect relationships between the simulation model input factors and the simulation output.*

3.4. Visual analytics

Results of simulation experiments can be shown visually by means number of charts. Proper selection of charts can reveal interesting patterns. Visual Analytics is a key technique of a knowledge discovery process for discrete event simulations.

- *In traditional simulation studies, techniques such as animation of process flow, time plots, and graphs of selected outputs are used for visually representations*
- *In Visual Analytics, Data mining algorithms and visualization are used to build up knowledge and draw conclusions from it.*

- *This approach is advantageous because the human mind is able to identify patterns and relations in visual representations quickly.*

Feldkamp *et al.*(2015a, b) lists the following visual tools combined with data mining tools to extract patterns in simulation experiment data(see Table 3). Few case studies on use of visual analytics in simulations in various application areas are found in recent literature. Table 4 lists some of them and the types of visual analytics representations used in such case studies.

Table 3: Knowledge discovery tools used in simulations

Visualization tools	Data Mining methods
• Box plots • Histograms	• Measures of central tendency and variation • Distribution analysis
• Scatter matrix and plots • Parallel coordinate plots • Spider charts	• Multidimensional patterns • Linear regression • Logistic regression
• Flowcharts • Heatmaps • Network graphs	• Correlation tables • Association rules • Bayesian networks • Classification trees

Table 4: Visualization tools and data mining methods for knowledge discovery in some recent literature

No	Reference	Author	Tools used
1	Using Simulation as a Knowledge Discovery Tool in An Adversary C2 Network	Ntuen *et al*, 2009	A hierarchical cluster tree
2	Knowledge Discovery Based Simulation System in Construction	Emad E, 2011	Fuzzy Clustering
3	Knowledge Discovery In Simulation Data: A Case Study Of a Gold Mining Facility	Feldkamp *et al.*, 2016	Correlation matrix of input and output parameters, Matrix scatter plot of selected parameters, Clustering, Linear regression model Radarplots 3D Scatterplot
4	Interactive Visual Analysis of Large Simulation Ensembles	Matkovic *et al*, 2015	Scatterplot Histogram
5	Visual Analytics of Manufacturing Simulation Data.	Feldkamp *et al*, 2015a&b	Correlation matrix of input and output parameters Matrix scatter plot of selected parameters Clustering Linear regression model Radarplots 3D Scatterplot

6	Improving Navy Recruiting with Data Farming	Hogarth *et al*, 2016	Scatter plots Regression models Partition trees
7	A data farming analysis of a simulation of Armstrong's stochastic salvo model	Kesler *et al,* 2019	Pairwise scatter plot Partition tree

4. Robust Design

Robust design is a system optimization and enhancement approach based on the idea that a system shouldn't be judged solely on its average performance. A "good" system must be somewhat insensitive to uncontrollable causes of variation in the system's environment, in addition to demonstrating acceptable mean performance. The purpose of robust design is to help people make better decisions

- it focuses the decision-making process on factors that are controllable in practice;
- it identifies levels and consistency of performance based on those controllable factors;
- robust configurations are more likely to yield better engineering implementations;
- those real-world implementations have in many cases achieved greater reliability and performance at a lower cost.

In the simulation context, robust design can be viewed from different perspectives as depicted in Table 5. This table shows a comparison of experiments on real systems vs computer simulation. Please note that sometimes analogous/physical models/prototypes are easier to experiment with and may draw better results. Because of the expense, effort, and dangers involved in making and observing changes in a real system, one view is that simulation is largely used as a surrogate for a real system; another view is that robust design is an inherent element of the simulation process.

Table 5: Robust design - comparison of experiment with a real system vs simulation

Robust Design: Experiment with a real system	**Robust Design: Experiment with a Simulation model**
Conduct experiments on the real system	Simulation is largely used as a surrogate for a real system
Expense, effort, and dangers involved in making and observing changes in a real system are considerable	Initial development of models involves lots of time and effort to develop a valid and credible model. Lots of calibration efforts are required to fine-tune the model
Changes in the system for experimentation are often difficult and risky	Changes to the model and experimentation are relatively easy.
A large number of inputs and factor levels may not be physically possible always	Large number of input and factor levels can be studied with ease
Running/completion of experiments can take long time and effort. Replication is difficult.	Total time to perform an experiment is significantly less. Replication means we get multiple experimental units (runs or batches) to gain a sense of the magnitude of the variability associated with response and replications are easy in a simulation model.

Randomization is used to guard against hidden or uncontrollable sources of bias.	Results from simulation experiments are perfectly repeatable, and randomization is not needed to guard against hidden or uncontrollable sources of bias.
Homogeneous (*i.e.*, constant) variance is commonly assumed for physical experiments	Heterogeneous (*i.e.*, non-constant) variance is pervasive in stochastic simulation. Consequently, we should not view response variability as merely a nuisance for estimating means or other output statistics, but as an important characteristic of the simulation's behavior.
Experiments on real system help system optimization and improvement process that springs from the view that a system should not be evaluated based on mean performance alone	Robust design can be seen as a process of simulation optimization, where the "best" answer is not overly sensitive to small changes in the system inputs. Kleijnen (2017) calls this "robust optimization." If robust configurations are identified, then the actual results are more likely to conform to the anticipated results after implementation.

Accuracy and Precision: The predicted value of the distribution of outcomes in relation to some desired aim is referred to as accuracy. For example, if our goal is to determine an object's genuine weight, a scale that produces readings with a distribution that has the true weight as its anticipated value is considered an accurate scale, even if individual readings differ by a significant amount. The dispersion of the outcome distribution is referred to as precision. It is considered to be an accurate scale if the measurements are firmly grouped. Figure 3 illustrates many possible combinations of accuracy and precision for somebody shooting at a target. Subplot (a) has shots with low precision because most points are spread from their center of mass. Subplot (b) has high accuracy and precision—the center of mass is on-target, without much the spread. Subplot (c) is precise but not accurate. Many other possibilities are also there.

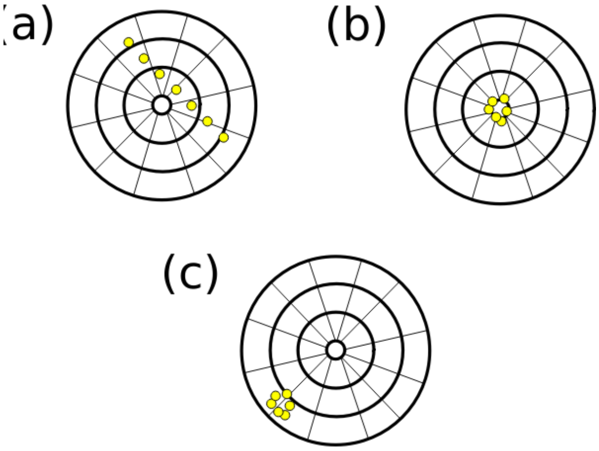
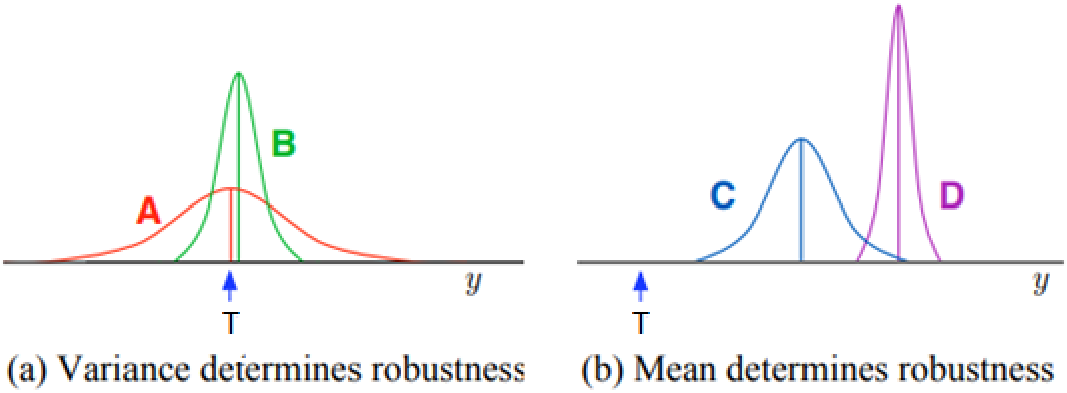


Figure 3: accuracy vs precision

Sanchez and Sanchez (2020) illustrate robustness with a nice example to make it clear that robustness is not solely determined by either the accuracy or the precision of the outcomes relative to the target. Consider response distributions for alternate configurations as depicted in Figure 4. Because the mean response for A is perfectly on target, it is the most accurate. If we were looking for the most accurate system, A would be the best option. System B, on the other hand, is a close second in terms of mean and precision—due to its smaller variance, it is far more likely to produce results close to. Based on this example, we could even argue that the means do not justify the ends in terms of robustness. Both C and D have mean performance above the target value, but when accuracy is taken into account, option D is significantly more likely to be farther from the target. The conclusion here is that robustness is not solely determined by either the accuracy or the precision of the outcomes relative to the target. Tradeoffs may be necessary.



(a) Variance determines robustness (b) Mean determines robustness

Figure 4: comparison of the robustness of systems A, B, C, and D

4.1. Loss function

Robust design analysis makes use of loss functions described by Taguchi. The quality loss function estimate costs associated when the product or process characteristics are shifted from the target value. Such functions help to assess the degree of risk associated with having outcomes that deviate significantly from the specified target. Risk should be non-negative, so loss functions are monotonically non-decreasing as the magnitude of deviations from the target increases. However, a loss function can be asymmetric about the target.

One such loss function commonly used is the quadratic loss function as given below.

$$L(y) = k(y - T)^2$$

where *y* is the observed outcome based on input *x*, *T* is the target value, and *k* is a scaling constant that is often used to adjust the loss to cost. When *k* is set to 1, it is referred to as scaled loss.

We get configurations with low loss when responses are consistently (as measured by variance) close to the target. We accept a tradeoff for small expected deviations from the target with sufficient improvement in consistency of the outcomes, or vice-versa and the variance is non-homogeneous.

4.2. Robust analysis and optimization with simulation metamodels

Computer simulation models of proposed or existing real systems are frequently used to make design decisions. Because it is impractical to build several prototype versions of the real system, or because the cost or other constraints prevent experimenting with the real system, analysts use the simulation model as a surrogate. As these models can be fairly complicated, simpler approximations are frequently created; models of the model, or metamodels. (Kleijnen, 2017). Simulation metamodels can take many forms like multiple regression, partition trees, and forests, or kriging.

Quadratic loss function may be used to fit metamodels of loss directly or fit separate metamodels for the mean and the variability. We often find it convenient to fit the standard deviation as our measure of variability, since it is on the same scale as the mean, but other options such as variance or log(variance) are possible. Fitting separate metamodels help identify which factors, interactions, or non-linear terms are the key causal drivers of average performance and variability. Numerical examples are presented in Sanchez and Sanchez (2020).

5. Conclusion

In this paper we discuss a number of issues related to simulations: how to develop discrete-event simulation models and perform designed experiments to identify significant variables, thus helping simulation optimization searches for optimal solutions. The focus is on some recent developments in the field of simulations which include ideas of visual analytics, data farming, knowledge discovery, and robust design.

Concepts in robust design and analysis in the context of simulation demonstrate how robustness often changes our perspective when contrasted with simulation optimization approaches. Robust solutions can be designed to yield consistently good performance even in the face of uncertainty and uncontrollable factors by incorporating those aspects of the system into the problem formulation.

The process of conceptual modelling is sometimes neglected by analysts and obviously, this can impact the credibility of the model. The author feels that there should be more research connecting the field of statistical modelling with soft operations research and soft systems methodologies. Research on data science tools such as visualisations and data mining applications is sparse in simulation literature and is an open area that requires more research inputs.

References

Emad, E. (2011). *Knowledge Discovery Based Simulation System in Construction, Unpublished Ph.D. Thesis,* Concordia University.

Feldkamp, N., Bergmann, S. and Strassburger, S. (2015a). Knowledge discovery in manufacturing simulations. *In Proceedings of the 2015 ACM SIGSIM PADS Conference,* 3–12.

Feldkamp, N., Bergmann, S. and Strassburger, S. (2015b). Visual analytics of manufacturing simulation data. *In Proceedings of the 2015 Winter Simulation Conference,* edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 779–790. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Feldkamp, N., Bergmann, S. and Strassburger, S. (2016). Knowledge Discovery In Simulation Data: A Case Study Of a Gold Mining Facility. *In Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 779–790. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gittins, P., McElwee, G. and Tipi, N. (2020). Discrete event simulation in livestock management. *Journal of Rural Studies*, Volume 78, August 2020, Pages 387-398.

Hogarth A. R., Lucas T. W., and McLemore C.S. (2016). Improving navy recruiting with data farming, Proceedings of the 2016 Winter Simulation Conference

Kesler, G., Lucas, T. W. and Sanchez, P. J. (2019). A data farming analysis of a simulation of Armstrong's stochastic salvo model, *Proceedings of the 2019 Winter Simulation Conference*

Kleijnen, J. P. C. (2017). Design and analysis of simulation experiments: a tutorial. *In Advances in Modeling and Simulation*, edited by A. Tolk, J. Fowler, G. Shao, and E. Yucesan, 135–158. Cham, Switzerland: Springer International Publishing AG.

Law, A. M. (2015). *Simulation Modeling and Analysis*, 5th edition. McGraw Hill, Boston.

Matkovic, K., Gracanin, D., Jelović, M. and Hauser, H. (2015). Interactive visual analysis of large simulation ensembles, *Proceedings of the 2015 Winter Simulation Conference*.

Ntuen, C. A., Alabi1, O. A., Seong, Y. and Park, E. H. (2009). Using Simulation as a Knowledge Discovery Tool in An Adversary C2 Network, *Proceeding of 14th International Command & Control Research and Technology Symposium C2 and Agility*, Washington, DC, June 15-17

Pidd, M. (2007). Making sure you tackle the right problem: linking hard and soft methods in simulation practice, *Proceedings of the 2007 Winter Simulation Conference*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

Robinson, S. (2011). Conceptual modeling for simulation. *In Encyclopedia of Operations Research and Management Science*, Edited by J.J. Cochran, New York: Wiley.

Robinson, S. (2014). *Simulation: The Practice of Model Development and Use*, Palgrave Macmillan, 2014

Rosenhead, J. and Mingers, J. (2004). Problem structuring methods in action. *European Journal of Operational Research*, **152**, 530 – 554

Sanchez, S M. (2018). Data farming: better data, not just big data, *Proceedings of the 2018 Winter Simulation Conference*, M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

Sanchez, S. M. and Sanchez, P. J. (2020). Robustness revisited: simulation optimization viewed through a different lens, *Proceedings of the 2020 Winter Simulation Conference* K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 69- 82

Method Based on State-Space Epidemiological Model for Cost-Effectiveness Analysis of Non-Medical Interventions- A Study on COVID-19 in California and Florida

Vishal Deo[1] **and Gurprit Grover**[2]
[1]*Department of Statistics, Ramjas College, University of Delhi, Delhi, India*
[2]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi, India*

Received: 02 June 2022; Revised: 12 July 2022; Accepted: 15 July 2022

Abstract

Non-medical containment measures, like quarantine, lockdown, travel restrictions, physical distancing *etc.*, are paramount towards containing the spread of a novel epidemic, especially at its initial stage when little is known about its transmission dynamics and the pathogen responsible for the infections. For these containment measures to be effective, timely identification of infectives through clinical testing is essential. To stress upon the importance of extensive random testing for breaking the chains of transmissions, we have designed a detailed framework for carrying out cost-effectiveness analysis (CEA) of extensive random testing in comparison to targeted testing (the testing policy followed by most countries). This framework can be easily extended to CEA of any other non-medical or even medical interventions for containing epidemics.

We have used the state-space susceptible-infected (quarantined/ free)-recovered-deceased model, which enables predictions of transmission dynamics in the presence of undetected cases, to forecast epidemiological parameters under the two scenarios being compared. The health outcomes have been measured in terms of the estimates of total number of deaths, and infections prevented because of the intervention. Since long-term generic health state measurement is not involved in this study, utility scores are not required for evaluating health benefits.

As a demonstration, the proposed methodology is applied to the COVID-19 data of California and Florida to carry out CEA of ‘extensive random testing’ over ‘targeted testing’ for containing the spread of the epidemic. During the period of the study, these two states were among the worst affected states in the USA, and also had very high percentages of positivity of COVID-19 tests, which raised speculations of inadequate testing capacity.

Key words: State-space epidemic model; Underreporting; MCMC; Cost-effectiveness analysis; Random testing; Non-medical interventions; SI(Q/F)RD model.

1. Introduction

Whenever we encounter an epidemic, the best medical intervention we can think about for containing its spread is a quick resort to mass vaccination of the susceptible population. However, when we face a pandemic like COVID-19, caused by the SARS-CoV-2 virus, the

Corresponding Author: Vishal Deo
Email address: vishaaldeo@gmail.com; vishal_deo@ramjas.du.ac.in

novelty of the virus puts up an arduous challenge before the scientists to develop an effective vaccine in a short span of time. Further, the necessary safety protocols underlining the testing and approval of vaccines, followed by the herculean task of manufacturing it in abundance, makes it practically impossible to get a potent vaccine within a year of the outbreak of the pandemic. Consequently, it is of paramount importance to strategically implement non-medical interventions, like physical distancing, quarantine, lockdown measures *etc.*, to minimize the spread of the infection. The rationale behind these non-medical containment measures is to break the chain of infections by bringing down the basic reproduction number/ rate, R_0, below one. R_0 is an important factor for risk assessment of any epidemic and is defined as the expected number of secondary cases that arise from a typical infectious index-case in a completely susceptible host population. When R_0 is less than one, an infected case is expected to produce less than one new infected. This marks the decline in the number of infecteds over time and, eventually, the epidemic dies out.

Success of any non-medical containment measure relies heavily on the ability to have sufficient testing capacity to identify and isolate the infected people. Even the strongest of the lockdown measures will fail to serve its purpose of breaking transmission chains unless it is accompanied with sufficient amount of random testing. Further, as also argued by the W.H.O, high positivity rate of testing potentially indicates insufficient testing capacity in the region (Deo and Grover (2021)). This leads to a significant amount of underreporting of cases. Significant underreporting of COVID-19 cases in various countries, including the U.S.A., has been reported by various scientific studies (Deo and Grover (2021), Wu *et al*. (2020), Lau *et al*. (2020)). Or, in other words, in the absence of sufficient testing capacity, lockdown measures can only succeed in delaying the spread of the epidemic. W.H.O has issued repeated appeals and advisories to all countries to employ extensive random testing (World Health Organisation (2020 a)). However, only a few countries showed any conviction to conduct adequate number of COVID-19 tests and confined their strategy to testing of symptomatic and high-risk people only. Citing these reasons, we have considered analysing the effectiveness of extensive random testing over targeted testing as a non-medical intervention in containing the spread of COVID-19- both in terms of effectiveness in reducing transmission rates and the associated costs. By the phrase ‘targeted testing’ we imply testing of only symptomatic and high-risk people. To perform the cost-effectiveness analysis (CEA), we have considered the case of two of the worst affected states of USA, California and Florida, which had very high percentages of positive tests. Since the level of testing, and protocols/ procedure of reporting of number of deaths vary between different state jurisdictions, the level of underreporting of deaths and cases can also be expected to vary between states. This is the reason that we have performed state-wise analyses, rather than analysing the combined data of the USA. For forecasting the transmission dynamics of the pandemic under different assumptions regarding prevalence of underreporting, we have used the state-space susceptible-infected (quarantined/ free) -recovered- deceased (SI(Q/F)RD) model given by Deo and Grover (2021). It should be noted that, although underreporting of cases can occur because of various other reasons, we have assumed that lack of sufficient testing is the primary reason for underreporting.

2. Methodology

To realize the objective of conducting CEA of extensive random testing against targeted testing, we propose the following sequence of steps, which are then implemented on the COVID-19 time-series data of California and Florida.

2.1. Predictions using the state-space SI(Q/F) RD model

The Dirichlet-Beta state-space SI(Q/F) RD model, proposed by Deo and Grover (2021), is defined as follows.

2.1.1. Defining transitions between different compartments of the model

The states and transitions of the compartmental set-up of the SI(Q/F)RD model can be visualised in Figure 1. Further, these transitions are quantified through the following set of differential equations.

$$\frac{d\theta_t^S}{dt} = -\left[\beta_1\theta_t^Q + \beta_2\theta_t^F\right]\theta_t^S \tag{1}$$

$$\frac{d\theta_t^I}{dt} = \left[\beta_1\theta_t^Q + \beta_2\theta_t^F\right]\theta_t^S - \gamma_1\theta_t^Q - \gamma_2\theta_t^F - d_1\theta_t^Q - d_2\theta_t^F \tag{2}$$

$$\frac{d\theta_t^R}{dt} = \gamma_1\theta_t^Q + \gamma_2\theta_t^F = \gamma\theta_t^I\ (if\ \gamma_1 = \gamma_2 = \gamma) \tag{3}$$

$$\frac{d\theta_t^D}{dt} = d_1\theta_t^Q + d_2\theta_t^F \tag{4}$$

$$where, \theta_t^Q = p_t\theta_t^I\ and\ \theta_t^F = (1 - p_t)\theta_t^I\ , and\ \theta_t^S + \theta_t^I + \theta_t^R + \theta_t^D = 1 \tag{5}$$

Here, $\theta_t^S, \theta_t^I, \theta_t^Q, \theta_t^F, \theta_t^R$ and θ_t^D are the true but unobserved (latent) prevalence of susceptibles, infecteds, infected and quarantined, infected and free (undetected), recovered, and deceased respectively. In other words, they are the probabilities of a person being in the respective compartments at time *t*.

2.1.2. Dirichlet-Beta state-space formulation of the SI(Q/F) RD model

Let $\boldsymbol{\theta}_t = (\theta_t^S, \theta_t^I, \theta_t^R, \theta_t^D)^T$ be the latent population prevalence, and $\boldsymbol{f}(\boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{d})$ be the solution of the set of differential equations for time *t*, where the function takes the values of the vectors $\boldsymbol{\theta}_{t-1}$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, $\boldsymbol{d} = (d_1, d_2)^T$and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$ as the arguments. Then the Bayesian hierarchical Dirichlet-Beta state-space SI(Q/F)RD is defined as follows [Deo and Grover (2021)].

$$Y_t^I|\boldsymbol{\theta_t}, \tau \sim Beta\left(\lambda^I\theta_t^I, \lambda^I(1 - \theta_t^I)\right) \tag{6}$$

$$Y_t^R|\boldsymbol{\theta_t}, \tau \sim Beta\left(\lambda^R\theta_t^R, \lambda^R(1 - \theta_t^R)\right) \tag{7}$$

$$Y_t^D|\boldsymbol{\theta_t}, \tau \sim Beta\left(\lambda^D\theta_t^D, \lambda^D(1 - \theta_t^D)\right) \tag{8}$$

$$\text{and, } \boldsymbol{\theta_t}|\boldsymbol{\theta_{t-1}}, \tau \sim Dirichlet(\kappa f(\boldsymbol{\theta_{t-1}}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{d})) \tag{9}$$

where, $\tau = \{\boldsymbol{\theta_0}, \kappa, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{d}, \lambda^I, \lambda^R, \lambda^D\}$, $\boldsymbol{\theta_0}$ is the baseline value of the vector $\boldsymbol{\theta_t}$, and $\lambda^I, \lambda^R, \lambda^D, \kappa > 0$ control the variances of the distributions defined in equations (6), (7), (8) and (9) respectively. Prior distributions of the model parameters are defined as follows.

$$\theta_0^I \sim Beta(1, (Y_1^I)^{-1}), \theta_0^R \sim Beta(1, (Y_1^R)^{-1}), \theta_0^D \sim Beta(1, (Y_1^D)^{-1}), \theta_0^S = 1 - \theta_0^I - \theta_0^R - \theta_0^D \tag{10}$$

$$R_i \sim LogN\left(\mu_{r_i}, \sigma_{r_i}^2\right), \sigma_{r_i}^2 = ln\left(\frac{V(R_i) + (E(R_i))^2}{(E(R_i))^2}\right) \text{ and } \mu_{r_i} = ln\left(E(R_i)\right) - \frac{\sigma_{r_i}^2}{2}, i = 1{,}2 \tag{11}$$

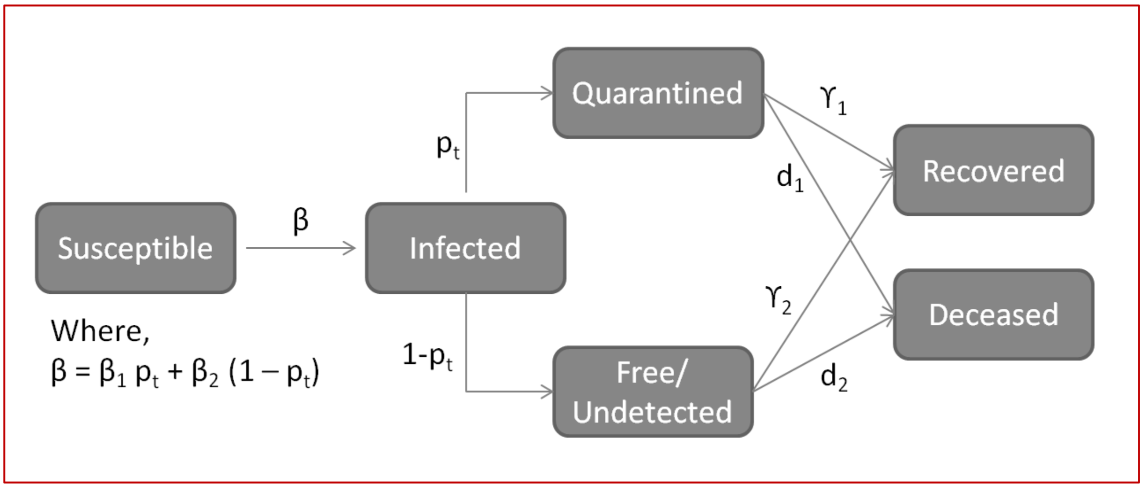
$$\gamma_i \sim LogN\left(\mu_{g_i}, \sigma_{g_i}^2\right), \sigma_{g_i}^2 = ln\left(\frac{V(\gamma_i) + (E(\gamma_i))^2}{(E(\gamma_i))^2}\right) \text{ and } \mu_{g_i} = ln\left(E(\gamma_i)\right) - \frac{\sigma_{g_i}^2}{2}, i = 1{,}2 \tag{12}$$

$$p_t \sim Beta\left(a_p, b_p\right), \forall t = 1{,}2 \ldots\ldots T \tag{13}$$

R_1 and R_2 are basic (average) reproduction rates associated with quarantined (Q) and undetected (F) infecteds, respectively. That is, $R_i = \frac{\beta_i}{(\gamma_i+d_i)}$, $i = 1,2$.

$$\kappa \sim Gamma(a_k, b_k), \lambda^I \sim Gamma(a_I, b_I), \lambda^R \sim Gamma(a_R, b_R), \lambda^D \sim Gamma(a_D, b_D) \qquad (14)$$

Elaborate procedure for the estimation of the parameters and hyper-parameters, and for forecasting using this model is outlined in Deo and Grover (2021). These procedures are used to predict the number of infections and deaths under the base intervention- targeted testing, using current estimates of underreporting based on the observed data.



Source: Deo and Grover (2021)

Figure 1: SI(Q/F)RD model structure- p_t is the proportion of infecteds detected and quarantined, $1\text{-}p_t$ is the proportion of infecteds who are undetected and roaming freely among the susceptible, β_1 is the transmission rate associated with quarantined infected and β_2 is the transmission rate associated with undetected infected, Υ_1 and d_1 are rate of recovery and rate of death for quarantined cases and Υ_2 and d_2 are rate of recovery and rate of death for undetected cases.

2.2. Prediction under the assumption of extensive random testing and CEA

Extensive random testing can be expected to result in a significant rise in expenditure on the testing kits and medical personnel. However, it can play a major role in breaking the chains of transmission and hence, result in a significant reduction in the overall number of infecteds and deaths due to the COVID-19 epidemic. The outcome of CEA will tell us how much additional overall cost is required to save one additional person from getting infected, or to save one additional person from dying due to the infection. That is, CEA will be conducted in terms of the outcomes, ‘infection’ and ‘death’. It should also be noted that, if the total duration of the epidemic is reduced drastically because of the recommended intervention ‘extensive random testing’, the overall expected cost may even come out to be lesser than the expected cost of using targeted testing strategy.

To derive the outcomes pertaining to the recommended intervention, *i.e.*, ‘extensive random testing’, following procedure is followed.

a. In terms of the SI(Q/F)RD model, the major difference between the outcomes of the two scenarios will rely on the difference in the proportion of infecteds being detected and quarantined, *i.e.*, p_t.

b. It will be impractical to assume that 100% infecteds can be detected using extensive random testing. This is because even popular tests like the reverse transcription polymerase chain reaction (RT-PCR), which is also recommended by the W.H.O. [World Health Organisation (2020 b)], do not have 100% sensitivity and specificity. Sensitivity and specificity may vary according to the laboratory settings and expertise levels of the medical practitioners. Different studies have reported varying levels of sensitivity and specificity of the RT-PCR test, mostly ranging from around 80% to 95% [(West *et al.* (2020), Padhye (2020), Tahamtan and Ardebili (2020)]. On a conservative note, we have assumed that the average proportion of detection of infecteds will be 80%, *i.e.*, p_t will have a mean of 0.8. Instead of assigning a fixed value to p_t, we have assumed p_t to follow Beta distribution to introduce realistic variability in the calculations. The mean of the distribution is taken as 0.8 and its variance is obtained from the results of the state-space model estimated by the method described in section 2.6.

c. To simulate a practically realistic situation, we have assumed that the extensive random testing can be applied only after first 30 days of the outbreak of the epidemic. This is because, extensive random testing requires procurement of testing kits and other logistic arrangements on a large scale, which need some time to be organized. To accommodate this assumption into calculations, the mean value of the distribution of p_t for the first 30 days can be based on the average of posterior estimates of p_t for the first 30 days obtained from the state-space SI(Q/F)RD model. That is, we are assuming that there will not be much difference in the outcomes and costs associated with the two interventions in the initial days of the epidemic.

d. Simulation exercise:

   i. At each t, t = 1, 2,..., T, L number of values are generated on the parameters R_1, R_2, γ_1 and γ_2 from their distributions defined in the equations (11) and (12). The parameters of these distributions are calculated on the basis of their posterior estimates obtained from the state-space SI(Q/F)RD model. The corresponding values on p_t are simulated from its distribution defined in the previous step c. Fixed values of the death rates, d_1 and d_2, are assumed to be same, as also for the state-space SI(Q/F)RD model.

   ii. For each combination *(t, l), t = 1,2,..., T* and *l = 1,2,...,L* , the respective simulated values of the parameters are used in the fourth degree Runge-Kutta approximation of the solution of the set of differential equations of the SI(Q/F)RD model to obtain $f(\boldsymbol{\theta}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{d})$, *i.e.*, the values of the latent process at time *t* as a function of their values at time *t-1*. At the start of the iteration, the initial values of these latent process variables are assigned as the vector $\boldsymbol{\theta}_0$**.** The mean of the L values of the latent process at a time t is taken as its estimate, *i.e.*, $\widehat{\boldsymbol{\theta}}_t = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{\theta}_{\boldsymbol{t}}^{(\boldsymbol{l})}$. Sample quantiles (0.025, 0.975) are used to obtain 95% credible intervals at each *t*.

   iii. At each *t*, *t* = 1,2,...,*T*, *L* values of λ^I, λ^R and λ^D are simulated from their respective Gamma distributions whose parameters are calculated from the posterior estimates of their means and variances obtained from the state-space SI(Q/F)RD model. At each combination (*t*, *l*), *t* = 1,2,..., *T* and *l* = 1,2,...,*L* , using the estimate of the latent prevalence process, $\widehat{\boldsymbol{\theta}}_t$**,** from the previous step and the generated values of λ^I, λ^R and λ^D, $(Y_t^{I(l)}, Y_t^{R(l)}, Y_t^{D(l)})$ are simulated from their respective Beta distributions. Finally, mean of these *L* values at a time *t* is taken as the estimate of the observed process at *t*. These proportions can be multiplied with the total number of susceptibles (total population of the state) and rounded to obtain the estimated counts of each compartment at time *t*, $t = 1,2,\ldots T$.

e. Total number of infected cases and total number of deaths, till the end of the epidemic, are calculated from the predictions for each case (interventions). These values give us the difference in outcomes (infection/ death) under two interventions. Let, (C_1, D_1) be the estimates of total number of infecteds and total number of deaths during the entire course of the epidemic for the base intervention, targeted testing, and (C_2, D_2) be the respective estimates for the recommended intervention, extensive random testing.

To obtain the estimate of total costs associated with the two interventions we will first need to estimate the total number of tests that will be conducted under the two testing strategies (interventions). For the base intervention of targeted testing, the current percentage of positivity of tests in the state can be used to obtain an estimate of the total number of tests to be conducted by the end of the epidemic. If r_1 is the current proportion of positive tests in the state, the estimate of total number of tests which will be conducted under the base intervention will be given as, $N_1 = \frac{Q_1}{r_1}$, where Q_1 is the number of infecteds who are detected and quarantined. For the second intervention of extensive random testing, the proportion of positive tests is taken as the probability that a person in the state got infected during the entire duration of the epidemic and is simply given as, $r_2 = \frac{C_2}{Total\ Population}$. Subsequently, the total number of tests under the second intervention is estimated as, $N_2 = \frac{Q_2}{r_2}$, where Q_2 is the number of infecteds who are detected and quarantined under the intervention extensive random testing. As an alternative, N_2 has also been taken as the total population, assuming that all individuals are tested (once) by the time the epidemic gets over in the state.

Let *Z* be the per unit average cost of COVID-19 test, then the incremental cost-effectiveness ratio (ICER) is calculated as the ratio of change in cost to the change in outcome as follows,

$$\text{ICER}_{\text{inf}} = \frac{(N_2 - N_1)Z}{(C_1 - C_2)} \quad \text{and} \quad \text{ICER}_{\text{death}} = \frac{(N_2 - N_1)Z}{(D_1 - D_2)} \tag{15}$$

3. Implementation and Results

3.1. Data

In this paper, we have used the same data for conducting the CEA which was used for demonstrating the estimation and prediction methodology of state-space SI(Q/F)RD model in Deo and Grover (2021). Description of the data is provided in Table 1.

3.2. Estimates and predictions for the base intervention- targeted testing

Posterior estimates of the parameters of the Dirichlet-Beta state-space SI(Q/F)RD model and the predicted values of number of infecteds and deaths based on these estimates are taken from the results of Deo and Grover (2021). These results are presented in the Appendix A in the Table A.1, Table A.2, Graph A.1, and Graph A.2.

Table 1: Data description

Sl. No.	Data	Source
1	Daily time-series data on total confirmed cases and total deaths for the states of California and Florida (Till 11 July 2020)	Github repository of the Centre for Systems Science and Engineering (CSSE), Johns Hopkins University, Maryland, USA [https://github.com/CSSEGISandData/COVID-19]
2	Weekly state-wise estimates of excess deaths associated with COVID-19 till 11 July 2020.	Website of CDC [https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.html]
3	Data on rates of positivity of COVID-19 testing for the two states, California and Florida [As on 29 July 2020]	Website of Johns Hopkins University [https://coronavirus.jhu.edu/ testing/testing-positivity].

Source: Deo and Grover (2021)

3.3. Predictions under the assumption of extensive random testing (recommended intervention)

Once the posterior estimates of the transmission parameters are obtained from the state-space SI(Q/F)RD model, predictions of observed process under the assumption of extensive random testing are carried out using the steps outlined in section 2.2. Based on the posterior mean and standard deviation of the parameters of the state-space model, following specifications are used for conducting the required simulations to predict the transmission dynamics of the epidemic.

California:

$$R_1 \sim LogN(-0.822, 0.495), E(R_1) = 0.497, V(R_1) = 0.069;\ \beta_1 = R_1(\gamma + d_1)$$

$$R_2 \sim LogN(0.376, 0.106), E(R_2) = 1.464, V(R_2) = 0.024;\ \beta_2 = R_2(\gamma + d_2)$$

$$\gamma \sim LogN(-2.68, 0.087), E(\gamma) = 0.069, V(\gamma) = 0.00004$$

$$p_t \sim Beta(4.49, 59.61),\ E(p_t) = 0.07, V(p_t) = 0.001\ \forall t \leq 30$$

$$p_t \sim Beta(15.2, 3.8), E(p_t) = 0.8, V(p_t) = 0.008\ \forall t > 30$$

$$\lambda^I \sim Gamma(1012524.75, 1.88e-06), \lambda^R \sim Gamma(1633152.503, 1.46e-05), \\ \lambda^D \sim Gamma(1355.195, 0.00262)$$

Florida:

$$R_1 \sim LogN(-1.19, 0.573), E(R_1) = 0.359, V(R_1) = 0.05;\ \beta_1 = R_1(\gamma + d_1)$$

$$R_2 \sim LogN(0.476, 0.06), E(R_2) = 1.612, V(R_2) = 0.009;\ \beta_2 = R_2(\gamma + d_2)$$

$$\gamma \sim LogN(-2.77, 0.063), E(\gamma) = 0.063, V(\gamma) = 0.00002$$

$$p_t \sim Beta(1.39, 8.55),\ E(p_t) = 0.14, V(p_t) = 0.011\ \forall t \leq 30$$

$$p_t \sim Beta(41.87,10.47), E(p_t) = 0.8, V(p_t) = 0.003\ \forall t > 30$$

$$\lambda^I \sim Gamma(999169.436,1.76e-06), \lambda^R \sim Gamma(1807366.511,1.35e-05),$$
$$\lambda^D \sim Gamma(1022.341, 0.011)$$

The entire simulation exercise for this section is implemented in R programming through self-written codes. Plots of predicted values of daily number of active infected cases and cumulative deaths, along with their 95% confidence intervals, are shown in Figures 2 and 3, respectively. For a comparative assessment of the predictions of transmission trajectory of the epidemic under the two interventions, daily counts of active infecteds and cumulative number of deaths for both cases are plotted together in Figures 4 and 5.

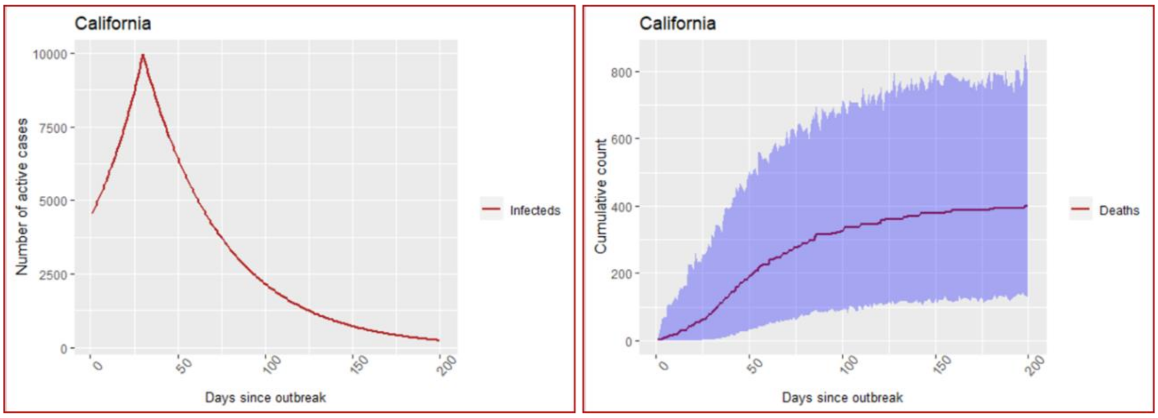


Figure 2: California - Predictions under the assumption of extensive random testing. The blue shaded region depicts the region of 95% confidence intervals based on simulated values. The confidence region for number of infecteds is too narrow to be visible in the graph.

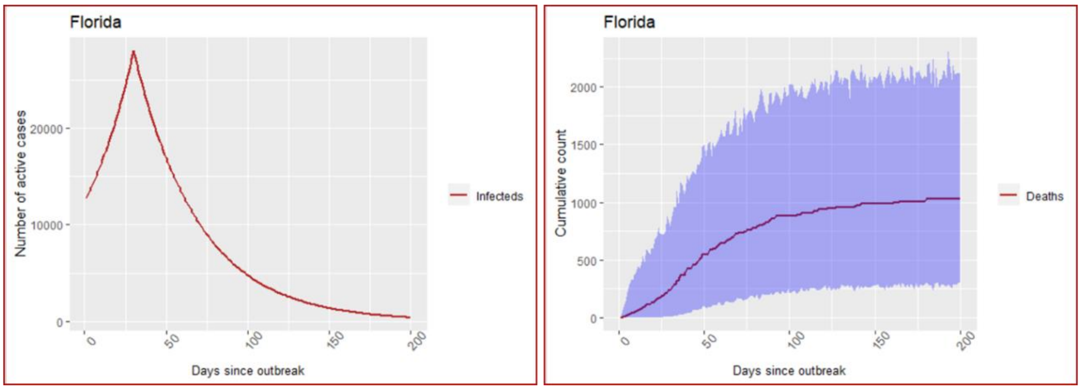


Figure 3: Florida - Predictions under the assumption of extensive random testing. The blue shaded region depicts the region of 95% confidence intervals based on simulated values. The confidence region for number of infecteds is too narrow to be visible in the graph.

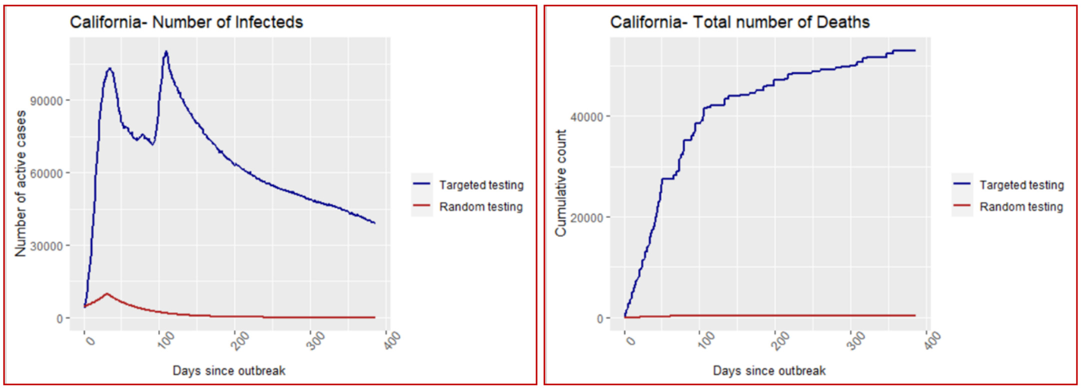


Figure 4: California - Comparative graphs of predictions of cases under both interventions

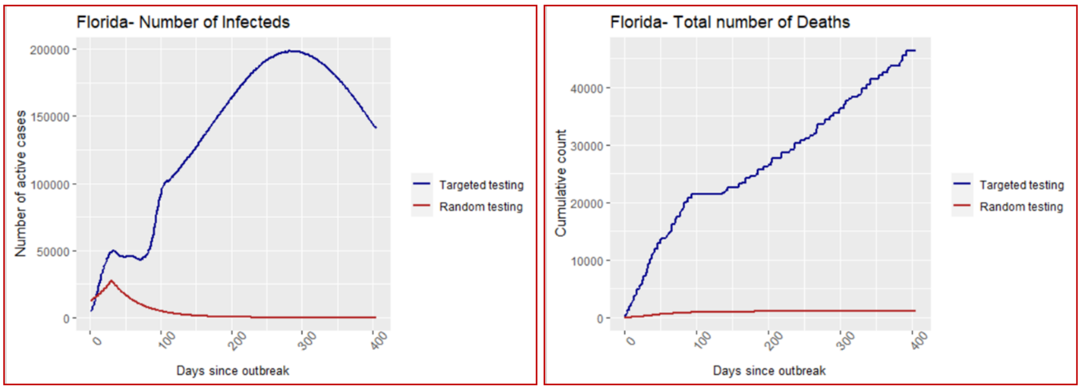


Figure 5: Florida - Comparative graphs of predictions of cases under both interventions

3.4. CEA of extensive random testing over targeted testing

To estimate the cost incremental, we first need the estimates of total number of tests to be conducted under both interventions. The rates of positivity of COVID-19 testing, as reported till 29 July 2020, were 7.47% in California and 18.96% in Florida. These percentages were taken as r_1 for estimating number of tests under the base intervention of targeted testing. The rates of positivity of tests under the assumption of extensive random testing, r_2 are obtained as $r_2 = \frac{C_2}{Total\ Population}$ for the two states and are provided in Table 2.

Table 2: Estimates of rates of positivity of tests under extensive random testing

Intervention= Extensive random testing	California	Florida
***C*₂**, total no. of infecteds by the end of the epidemic	45,819	108,290
Total no. of susceptibles at the start of the epidemic (Taken as total population of the state)	39,512,223	21,477,737
***r*₂** (considered as the probability that a person in the state got infected during the entire duration of the epidemic)	= 0.0012 (0.12%)	= 0.005 (0.5%)

Values of cost incremental, changes in outcome measures (both in terms of number of infections and number of deaths), and ICERs are calculated by implementing these values of

r_1 and r_2 in the steps discussed in part e of section 2.2. Cost of COVID-19 test (RT-PCR) varies considerably across USA. However, leaving out some extreme cases, the average cost per unit of the RT-PCR test is around $100 in USA [Kliff (2020)]. We have used this average cost for evaluating cost incremental owing to increment in the number of tests. Results on the difference in number of tests, cost increment (or decrement), and changes in the outcomes of number of infections and number of deaths, on using the proposed intervention 'extensive random testing' over the base intervention 'targeted testing', are furnished in Table 3. Further, incremental cost effectiveness ratios associated with extensive random testing as compared to targeted testing are presented in Table 4.

Table 3: Changes in outcomes and costs on using extensive random testing instead of targeted testing as the intervention to contain the spread of SARS-CoV-2 infections

State	Total Confirmed Cases	Total Detected Cases	Number of Tests	Number of Deaths
Predictions under the base case (Base intervention = Targeted testing)				
California	2384143	1328158	17779893	58292
Florida	4793903	1873747	9882632	58937
Predictions under the ideal case (Recommended intervention = Extensive random testing) (Case A- Number of tests estimated using r_2)				
California	45819	41237	35560914	405
Florida	108290	97461	19329963	1039
Predictions under the ideal case (Recommended intervention = Extensive random testing) (Case B-Assuming that everyone was tested by the end of the epidemic)				
California	45819	41237	39512223	405
Florida	108290	97461	21477737	1039
Changes in Cost and Outcomes				
State	Reduction in occurrence of infection	Reduction in occurrence of death	Tests incremental	Testing cost incremental ($)
Case A-Cost-effectiveness of extensive random testing with respect to targeted testing				
California	2338324	57887	17781021	1778102100
Florida	4685613	57898	9447331	944733100
Case B-Cost-effectiveness of extensive random testing with respect to targeted testing				
California	2338324	57887	21732330	2173233000
Florida	4685613	57898	11595105	1159510500

Table 4: Incremental cost effectiveness ratios associated with extensive random testing as compared to targeted testing

Incremental Cost Effectiveness Ratio- ICER		
State	Number of extra tests per unit reduction in infection	
	Case A*	Case B**
California	8	9
Florida	2	2
	Number of extra tests per unit reduction in death	
	Case A	Case B
California	307	375
Florida	163	200

ICER (infections)	Additional testing cost per unit reduction in infection ($)	
	Case A	Case B
California	760	929
Florida	202	247
ICER (deaths)	Additional testing cost per unit reduction in death ($)	
	Case A	Case B
California	30717	37543
Florida	16317	20027

*Case A- Number of tests estimated using r_2.

**Case B- Assuming that everyone was tested by the end of the epidemic.

4. Discussion

As per the posterior estimates obtained from the state-space SI(Q/F)RD model, around 43% infected cases in California and 61% infected cases in Florida, on an average, go unreported [Deo and Grover (2021)]. Further, the significantly higher posterior estimates of average reproduction number associated with undetected infecteds [California: 1.464 (sd: 0.155) and Florida: 1.612 (sd: 0.097)] as compared those for quarantined infecteds [California: 0.497 (sd: 0.262) and Florida: 0.359 (sd: 0.224)] stresses upon the necessity for conducting the CEA proposed in this study.

For both states, CEA of extensive random testing over targeted testing has yielded very strong results in favour of the former; refer Table 3 and Table 4. Citing uncertainties because of some unknown factors and leaving some space for errors in testing, even if we assume that 80% of the infecteds can be detected and quarantined using extensive random testing, a total of around 2.3 million people in California and 4.7 million people in Florida could be saved from the infection by the end of the epidemic if extensive random testing was used instead of targeted testing. Further, it is estimated that around 58 thousand deaths due to COVID-19 could be averted in each state if the states resorted to extensive random testing (after first month of outbreak) instead of targeted testing. These are huge expected gains for humanity, especially when every single life matter for us. The ICER values (in terms of number of tests) suggest that, on an average, only around 9 and 2 additional number of tests would be required in total to save one extra person from getting infected in California and Florida, respectively, by the time the epidemic ends. That is, around 760- 929 USD (California) and 202- 247 USD (Florida) additional expenditure on COVID-19 tests would be required to save every additional person from getting infected. Number of additional tests required to save one additional death from COVID-19 is estimated to be around 307- 375 for California and 163- 200 for Florida. That is, on using extensive random testing over targeted testing, one extra loss of life due to COVID-19 can be averted on an additional expenditure of around 30717- 37543 USD in California and around 16317- 20027 USD in Florida.

5. Conclusion

We have provided a comprehensive framework for conducting CEA of non-medical interventions for containing epidemics like COVID-19. To the best of our knowledge, there is no standard procedure available in the literature for conducting such analysis.

Results of the CEA conclude that extensive random testing, which has been strongly recommended by WHO, is significantly cost-effective over targeted testing. Since the R_0 values associated with quarantined infecteds in both states are estimated to be below 1, extensive random testing, resulting in quarantining of at least 80% infecteds, is expected to result in the epidemic to end quite quickly as compared to the case of targeted testing. So, targeted testing may imply a smaller number of tests over a much longer period of time, while extensive testing means a very high number of tests for a much shorter period of time. This simple logic is corroborated by the ICER values obtained from the CEA of extensive random testing over targeted testing. For California, if the state is willing to conduct around 9 extra tests (or spend around 900 USD extra amount on testing) for saving one additional person from getting infected, or if the state is willing to conduct around 375 extra tests (or spend around 37500 USD extra amount on testing) for saving one additional person from dying due to COVID-19, extensive random testing can be considered as cost-effective over targeted testing. While for Florida, willingness to spend an extra amount of around 200 USD (2 extra tests) for saving one additional person from getting infected, or willingness to spend an extra amount of around 20,000 USD (200 extra tests) for saving one additional person from dying due to COVID-19, renders extensive random testing as cost-effective over targeted testing.

Acknowledgements

The authors are grateful to the editor and anonymous reviewers for their invaluable suggestions towards improving the paper.

References

Deo, V. and Grover, G. (2021). A new extension of state-space SIR model to account for Underreporting – An application to the COVID-19 transmission in California and Florida. *Results in Physics*, **24**, 104182. https://doi.org/10.1016/j.rinp.2021.104182

Kliff, S. (2020, June 16). Most Coronavirus Tests Cost About $100. Why Did One Cost $2,315? *The New York Times*. Retrieved from https://www.nytimes.com/2020/06/16/upshot/coronavirus-test-cost-varies-widely.html

Lau, H., Khosrawipour, T., Kocbach, P., Ichii, H., Bania, J., and Khosrawipour, V. (2020). Evaluating the Massive Underreporting and Undertesting of COVID-19 Cases in Multiple Global Epicenters. *Pulmonology*, **27(2)**, 110-115. https://doi.org/10.1016/j.pulmoe.2020.05.015.

Padhye, N. S. (2020). Reconstructed diagnostic sensitivity and specificity of the RT-PCR test for COVID-19. *MedRxiv (Preprint)*. doi:https://doi.org/10.1101/2020.04.24.20078949

Tahamtan, A., and Ardebili, A. (2020). Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Review of Molecular Diagnostics*, **20(5)**, 453-454. doi:https://doi.org/10.1080/14737159.2020.1757437

West, C. P., Montori, V. M., and Sampathkumar, P. (2020). COVID-19 Testing: The Threat of False-Negative Results. *Mayo Clinic Proceedings*, **95(6)**, 1127-1129. doi:https://doi.org/10.1016/j.mayocp.2020.04.004

World Health Organisation. (2020 a). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 16 March 2020.* Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020

World Health Organisation. (2020 b). *Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19).* Retrieved from https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., *et al.* (2020). Substantial Underestimation of SARS-CoV-2 Infection in the United States. *Nature Communications*, **11**, 4507. https://doi.org/10.1038/s41467-020-18272-4

Appendix A

Table A.1: Posterior estimates of time-invariant parameters of the state-space SI(Q/F) RD model, along with their standard deviations and 95% credible intervals-California

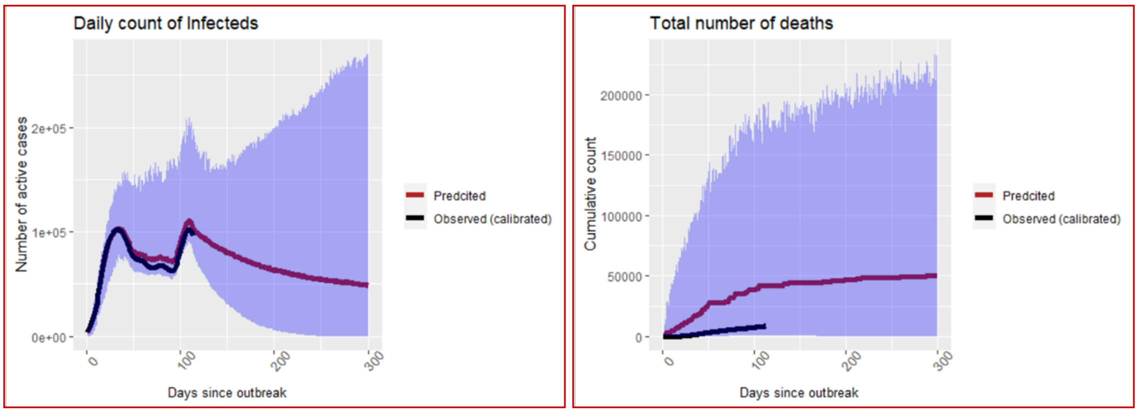
Parameter	Posterior mean	Posterior standard deviation	95% credible interval
R_1	0.497	0.262	[0.068, 1.004]
R_2	1.464	0.155	[1.214, 1.813]
γ	0.069	0.006	[0.056, 0.081]
κ	336063.593	47259.956	[243264.879, 431918.329]
λ^D	1355.195	718.277	[397.588, 2632.883]
λ^I	1012524.750	734717.729	[1349.955, 2006982.462]
λ^R	1633152.503	334437.988	[1073803.103, 2360304.964]
$\widehat{\beta}_1 = \widehat{R}_1(\widehat{\gamma} + d_1)$		0.035	
$\widehat{\beta}_2 = \widehat{R}_2(\widehat{\gamma} + d_2)$		0.102	

Source: Deo and Grover (2021)

Table A.2: Posterior estimates of time-invariant parameters of the state-space SI(Q/F) RD model, along with their standard deviations and 95% credible intervals-Florida

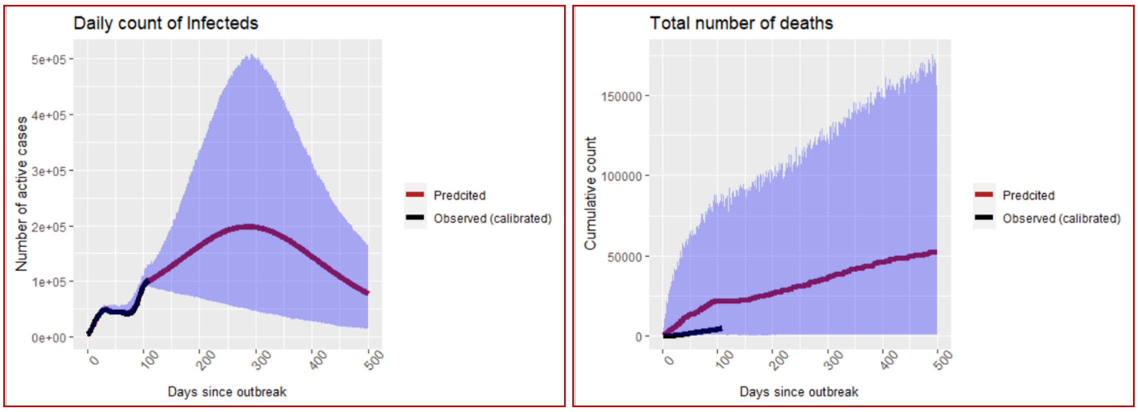
Parameter	Posterior mean	Posterior standard deviation	95% credible interval
R_1	0.359	0.224	[0.052, 0.880]
R_2	1.612	0.097	[1.416, 1.799]
γ	0.063	0.004	[0.054, 0.071]
κ	500800.490	94547.445	[327261.995, 679843.447]
λ^D	1022.341	303.916	[539.044, 1629.331]
λ^I	999169.436	753473.727	[4778.595, 2403835.884]
λ^R	1807366.511	365988.299	[1164580.155, 2616920.665]
$\widehat{\beta}_1 = \widehat{R}_1(\widehat{\gamma} + d_1)$		0.0229	
$\widehat{\beta}_2 = \widehat{R}_2(\widehat{\gamma} + d_2)$		0.102	

Source: Deo and Grover (2021)



Source: Deo and Grover (2021)

Figure A.1: Predictions of number of infected and number of deaths in California under the base case/ intervention of targeted testing. The blue shaded ribbon is the region of 95% credible intervals.



Source: Deo and Grover (2021)

Figure A.2: Predictions of number of infected and number of deaths in Florida under the base case / intervention of targeted testing. The blue shaded ribbon is the region of 95% credible intervals.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 83-90

One-inflated Intervened Poisson Distribution: Stochastic Representations and Estimation

V.S.Vaidyanathan and Jahnavi Merupula
Department of Statsitics
Pondicherry University, Puducherry, India.

Received: 01 July 2022; Revised: 16 July 2022; Accepted:20 July 2022

Abstract

Count data modelling using Poisson distribution has applications in medicine, biology, physical sciences etc,. For example, the number of people affected by a strain of virus, number of gamma ray emissions etc., can be modelled using Poisson distribution. Sometimes, it is necessary to alter the rate of occurrence of counts through intervention, like administering vaccines that alters the rate of people getting affected by a virus. Such an alteration of Poisson counts results in what is generally called in the literature as an intervened Poisson distribution whose support is the set of positive integers. There are situations in which these counts can occur with more frequency than what is expected from the underlying distribution. For example, the number of visits to a physician has more frequency of 1's. This can happen either due to people visiting for a general health checkup or treatment of any ailment. Inflated count data models are often used to model count data with excess counts. Popular inflated count data models include inflated Poisson and negative binomial distributions. In this paper, an intervened Poisson distribution with one inflated count is developed. Also, two stochastic representations of the model are discussed. The moment generating function of the model is derived, and parametric estimation using the frequentist approach is carried out. A real-life application of the model is also discussed.

Key words: EM algorithm; Intervened Poisson; Maximum likelihood estimation; Moment generating function; One inflation; Zero-truncated Poisson.

AMS Subject Classifications: 60E05, 62F10.

1. Introduction

Poisson distribution is one of the oldest distributions for modelling count data. Over the years, this distribution has evolved into various forms like truncated, intervened, inflated and generalized Poisson distribution. Applications of Poisson distribution can be seen in medical, epidemiological, environmental, physical sciences etc. For a detailed discussion on various Poisson models and their applications, one may refer to Johnson et al. (2005). Inflated count models are used when a particular count frequency is more prominent than expected from the model. The excess count frequencies are attributed to having come from other generating processes. Inflated Poisson models can be used to model count data with excess counts by considering them to be generated from a degenerate distribution. Lambert (1992) introduced

Corresponding Author: V.S.Vaidyanathan
Email: vaidya.stats@gmail.com

the zero-inflated Poisson distribution for modelling the number of defects of manufacturing equipment. Following Lambert (1992), many researchers have developed various inflated Poisson distributions. Godwin and Bohning (2017) have developed a one-inflated positive Poisson model to estimate the population size of an animal species. Melkersson and Olsson (1999) have proposed a zero-one-inflated Poisson model to analyze the number of visits to a dentist.

In certain situations, the count data generating process is altered due to an intervention mechanism. It is to be noted that the intervention mechanism is activated when at least one event has occurred. For example, the intervention mechanism can be administering a drug to control the spread of disease, adjusting the specifications of a manufacturing process to reduce the number of defects etc. To accommodate the effect of the intervention on the mean of the Poisson distribution, Shanmugam (1985) introduced an intervened Poisson distribution whose probability mass function (pmf) is as given in equation (1) using a zero-truncated Poisson distribution. The intervention parameter ρ alters the mean of the Poisson distribution after the intervention mechanism.

When the intervention mechanism decreases the mean of the underlying Poisson distribution, one can expect the frequency of the smaller counts to be high. As a consequence, there might be a surge in the one counts. Also, assuming some of these 1's to be arising from a degenerate distribution outside the intervened Poisson model, the overall counts can be modelled by a one-inflated intervened Poisson distribution (OIIPD). For example, in the context of controlling the spread of the SARS-CoV-2 virus through vaccination, it is observed that people can still be infected by the virus even after vaccination. Thus, if we consider the number of individuals infected exactly once, they belong either to the vaccinated group or the unvaccinated group. Thus, the number of 1's is from two generating processes.

The rest of the paper is organized as follows. In Section 2, the pmf of the OIIPD is derived and its distributional properties are presented. Two stochastic representations (SR) of the proposed distribution are constructed in Section 3, and their equivalence is shown. In Section 4, the stochastic representations are used to derive the moment generating function and the moments of OIIPD. The estimation of parameters of the OIIPD is discussed in Section 5 through maximum likelihood (ML) estimation and EM algorithm. A Numerical illustration of the estimation procedure is presented in Section 6 using real-life data. The conclusion of the paper is given in Section 7.

2. Model Formulation and Properties

The pmf of zero-truncated Poisson distribution for positive integer-valued random variable T with mean λ is given by

$$P(T=t)=\frac{\lambda^t}{t!(e^\lambda-1)}; \quad t=1,2,\ldots,\lambda>0.$$

After some intervention mechanism, let us suppose that the mean changes from λ to $\rho\lambda$. The parameter ρ, $0\leq\rho<\infty$ is called the intervention parameter. Let V denote a Poisson random variate with mean $\rho\lambda$. Define $X=T+V$. The pmf of X is then obtained using convolution and is given by (Shanmugam (1985))

$$\begin{aligned} P(X=x) &= \sum_{l=0}^{x-1} P(T=x-l)P(V=l|T=x-l) \\ &= \frac{e^{-\rho\lambda}}{(e^{\lambda}-1)}\left[\frac{(\rho+1)^x-\rho^x}{x!}\right]\lambda^x; \quad x=1,2,\ldots \end{aligned} \tag{1}$$

X defined above is said to have intervened Poisson distribution (IPD). The first two moments of X are respectively given by

$$E(X)=\lambda\left[\rho+\frac{e^{\lambda}}{(e^{\lambda}-1)}\right] \tag{2}$$

and

$$E(X^2)=\left[\frac{\lambda}{(e^{\lambda}-1)}((\rho+1)e^{\lambda}(1+\lambda(\rho+1))-\rho(1+\rho\lambda)\right]. \tag{3}$$

To obtain the pmf of OIIPD, we proceed as follows. Let $\pi \in (0,1)$ denote the proportion of 1's obtained from outside the generating process. Thus, $(1-\pi)$ is the proportion of counts obtained from the IPD. The pmf of a random variable Y having OIIPD can thus be written as

$$P(Y=y)=\begin{cases} \pi+(1-\pi)\dfrac{e^{-\rho\lambda}\lambda}{(e^{\lambda}-1)} & , y=1 \\ (1-\pi)\dfrac{e^{-\rho\lambda}}{(e^{\lambda}-1)}\left[\dfrac{(\rho+1)^y-\rho^y}{y!}\right]\lambda^y & , y=2,3,\ldots \end{cases} \tag{4}$$

From equation (4), it is seen that OIIPD has three parameters, namely, $\lambda > 0$ that denotes the location parameter, $\rho \in [0,\infty)$ that denotes the intervention parameter and $\pi \in (0,1)$ that denotes the inflation parameter.

2.1. Distributional Properties

Using the pmf given in equation (4), the following properties of OIIPD are obtained. The moment generating function (mgf) and the probability generating function (pgf) of Y are respectively given by

$$M_Y(t)=\pi e^t+\frac{(1-\pi)}{(e^{\lambda}-1)}e^{\lambda\rho(e^t-1)}(e^{\lambda e^t}-1) \tag{5}$$

and

$$P_Y(s)=s\pi+(1-\pi)\frac{e^{s\lambda\rho(e^{\lambda s}-1)}}{(e^{\lambda}-1)e^{\rho\lambda}}.$$

Using equation (5), the mean and variance of Y are obtained, respectively as

$$E(Y)=\mu=\pi+(1-\pi)\lambda\left[\rho+\frac{e^{\lambda}}{(e^{\lambda}-1)}\right]$$

and

$$V(Y)=\mu(1-\pi)\left[1-\lambda\rho+\frac{\lambda}{(e^{\lambda}-1)}(\lambda e^{\lambda}-e^{\lambda}+\rho^2)\right].$$

The r^{th} factorial moment of OIIPD is obtained as

$$\mu_{[r]} = \pi I_r + \frac{(1-\pi)}{(e^{\lambda}-1)}\lambda^r \left[(\rho+1)^r e^{\lambda} - \rho^r\right],$$

where $I_r = 1$ when $r = 1$ and $I_r = 0$ if $r > 1$.

3. Stochastic Representations

In this section, two SRs for the pmf given in equation (4) are presented, and their equivalence is discussed. Zhang et al. (2016) contain SRs for zero-one inflated Poisson distribution. The same methodology is adopted in the sequel.

3.1. First SR

Let Z denote a Bernoulli random variable having outcomes Z_1, Z_2. Suppose the probability of Z_1 happening is ϕ_1 and the probability of Z_2 happening is ϕ_2 i.e., $P(Z_1 = 1) = \phi_1$, $P(Z_2 = 1) = \phi_2$, $\phi_1 + \phi_2 = 1$. Let $X \sim IPD(\lambda, \rho)$ with pmf as defined in equation (1) and let $Y \sim OIIPD(\phi_1, \lambda, \rho)$. The first SR of Y is given by

$$Y = Z_1 + Z_2 X. \tag{6}$$

Note that Y takes the value one when $Z_1 = 1$ or $\{Z_2 = 1 \text{ and } X = 1\}$. Also Y takes value other than one when $\{Z_2 = 1 \text{ and } X = y\}$. Assuming X and Z are independent, the pmf of Y is obtained as

$$P(Y=y) = \begin{cases} \phi_1 + \phi_2 \dfrac{e^{-\rho\lambda}\lambda}{(e^{\lambda}-1)} & , y = 1 \\ \phi_2 \dfrac{e^{-\rho\lambda}}{(e^{\lambda}-1)} \left[\dfrac{(\rho+1)^y - \rho^y}{y!}\right]\lambda^y & , y = 2, 3, \ldots \end{cases} \tag{7}$$

Note that the pmf in equation (7) obtained through the first SR is the same as the pmf given in equation (4). The advantage of using the SR in equation (6) is that the moments of Y can be obtained easily as discussed in the next section.

3.2. Second SR

Let Z and η be two Bernoulli random variables such that $P(Z = 1) = 1 - \phi$ and $P(\eta = 1) = p$. Let $X \sim IPD(\lambda, \rho)$. Also Z, η and X are assumed to be independent. The second SR of Y is given by

$$Y = (1 - Z)\eta + ZX. \tag{8}$$

Note that Y takes the value one when $\{Z = 0, \eta = 1\}$ or $\{Z = 1, X = 1\}$. Also, Y takes value other than one when $\{Z = 1, X = y\}$.

$$P(Y=y) = \begin{cases} \phi p + (1-\phi) \dfrac{e^{-\rho\lambda}\lambda}{(e^{\lambda}-1)} & , y = 1 \\ (1-\phi) \dfrac{e^{-\rho\lambda}}{(e^{\lambda}-1)} \left[\dfrac{(\rho+1)^y - \rho^y}{y!}\right]\lambda^y & , y = 2, 3, \ldots \end{cases} \tag{9}$$

It can be observed from the right-hand side of equations (7) and (9) that,

$$\begin{cases} \phi p = \phi_1 \\ (1-\phi) = \phi_2 \end{cases} \Longleftrightarrow \begin{cases} \phi = \phi_1 \\ p = 1. \end{cases}$$

Hence the equivalence of the two SRs.

4. Moment Generating Function based on SRs

Consider the second SR of OIIPD given in equation (8). The mgf of Y is given by

$$\begin{aligned} M_Y(t) &= E(\exp(tY)) \\ &= E\left\{\exp[t(1-Z)\eta + tZX]\right\} \\ &= E_Z\left[E_Y\left\{\exp[t(1-Z)\eta + tZX]|Z\right\}\right] \\ &= E_Z\left[E_Y\left(e^{t(1-Z)\eta}e^{tZX}|Z\right)\right] \\ &= E_Z\left[M_\eta(t(1-Z))M_X(tZ)\right] \\ &= E_Z\left[\{(1-p)+pe^{t(1-Z)}\}\frac{e^{\rho\lambda(e^{tZ}-1)}(e^{\lambda e^{tZ}}-1)}{(e^\lambda-1)}\right] \\ &= \phi[(1-p)+pe^t] + (1-\phi)\frac{e^{\rho\lambda(e^t-1)}(e^{\lambda e^t}-1)}{(e^\lambda-1)}. \qquad (10) \end{aligned}$$

Using the equivalence of the two SRs, substituting $p = 1$ and taking $\phi = \phi_1$ in equation (10), the mgf of Y based on the first SR can be obtained as below.

$$M_Y(t) = \phi_1 e^t + \phi_2\frac{e^{\rho\lambda(e^t-1)}(e^{\lambda e^t}-1)}{(e^\lambda-1)}.$$

From equation (6), using the binomial expansion, we get

$$E(Y^r) = \phi_1 + \phi_2 E(X^r), r = 1, 2, \ldots \qquad (11)$$

Using the equations (2), (3) and (11), the first two moments of Y are respectively obtained as

$$E(Y) = \phi_1 + \phi_2\lambda\left[\rho + \frac{e^\lambda}{(e^\lambda-1)}\right]$$

and

$$E(Y^2) = \phi_1 + \phi_2\left[\frac{\lambda}{(e^\lambda-1)}((\rho+1)e^\lambda(1+\lambda(\rho+1)) - \rho(1+\rho\lambda)\right].$$

Thus,

$$V(Y) = E(Y)\phi_2\left[1 - \lambda\rho + \frac{\lambda}{(e^\lambda-1)}(\lambda e^\lambda - e^\lambda + \rho^2)\right].$$

5. Parametric Estimation

5.1. Method of Maximum Likelihood

Let $\vec{y} = (y_1, y_2, \ldots y_n)$ be a sample of n iid observations from $OIIPD(\pi, \lambda, \rho)$. Let m denote the number of 1's in the sample and $(n-m)$ denote the number of observations taking values other than one. The likelihood function of (π, λ, ρ) corresponding to the pmf given in equation (4) is

$$L(\pi, \lambda, \rho|\vec{y}) = \left[\pi + (1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}\right]^m \times (1-\pi)^{n-m} \prod_{i=1}^{n-m} \frac{e^{-\rho\lambda}}{(e^{\lambda}-1)} \left[\frac{(\rho+1)^{y_i} - \rho^{y_i}}{y_i!}\right] \lambda^{y_i}.$$

The corresponding log-likelihood function is

$$\begin{aligned} l(\pi, \lambda, \rho|\vec{y}) = {} & m \ln\left[\pi + (1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}\right] + (n-m)\left[\ln(1-\pi) - \rho\lambda - \ln(e^{\lambda}-1)\right] \\ & + \ln(\lambda)\sum_{i=1}^{n-m} y_i + \sum_{i=1}^{n-m} \ln\left((\rho+1)^{y_i} - \rho^{y_i}\right) - \sum_{i=1}^{n-m} \ln(y_i!). \end{aligned}$$

The score functions of the parameters (π, λ, ρ) are respectively obtained as below.

$$\frac{\partial l}{\partial \pi} = \frac{m\left(1 - \frac{\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}\right)}{\pi + \frac{\lambda \mathrm{e}^{-\rho\lambda}(1-\pi)}{(\mathrm{e}^{\lambda}-1)}} - \frac{n-m}{1-\pi}, \tag{12}$$

$$\frac{\partial l}{\partial \lambda} = \frac{m\left(-\frac{(1-\pi)\lambda \mathrm{e}^{\lambda(1-\rho)}}{\left(\mathrm{e}^{\lambda}-1\right)^2} - \frac{(1-\pi)\rho\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)} + \frac{(1-\pi)\mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}\right)}{\pi + \frac{(1-\pi)\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}} + (n-m)\left(-\frac{\mathrm{e}^{\lambda}}{(\mathrm{e}^{\lambda}-1)} - \rho\right) + \frac{1}{\lambda}\sum_{i=1}^{n-m} y_i, \tag{13}$$

$$\frac{\partial l}{\partial \rho} = \frac{m\,(1-\pi)\,\lambda^2 \mathrm{e}^{-\lambda\rho}}{(\mathrm{e}^{\lambda}-1)\left(\pi + \frac{(1-\pi)\lambda \mathrm{e}^{-\lambda\rho}}{(\mathrm{e}^{\lambda}-1)}\right)} + \sum_{i=1}^{n-m}\left[\frac{y_i\,(\rho+1)^{y_i-1} - y_i \rho^{y_i-1}}{(\rho+1)^{y_i} - \rho^{y_i}}\right] - (n-m)\,\lambda. \tag{14}$$

Equating the score functions in equations (12), (13) and (14) to zero and solving them simultaneously, the ML estimates of the parameters (π, λ, ρ) are obtained provided the Hessian matrix evaluated at the ML estimates is negative definite. Since the score functions are nonlinear in the parameters, one has to use numerical methods to obtain the ML estimates. To ease out the computation, in the sequel, the parameters are estimated using the EM algorithm by treating the 1's coming from the degenerate distribution as latent.

5.2. ML Estimation via EM Algorithm

Let us assume the 1's from OIIPD are from two distributions, namely, the degenerate distribution and the IPD. Let U be the latent variable that denotes the number of 1's from the degenerate distribution. Suppose there are a total of m 1's observed, then $(m-U)$ 1's are from the IPD. Thus, the distribution of U given Y is Binomial(m, p) , where

$$p = \frac{\pi}{\pi + (1-\pi)\dfrac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}}.$$

The likelihood based on the complete sample $Y_{comp} = (Y, U)$ is proportional to

$$L(\pi, \lambda, \rho | Y_{comp}) \propto \pi^u \left[(1-\pi) \frac{\lambda e^{-\rho\lambda}}{(e^\lambda - 1)} \right]^{m-u} \times (1-\pi)^{n-m} \left(\frac{e^{-\rho\lambda}}{(e^\lambda - 1)} \right)^{n-m} \lambda^N \prod_{i=1}^{n-m} \left[(\rho+1)^{y_i} - \rho^{y_i} \right],$$

taking $N = \sum_{i=1}^{n-m} y_i$. The corresponding log-likelihood function is thus proportional to

$$\begin{aligned} l(\pi, \lambda, \rho | Y_{comp}) \propto\ & u \ln(\pi) + (n-u) \ln(1-\pi) + (n-u)(-\rho\lambda) - (n-u) \ln(e^\lambda - 1) \\ & + (N+m-u) \ln(\lambda) + \sum_{i=1}^{n-m} \ln([(\rho+1)^{y_i} - \rho^{y_i}]). \end{aligned} \tag{15}$$

In the E-step of the EM algorithm, the latent U is estimated as

$$\hat{u} = \frac{m\pi}{\pi + (1-\pi)\dfrac{\lambda e^{-\rho\lambda}}{(e^\lambda - 1)}}. \tag{16}$$

The ML estimates of the parameters, namely, $\hat{\pi}, \hat{\lambda}$ and $\hat{\rho}$ are obtained using the complete log-likelihood given in equation (15) through the M-step of the EM algorithm by solving the following simultaneous equations.

$$\hat{\pi} = \frac{\hat{u}}{n}, \tag{17}$$

$$\hat{\lambda} = \frac{1}{(n-\hat{u})} \sum_{i=1}^{n-m} \frac{[y_i(\hat{\rho}+1)^{y_i-1} - y_i \hat{\rho}^{y_i-1}]}{[(\hat{\rho}+1)^{y_i} - \hat{\rho}^{y_i}]}, \tag{18}$$

and

$$\hat{\rho} = \frac{(N-m-\hat{u})}{(n-\hat{u})\hat{\lambda}} - \frac{e^{\hat{\lambda}}}{e^{\hat{\lambda}} - 1}. \tag{19}$$

The E and the M steps in the equations (16) to (19) are repeated till the estimates converge. To start the iterative procedure, initial values of the parameters, say $\pi^{(0)}, \lambda^{(0)}$ and $\rho^{(0)}$ need to be specified. The advantage of using the EM algorithm is that the estimators have closed-form expressions, unlike the ML method, making the computations easier.

6. Numerical Illustration

The application of the proposed OIIPD to a real-life dataset is illustrated in this section. We consider the data on an epidemic of cholera in a village in India used in Shanmugam (1985) to fit a intervened Poisson distribution. The data relate to the spread of cholera in an Indian village and was earlier reported in McKendrick (1926). The data was observed when preventive treatment to contain the spread of cholera had been initiated. The data excluding the households not affected by cholera is tabulated below.

x	1	2	3	4+	Total
f_x	32	16	6	1	55

Here, x denotes the number of cholera cases, and f_x denotes the number of households with x cases. The primary reason for many households having cholera cases was attributed to one particular infected well which was used by a large section of the people in the village. However, other wells near its vicinity can also be the source of infection. Since the frequency of the number of households having one cholera case is large, OIIPD model is used to fit the data. The EM-algorithm steps given in the previous section are implemented to estimate the parameters (π, λ, ρ) by fixing their initial values as $(0.01, 0.5, 0.5)$ respectively. The initial values of the parameters λ and ρ were fixed near to their moment estimates obtained thorough the intervened Poisson model. The difference between the proportion of the observed and the expected 1's (rounded to two decimals) based on intervened Poisson model is taken as the initial value of π. The final estimates of the parameters (π, λ, ρ) are obtained as $(0.0099, 0.7050, 0.2492)$. From the estimate of π, it is clear that the proportion of 1's emerging from outside the IPD is small. This means that the primary source of the spread of cholera among the people in the village is the particular infected well. Also, a small value of the estimate of the intervention parameter ρ suggests that the preventive mechanism had a considerable effect in bringing down the number of cholera cases per household.

7. Concluding Remarks

The OIIPD introduced in this paper not only accounts for the excess 1's but also provides information on the effectiveness of the intervention mechanism. The two equivalent stochastic representations of the model given in this work provide an efficient way to derive the moment generating function and moments of OIIPD. EM algorithm approach is used to estimate the model parameters, circumventing the need to solve simultaneously the nonlinear equations given by the ML method. The proposed distribution can be used to model count data process altered by an intervention mechanism resulting in 1's with high frequency.

References

Godwin, R. T. and Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66(2)**, 425-448.

Johnson, N. L., Kotz, S. and Kemp, A. W. (2005). *Univariate Discrete Distributions.* John Wiley & Sons.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34(1)**, 1-14.

Melkersson, M. and Olsson, C. (1999). Is visiting the dentist a good habit?: Analyzing count data with excess zeros and excess ones. *Umeå Economic Studies*, **492**, 1–18.

McKendrick, A.G. (1926). Application of Mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society*, **44**, 98-130.

Shanmugam, R. (1985). An intervened Poisson distribution and its medical application. *Biometrics*, 1025-1029.

Zhang, C., Tian, G. L. and Ng, K. W. (2016). Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods. *Statistics and its Interface*, **9(1)**, 11-32.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 91-106

Profit to Rice Growers at National Level and Jammu Region of J&K UT: Implication to Doubling the Farmers Income

Sunali Mahajan and Manish Sharma
Division of Statistics and Computer Science, SKUAST- Jammu (*180009*)

Received: 15 June 2022; Revised: 15 July 2022; Accepted: 21 July 2022

Abstract

Quantile regression (QR) has more scope in economic analysis. Economics data are usually contaminated, and the error assumptions are generally violated. QR can characterize the entire conditional distribution of the outcome variable, may be more robust to outliers and misspecification of error distribution, and provides robust and complete estimates compared to the mean regression. These advantages make QR attractive and are extended to apply for different types of data sets like agricultural, economic, and financial. The QR method has been applied to study the production and profit of rice as endogenous variable(s) and observed that by focusing on the variables fertilizer consumption(FC), quality seed rate (QSR), electricity consumption (EC), sale of power tillers (SPT), cost of seed (CS), cost of fertilizer and manure (CFAM), cost of insecticide (COI) and irrigation charges (IC), the production and profit of rice may be maximized at national level as well as in Jammu region. Further, the time series model has been applied on the data and observed that the best fitted model for production and profit were ARIMA (0 2 2) and ARIMA (0 1 1) based on AIC and SBIC criterion. The gain in profit percentage w.r.t cost has also been calculated for the year 2025 and shall be 136.62 percent. It is observed that the government is taking initiative by implementing the policies to enhance the income of the famers. From the study, it has been observed that, the variables cost of seed (CS), cost of fertilizer and manure (CFAM), cost of insecticide (COI) and irrigation charges (IC) may help the government to enhance the income two - three times in the coming years.

Key words: ARIMA; Box-Jenkins; Quantile regression model; Profit function; Fixed cost; Forecast.

1. Introduction

Agriculture plays a key role in the overall economic and social well-being in India. Economic growth in India has been broadly on an accelerating path. In 1951, the total National income coming from agriculture sector was 55 percent but now-a-times, it is stuck to mere 24 percent. The value addition per worker in agriculture grew slowly and income per farmer never crossed one-third of the income of a non-agriculture worker since 1980s. More than 65 percent population of India depends on rice (staple food), contributing 40 percent of the total food grain and plays a major role in diet, economy, employment, culture, and history. Total production of Rice during 2020-21 is estimated at record 121.46 million tonnes. It is higher by 9.01 million tonnes than the last five years' average production of 112.44 million tonnes.

Corresponding author: Sunali Mahajan
Email: sunali12mahajan@gmail.com

As far as Jammu is concerned, Jammu and Kashmir is basically an agrarian UT with 80 percent of the people engaged in agriculture for their livelihood. The current situation is not satisfactory in terms of food grains as the area under these crops have shown the disturbing trends and around 3.5 hectares area has been converted for commercial and other purposes which is being revealed by the Department of Agriculture, despite the fact that there is a sealing act on paddy land by the government for its conversion to some other activities which has caused the food deficiency in J&K and has already touched to 40 percent. Further, the increase in the area under these crops is less as they are seeming to be profitable (Present Status and Future Prospectus of Agriculture in Jammu and Kashmir, January 2021 DOI:10.9790/0837-201136267). As a matter of fact, J&K is not sufficient to feed its own people as a result a large quantity of rice (on an average 4.97 lakh tones) in a year are drawn from central pool to meet the deficient requirement of the UT. Above data reveals the clear deficiency of food grains in Jammu region of J&K (UT) as compared to National Level.

So far considering the above points, to overcome the deficiency of rice and to identify the parameters which will double the farmers' income at National as well as Jammu region. One must evaluate and study the regressor variables to maximize the production and profit of rice at National level and in Jammu region through robust models. Further, the constraints have been identified which the farmers' faced during the production and profit maximization of rice in Jammu region and to forecast the value(s) of production and profit of rice crop at National level.

2. Materials and Methods

In the present study, the data have been collected into two ways *i.e.*, Primary for Jammu region and Secondary at National level. The time series data over decades have been collected with respect to rice from various published sources/portals such as: Ministry of Agriculture, Government of India; Directorate of Economics and Statistics, Government of India; Reserve Bank of India; IASRI (Data Book) and Indiastat. For Primary data, from the Jammu region, two districts (Jammu and Kathua) selected purposively due to maximum rice growers; from each district 6 villages have been selected randomly and then from each village 10 farmers have been selected randomly. So, there were 120 farmers who have been surveyed through multistage sampling procedure for Jammu region of J&K UT. Ordinary least square (OLS), Quantile Regression (QR) and ARIMA technique(s) have been applied on the annual decadal data of production and profit of rice crop. QR presents a complete understanding of the effects of exogenous variables when a set of percentiles are modeled. QR is particularly useful when the rate of change in the conditional quantile, revealed by the coefficients of the regression, and extreme values are very important (Benhin 2008, Cameron and Trivedi 2009). QR model have been used to identify variables which may maximize the profit for rice growers. Just as the sample mean that minimizes the sum of squared residuals: $\hat{\varepsilon} = arg\ \min_{\varepsilon\epsilon\mathrm{R}} \sum_{i=1}^{n}(y_i - \varepsilon)^2$, *is* extended to the linear mean function $\mathrm{E}(Y|X = x) = x'/\beta$ by the solution of $\hat{\beta} = argmin_{\beta\epsilon\mathbb{R}^{\mathrm{p}}} \sum_{i=1}^{n}(y_i - x_i'\beta)^2$. The linear conditional quantile function, $(\tau|X = x) = x'\beta(\tau)$ can be estimated by the solution of $\hat{\beta}(\tau) = arg\ \min_{\beta\epsilon\mathbb{R}^{\mathrm{p}}} \sum_{i=1}^{n}\rho_\tau(y_i - x_i'\beta)$, for any quantile $\tau \epsilon (0,1)$.

Moreover, Information regarding the problems faced by the farmers in rice production, and profit have been surveyed. Farmers were being asked to choose one of the five choices *i.e.*, strongly disagree, disagree, neutral, agree, strongly agree, then rank the problems which were proposed to them in a schedule and were identified by giving them ranks using Garrett's Ranking Technique (GRT). Garrett's formula for converting ranks into percent is

$\text{Percent position} = 100 * \frac{R_{ij}-0.5}{N_j}$ where, R_{ij} = rank given for i^{th} constraint by j^{th} individual, N_j = number of constraints ranked by j^{th} individual. The percent position of each rank be converted into scores by using table of Garrett and Woodworth (1969). For each factors, the scores of individual respondents be added together and divided by the total number of the respondents for whom scores be added. These mean scores for all the constraints be arranged in descending order; the constraints were accordingly ranked. After this, the main issue(s) of our concern is that how to select an appropriate model that can produce accurate forecast based on a description of historical pattern in the data and how to determine the optimal model orders. Box and Jenkins (1976) developed a practical approach to build ARIMA model, which is helpful in best fitting the given time series and satisfy the parsimony principle. In ARIMA models, a non-stationary time series is made stationary by applying finite differencing of the data points. The mathematical formulation of the ARIMA ($p\ d\ q$) model using lag polynomials is

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \ i.e;\ 1-\sum_{i=1}^{p}\varphi_i L^i\ (1-L)^d y_t = 1+\sum_{j=1}^{q}\theta_j L^j\ \varepsilon_t$$

where

$$AR(p)model\colon \varepsilon_t = \varphi(L)y_t;$$
$$MA(q)model\colon y_t = \theta(L)\varepsilon_t$$
$$ARMA(p\ q)\ model\colon \varphi(L)y_t = \theta(L)\varepsilon_t.$$

Here, $\varphi(L) = 1-\sum_{i=1}^{p}\varphi_i L^i\ and\ \theta(L) = 1+\sum_{j=1}^{q}\theta_j L^j$. The Box-Jenkins methodology does not assume any pattern in the historical data of the series to be forecasted. Rather, it uses a three-step iterative approach of model identification, parameter estimation and diagnostic checking to determine the best parsimonious model from a general class of ARIMA models. These three-step processes were repeated several times until a satisfactory model finally selected. A crucial step in an appropriate model selection is the determination of optimal model parameters. One criterion is that the sample ACF and PACF, other widely used measures for model identification are Akaike Information Criterion (AIC) and Schwartz Bayesian Information Criterion (SBIC) which are defined as AIC(p) = $n\ln(\hat{\sigma}_e^2/n) + 2p$ and SBIC(p) = $n\ln(\hat{\sigma}_e^2/n) + p + p\ln(n)$. Here, n is the number of effective observations, used to fit the model, p is the number of parameters in the model and $\hat{\sigma}_e^2$ is the sum of sample squared residuals. The optimal model order is chosen by the number of model parameters, which minimizes AIC and BIC. The final selected model is used for forecasting future values of the time series.

3. Results and Discussion

The summary statistics for exogenous variables area under rice (AUR), fertilizer consumption (FC), quality seed of rice (QSR), electricity consumption (EC), annual rainfall (AR), pesticide consumption (PC), sale of tractors (ST) and sale of power tillers (SPT) in case of production of rice (POR) whereas in case of profit of rice (PrOR), the exogenous variables were cost of machine labour (CML), cost of seed (CS), cost of fertilizer and manure (CFAM), cost of insecticides (COI), irrigation charges (IC) and fixed cost (FC) with the values of mean, standard error as well as coefficient of variation as shown by table 1.

Table 1: Summary statistics of exogenous variables for the production and profit of rice in India

Production(Million ton)				Profit (Rs/hectare)			
Variable (Unit)	**Mean**	**±SE $(\bar{X})$**	**CV (in %)**	**Variable (Rs/hectare)**	**Mean**	**±SE $(\bar{X})$**	**CV (in %)**
AUR (Million hectare)	43.40	0.23	2.64	**CML**	19721.81	2387.11	40.14
FC (Thousand ton)	19870.53	1072.45	26.44	**CS**	3039.09	380.25	41.49
QSR (Lakh quintal per hectare)	38.73	4.97	62.89	**CFAM**	1548.38	169.27	36.25
EC (Giga-watt)	102008.80	5546.62	26.63	**COI**	3135.96	325.94	34.47
AR (Millimeter)	1154.82	17.51	7.42	**IC**	633.54	84.49	44.23
PC (Million ton)	50335.91	1716.31	16.70	**FC**	1009.57	97.86	32.14
ST (Number)	331702.80	33383.87	49.30				
SPT (Number)	26197.08	3526.91	65.95				

Here, variable sale of power tillers (SPT) has maximum CV (65.95 percent) followed by quality seed of rice (QSR) whereas the variable area under rice (AUR) has minimum CV (2.64 percent) followed by the variable annual rainfall (AR) which clearly indicates lot of variation among the independent variables may be due to the presence of influential observations. In case of profit, the variable IC has maximum value of CV (44.23 percent) followed by CML and the variable FC has minimum CV (32.14 percent) followed by COI.

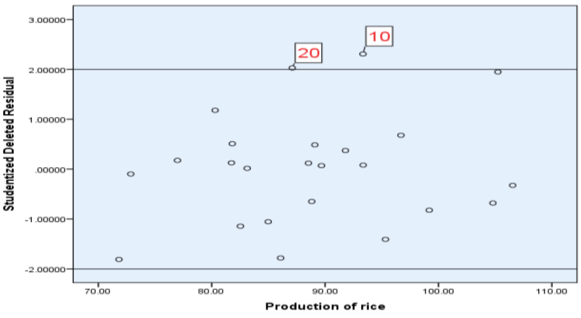


Figure 1: Graph of studentized deleted residual against endogenous variable

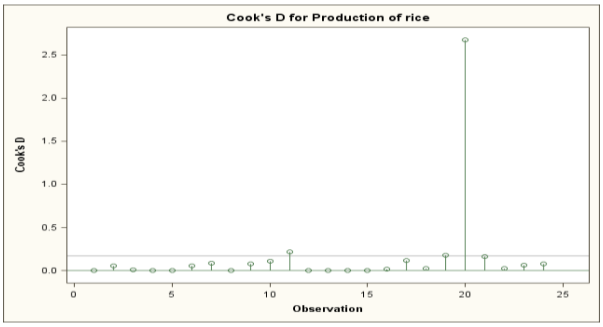


Figure 2: Graph of cook's distance for the production of rice in India

The behaviour of the data has been evaluated through studentized deleted residual, Cook's Distance, Breusch-Pagan test and Durbin Watson test. It was concluded from figure 1 that, the observations which were coming out of the range -2 to +2 were the outliers by studentized deleted residual and from figure 4, the observations which showed sudden jumps in the graph were influential observations by Cook's Distance. Moreover, Breusch-Pagan test and Durbin Watson test showed that there is no heteroscedasticity and no autocorrelation present in the data as their Lagrange multiplier (LM) value is 0.96 (p-value = 0.99) which was non-significant and the value of D = 2.36, respectively. Moreover, the values of R^2 and adj. R^2 be 0.93 and 0.89, which depicts the model is going to be good fitted and the F-value (26.60**) which means the model is adequate.

Table 2: Estimation of regression coefficients through OLS and quantile regression at different quantiles using Cobb-Douglas production function for India

Variable	Regression coefficients				Gain in Percentage
	OLS (SE)	Quantile regression			
		τ = 0.50 (SE)	τ = 0.75 (SE)	τ = 0.90 (SE {E-08})	τ = 0.90 w.r.t OLS
Constant	−0.9372 (1.9536)	−1.1684 (0.6752)	1.8827 (0.0892)	−2.8305 (0.0186)	
AUR	1.2634** (0.3146)	1.3178** (0.1087)	0.9218** (0.0144)	1.8009** (0.2990)	42.5439
FC	−0.1799 (0.1698)	−0.2175** (0.0587)	−0.2697** (0.0078)	−0.2296** (0.1610)	27.6264
QSR	0.2246** (0.0390)	0.2115** (0.0135)	0.2093** (0.0019)	0.3486** (3.7100)	55.2092
EC	−0.1050 (0.0885)	−0.0776* (0.0306)	−0.0673** (0.0040)	−0.1510** (8.4200)	43.8095
AR	0.3936** (0.1145)	0.4216** (0.0396)	0.2753** (0.0052)	0.3853** (0.1090)	-2.1087
PC	−0.0424 (0.0703)	−0.0681* (0.0243)	−0.0688** (0.0032)	0.0976** (6.6800)	-130.1886
ST	0.0333 (0.0664)	0.0502* (0.0229)	0.0312** (0.0030)	−0.0520** (6.3100)	-256.1561
SPT	0.0121 (0.0756)	0.0114 (0.0261)	0.0257** (0.0035)	0.0230** (7.1900)	90.9090
Returns to scale	1.9270	2.0125	1.4633	2.6555	

***= significant at 5% and **= significant at 1%**

The regression coefficient by OLS indicates that the variables *i.e.*; AUR, QSR and AR were statistically significant. A detailed representation of the parameter coefficients at quantiles 0.50, 0.75 and 0.90 have been revealed by Table 2. The variables AUR, FC, QSR, EC and SPT were significant and maximum at τ = 0.90th as compared to OLS. The gain in magnitude for the variables AUR, FC, QSR, EC and SPT were as 42.54 percent, 27.62 percent, 55.21 percent, 43.80 per cent and 90.91 percent which can maximize the production of rice in India. Moreover, after the 0.90th quantile, the estimate of the parameter remains constant indicated that there is no more effect of multicollinearity, outliers and influential observations on the data. The value of returns to scale at τ = 0.90th was 2.65 which means production increases with the increase in all inputs at τ = 0.90th.

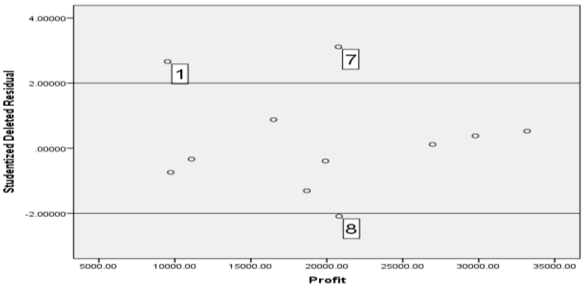


Figure 3: Graph of studentized deleted residual against endogenous variable

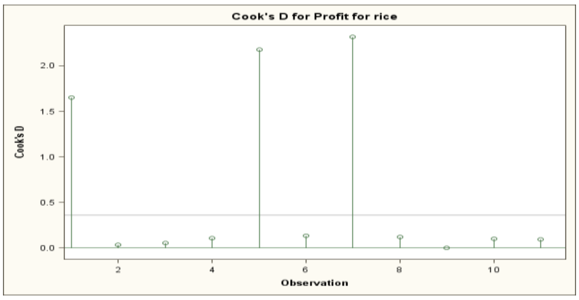


Figure 4: Graph of cook's distance for profit of rice crop in India

It has been observed from the figures 3 and figure 4 that, the observations which were coming out of the range -2 to +2 were the outliers by studentized deleted residual and the observations which showed the sudden jumps in the graph were influential observations by Cook's distance. Breusch-Pagan test and Durbin Watson test showed that there was no heteroscedasticity present in the data as its LM value is 3.98 with p value = 0.67 but the autocorrelation was present in the data as the value of *D* = 2.72 respectively. Moreover, the values of R^2, adj. R^2 and F-value were 0.99, 0.98 and 83.73**, which is significant, and the model is adequate for study.

Table 3: Estimation of regression coefficients through quantile regression at different quantile using Cobb-Douglas profit function for India

Variable	Regression coefficients			Gain in Percentage
	OLS (SE)	Quantile regression		
		τ = 0.5 (SE)	τ = 0.75 (SE)	τ = 0.75 w.r.t OLS
Constant	−3.9688 (1.7781)	−6.2959** (0.3313)	−2.4551* (0.6842)	
CML	−1.2604 (0.4818)	−1.9825** (0.0898)	−1.2361** (0.1854)	-1.9280
CS	0.7717 (0.3917)	0.8347** (0.0730)	0.9626** (0.1507)	24.7376
CFAM	0.4402 (0.2208)	0.6098** (0.0411)	0.5750** (0.0850)	30.6224
COI	0.1002 (0.1519)	0.1278* (0.0283)	0.1866* (0.0584)	86.2275
IC	−0.7957** (0.1577)	−0.8360** (0.0294)	−0.8987** (0.0606)	12.9446
FC	2.0879* (0.4923)	2.7670** (0.0917)	1.6602** (0.1894)	-20.485
Returns to scale	3.4000	4.3393	3.3844	

***= significant at 5% and **= significant at 1%**

After the model coming out to be significant, the regression coefficient has been evaluated by OLS and at different quantiles. The variables IC and FC were statistically significant by OLS with the value of 0.79 and 2.08. The result in Table 3 showed that the cost of variables CS, CFAM, COI and IC was significant and the increase in magnitude of these variables at $\tau = 0.75^{th}$ as compared to OLS were 24.73, 30.62, 86.22 and 12.94 percent may maximize the profit of rice in India. Moreover, after the 0.75^{th} quantile, the estimates of the parameter remain constant. The value of returns to scale at $\tau = 0.75^{th}$ was 3.38 indicating that the profit increases with the increase in all inputs at 0.75^{th} quantile.

After the identification of production and profit of rice variables at National level, the survey has been conducted in the two districts (Jammu and Kathua) of Jammu region of J&K UT. From these two districts, 120 farmers have been surveyed and their socio-economic status is as:

Table 4: Socio-economic status of the farmers of rice crop in Jammu and Kathua districts of Jammu region

Variable	**Category**	**District**		**χ^2-test (*p*-value)**
		Jammu (%)	**Kathua (%)**	
Age (in years)	<30	2 (1.70)	1 (0.80)	6.86 (0.03)
	30-50	17 (14.20)	31 (25.80)	
	>50	41 (34.20)	28 (23.30)	
Schooling (Classes passed)	0-8	18 (15.00)	9 (7.50)	5.02 (0.17)
	Below matric	13 (10.80)	13 (10.80)	
	Matric	17 (14.20)	18 (15.00)	
	Above matric	12 (10.00)	20 (16.70)	
Occupation (Agricultural Farming)	Primary	38 (31.70)	39 (32.50)	0.04 (0.84)
	Secondary	22 (18.30)	21 (17.50)	

The socio-economic status of farmers of Jammu and Kathua districts of Jammu region of Jammu and Kashmir UT has been presented in Table 4. The result showed that the variable age was significant as their χ^2 value was 6.86. The farmer with age more than 50 were more indulged in farming *i.e.*, 34.20 percent followed by age group 30-50 in Jammu district, the reason behind this may be that the youth is moving towards private sector rather than agriculture sector whereas in Kathua district, the farmers with age group 30-50 were more indulged in farming *i.e.*, 25.8 percent followed by age group more than 50. The farmers with age less than 30 were less indulged in farming in both the districts.

Further, the variable schooling was non-significant with χ^2 value as 5.02 but associated with the districts. Also, the variable occupation was divided into two categories as primary farming and secondary farming which was again not significant with χ^2 value as 0.04 and is associated with districts. The production and profit of rice in Jammu division have been studied and the variables have been identified for knowing the status in case of rice in J&K UT.

The exogenous variables in case of production were area under rice (AUR), seed rate (SR), fertilizer consumption (FC), labour for rice (LFR) and herbicide consumption (HC) whereas in case of profit, the variables were cost of harvesting, threshing & winnowing (CHTW), cost of seed (CS), cost of fertilizer (COF), cost of labour (COL), cost of herbicide (COH), cost of irrigation (COI) and cost of transplanting (COT). The graph of studentized deleted residuals plotted against endogenous variable as shown in figure 5, it was concluded that the observations which were coming out of range -2 to +2 were outliers.

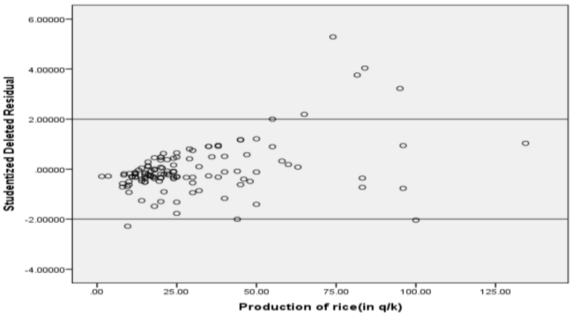


Figure 5: Graph of studentized deleted residual against endogenous variable

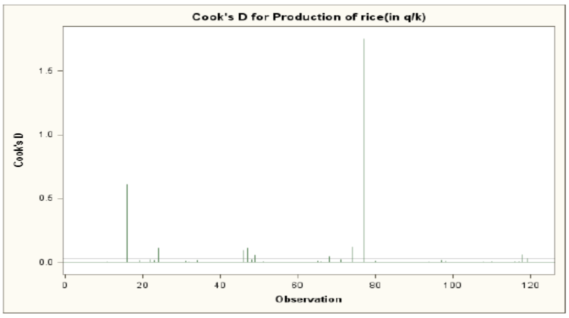


Figure 6: Graph of cook's distance for the production of rice in Jammu region

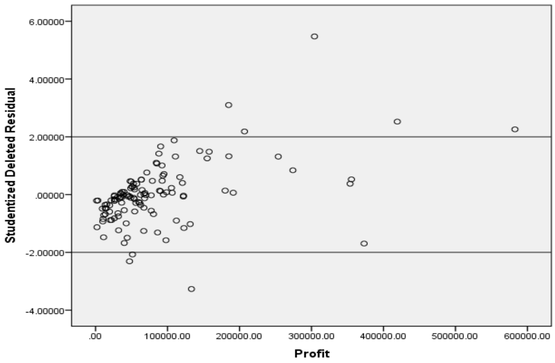
Also, from the graph of Cook's distance as shown in figure 6, it can be observed that some of the observations have the large value of cook's distance and also showed sudden jumps, so were influential observations. The observations (16, 47 and 74) which were both outliers and influential observations affect both the slope and intercept. Breusch-Pagan test indicated that the heteroscedasticity was present in the data as their LM value was 23.47 (*p*-value=0.0001) which was statistically significant and Durbin Watson test showed that the autocorrelation was not present in the data as the D value = 1.56. The values of R^2 and adj R^2 be 0.81, 0.80 and the F-value was 96.47** which showed that the model was significant. Thus, the exogenous variables considered for the study were adequate.

The behaviour of the data disobey the assumptions of error term and the regression coefficient by OLS indicated that the variable SR was statistically significant with the value 0.14 as shown in Table 5 whereas; quantile regression illustrated a positive and statistically significant effect of AUR on the production at quantiles 0.5th and 0.95th. Moreover, after the 0.95th quantile, the estimates of the parameter remain constant indicated that there was no more effect of multicollinearity, outliers and influential observations on the data. The variables AUR was significant based on sign, size and significance and maximum at τ = 0.95th with the increase in magnitude as 74.73 percent as compared to OLS which can maximize the production of rice in Jammu division. The value of returns to scale at τ = 0.95th was 2.45 which increases the production of rice with the increase in all inputs at 0.95thquantile. The behaviour of profit of rice growers in Jammu region have also been studied with the following graphs and tests.

Table 5: Estimation of regression coefficients through OLS and quantile regression at different quantiles using Cobb-Douglas production function for Jammu region of J&K UT

Variable (per kanal)	Regression coefficients				Gain in Percentage
	OLS	Quantile regression			
		τ = 0.50	τ = 0.75	τ = 0.95	τ = 0.95 w.r.t. OLS
Constant	1.1800	0.6600	1.4902	3.4188*	
AUR (Quintal)	1.1860	0.9469*	1.4138	2.0724**	74.7386
SR (Kilogram)	0.1496*	0.1682**	0.0841	0.1466	-2.0053
FC (Kilogram)	0.0797	0.1599	0.2378	0.2397	200.7528
LFR (Number)	−0.5645	−0.3960	−0.9095	−1.4711	160.6023
HC (Milliliter)	0.0395	0.0380	0.0485	−0.1086	-374.937
Returns to scale	1.4548	1.3130	1.7842	2.4587	

***= significant at 5% and **= significant at 1%**



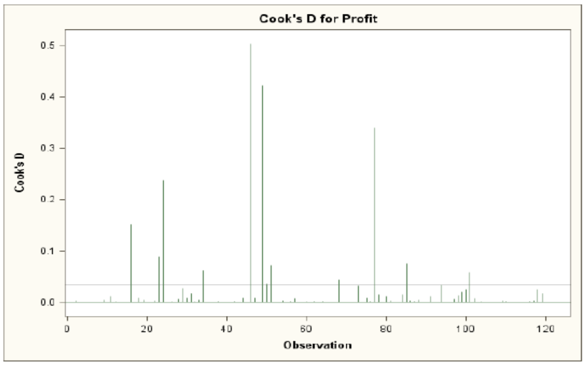


Figure 7: Graph of studentized deleted residual against endogenous variable

Figure 8: Graph of cook’s distance for profit of rice crop in Jammu region

The values of studentized deleted residual were used to construct the plot against endogenous variable and resulted that the some of the observations were the outliers as shown in Figure 7 whereas, from the Figure 8, it can be observed that the observations were coming out of cut off line were influential observations as these observations showed sudden jumps in the graph. The some observations (24, 46, 49 and 77) were both outliers and influential observations which affect both the slope and intercept. Breusch-Pagan test and Durbin Watson test showed that both heteroscedasticity and autocorrelation were present in the data with LM value as 102.96 (*p*-value = 0.0001) and D = 1.22 respectively. Moreover, the values of R^2 and adj R^2 are 0.61 and 0.63, depicting that the model was going to be good fit and the F-value was (27.03), which was found to be significant.

Table 6: Estimation of regression coefficients through OLS and quantile regression at different quantiles using Cobb-Douglas profit function for Jammu region of J&K UT

Variable (per kanal)	Regression coefficients				Gain in Percentage
	OLS	Quantile regression			
		τ = 0.50	τ = 0.75	τ = 0.95	τ = 0.95 w.r.t. OLS
Constant	4.4279**	4.6597**	3.5393**	5.1133**	
CHTW	0.0708	0.0557	0.1709	0.1115**	-88.85
COS	−0.0396	−0.1201*	−0.0668	−0.1530**	286.36
COT	0.9707**	1.3896**	0.7104**	0.3975**	-60.25
COF	0.3543	−0.0016	0.0835	0.3520**	-64.80
COI	0.1445	0.0737	0.2619**	0.1668**	-83.32
COL	−0.5076**	−0.6375**	−0.1540	0.0323	-106.36
COH	−0.1210	0.0144	−0.0798	−0.0946**	-21.81
Returns to scale	0.8721	0.7742	0.9261	0.8125	

***= significant at 5% and **= significant at 1%**

The behaviour of the data showed that there is a violation of the error assumptions and the regression coefficients by OLS method mislead the results as shown in Table 6. So, to overcome from such situation, quantile regression has been applied and illustrated a positive, negative and statistically significant effect of the variables at 0.95 quantile. The overall result reveals that the variable COS was significant at τ = 0.95th and the gain in magnitude was286.36 as compared to OLS which can maximize the profit to rice growers in Jammu region of J&K UT. Moreover, after the 0.95th quantile, the estimates of the parameter remain constant. The value of returns to scale at τ = 0.95th was 0.81 which means there is slight increase in profit with the increase in all inputs at 0.95th quantile.

The constraints have also been studied which the farmers' faced during the time of production and profit of the rice crop in Jammu and Kathua districts as shown in table 7. The foremost constraint faced by farmer were scarcity of labour during peak season in both the districts with Garrett values as 67.60 and 67.00 followed by non-availability of power machines at appropriate time and lack of electricity whereas; in case of profit, the major constraint faced by the farmer was rate fluctuation with Garrett values as 70.40 followed by lack of storage facility and high rate of interest in both the districts.

Table 7: Garrett ranking of the constraints for production and profit of rice in Jammu region

Production			Profit		
Factors	**Constraints**	**Rank (Garrett value)**	**Factors**	**Constraints**	**Rank (Garrett value)**
F1	Shortage of quality seed	4 (53.70)	F1	Crop failure	9 (44.00)
F2	Non availability of suitable HYV	5 (51.65)	F2	Lack of market facility	8 (55.15)
F3	Non availability of fertilizer in adequate quantity	7 (47.05)	F3	Intervention of middle man	4 (63.90)
F4	Shortage of herbicide	8 (42.65)	F4	Lack of storage facility	2 (67.90)
F5	Scarcity of labour during peak season	1 (67.60)	F5	Lack of transport facility	7 (57.10)
F6	Lack of irrigation facility	6 (49.65)	F6	Village far from market	6 (59.35)
F7	Lack of electricity	3 (56.05)	F7	High rate of interest	3 (66.30)
F8	Non availability of power machines at appropriate time	2 (59.35)	F8	Lack of delivery system at village level	5 (61.15)
			F9	Rate fluctuation in crops	1 (70.40)

After this, the main concern is to select an appropriate model for production and profit forecast. The behaviour, correlogram and the best model for forecasting of production and profit have been studied through ARIMA modeling

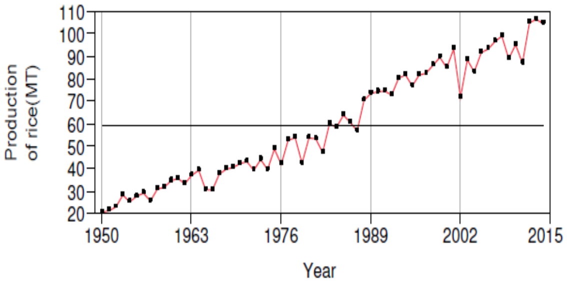
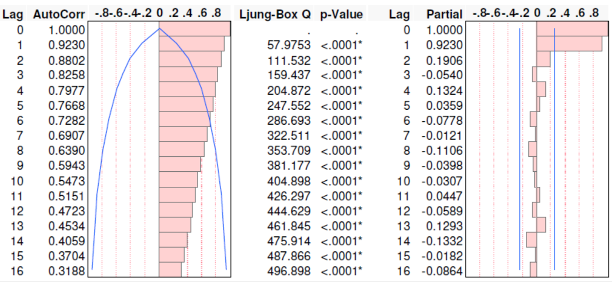


Figure 9: Behaviour of the annual production of rice in India



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.9230	57.9753	<.0001*	1	0.9230
2	0.8802	111.532	<.0001*	2	0.1906
3	0.8258	159.437	<.0001*	3	-0.0540
4	0.7977	204.872	<.0001*	4	0.1324
5	0.7668	247.552	<.0001*	5	0.0359
6	0.7282	286.693	<.0001*	6	-0.0778
7	0.6907	322.511	<.0001*	7	-0.0121
8	0.6390	353.709	<.0001*	8	-0.1106
9	0.5943	381.177	<.0001*	9	-0.0398
10	0.5473	404.898	<.0001*	10	-0.0307
11	0.5151	426.297	<.0001*	11	0.0447
12	0.4723	444.629	<.0001*	12	-0.0589
13	0.4534	461.845	<.0001*	13	0.1293
14	0.4059	475.914	<.0001*	14	-0.1332
15	0.3704	487.866	<.0001*	15	-0.0182
16	0.3188	496.898	<.0001*	16	-0.0864

Figure 10: Correlogram of annual production of rice in India

The line chart of rice production for India showed that the time is on x-axis and rice production on y axis as shown in Figure 9. Long-term increasing pattern of data indicates that it is non-stationary with mean (59.36), standard deviation (25.46) and trend augmented Dickey-Fuller (–6.05). Further, the augmented Dickey-Fuller showed the non-significant results *i.e.*, accept the null hypotheses of non-stationarity. The correlogram for yearly rice production data through ACF and PACF plots in which spikes are coming outside from the insignificant zone and fails to follow the assumption of randomness of the data as shown in Figure 10. So, forecasting rice production through ARIMA model has been estimated after transforming the variable. The logarithmic and differenced technique is utilized to make variable stationary. Firstly, the 1st order difference of the variable has been utilized which depicts that the data was still non-stationary. Thereafter, to achieve the stationarity, differencing of order 2 has been taken. After differencing, the mean and standard deviation were constant whereas, the trend ADF (–17.41).

Table 8: Different models for annual production of rice in India

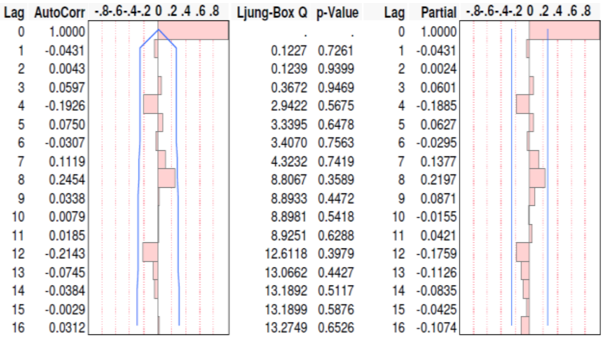
Model	**Intercept**	**Significance of parameters/model**	**AIC**	**SBC**	**R^2**
ARIMA(0 2 2)	Yes	Significant	–216.40	–209.97	0.95
ARIMA(1 2 2)	Yes	Non-significant	–214.44	–205.87	0.95
ARIMA(2 2 2)	Yes	Non-significant	–212.45	–201.73	0.95
ARIMA(2 2 1)	No	Significant	–210.82	–204.39	0.94
ARIMA(2 0 0)	Yes	Significant	–205.21	–198.69	0.89

Several ARIMA models were developed based on Box-Jenkins methodology. Among them the best five models have been proposed based on minimum AIC (Akaike Information Criterion), SBIC (Schwartz Bayesian Information Criterion) and the value of R^2 as shown in Table 8. The model ARIMA (0 2 2) had lowest AIC (–216.40) and SBIC (–209.97) values with the value of R^2 (0.95).

Table 9: Parameter estimates of ARIMA (0 2 2) for annual production of rice in India

Term	**Lag**	**Estimate**	**Std error**	***t* ratio**	***p*-value**	**MAPE**	**–2logliklehood**	**Constant estimate**
MA1	1	1.7833	0.1112	16.04	<.0001	1.8514	–222.3997	–0.0002
MA2	2	–0.7834	0.1053	–7.44	<.0001			
Intercept	0	–0.0002	0.0001	–2.27	0.0268			

The estimates of the parameter are shown in table 9 having MA (1) as 1.78, MA (2) as –0.78 and intercept as –0.0002 which were positively and negatively significant respectively. Further, ARIMA (0 2 2) model has also been selected because of minimum values of MAPE (1.85) and – 2 log likelihood (–222.39) which usually indicate a best fitted model according to the above three stages. The model verification is concerned with checking the residual of the model to see if they contain any systematic pattern which still can be removed to improve on the chosen ARIMA model.



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0431	0.1227	0.7261	1	-0.0431
2	0.0043	0.1239	0.9399	2	0.0024
3	0.0597	0.3672	0.9469	3	0.0601
4	-0.1926	2.9422	0.5675	4	-0.1885
5	0.0750	3.3395	0.6478	5	0.0627
6	-0.0307	3.4070	0.7563	6	-0.0295
7	0.1119	4.3232	0.7419	7	0.1377
8	0.2454	8.8067	0.3589	8	0.2197
9	0.0338	8.8933	0.4472	9	0.0871
10	0.0079	8.8981	0.5418	10	-0.0155
11	0.0185	8.9251	0.6288	11	0.0421
12	-0.2143	12.6118	0.3979	12	-0.1759
13	-0.0745	13.0662	0.4427	13	-0.1126
14	-0.0384	13.1892	0.5117	14	-0.0835
15	-0.0029	13.1899	0.5876	15	-0.0425
16	0.0312	13.2749	0.6526	16	-0.1074

Figure 11: ACF and PACF plots of ARIMA (0 2 2) for annual production of rice in India

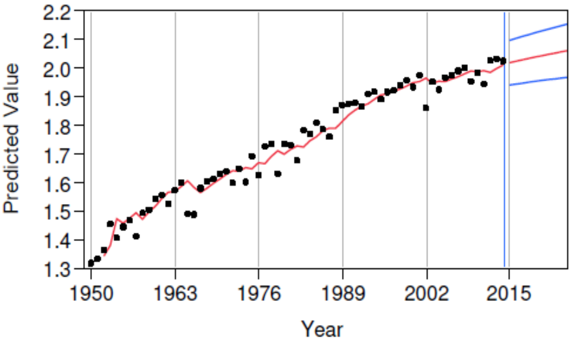


Figure 12: Forecasting graph of ARIMA (0 2 2) for annual production of rice in India

The ACF and PACF plots of the residual indicate 'good fit' of the model as shown in Figure 11. Moreover, *p*-values of Ljung-Box Q test are greater than the significance level (0.05), so we can conclude that the residuals are independent, and the model meets the assumption of randomness. ARIMA models are developed basically to forecast the corresponding variable. Figure 12 presents the graphical representation for forecasting of ARIMA (0 2 2) model which depicts that the trend is upward.

The line chart for profit of rice crop in India has been represented by Figure 13 and indicates that it is non-stationary with mean (19721.81), standard deviation (7548.71) and trend augmented Dickey-Fuller (-2.78).

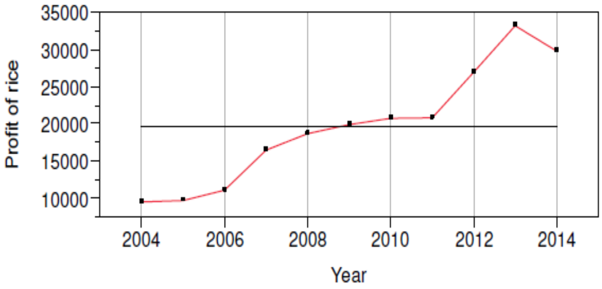
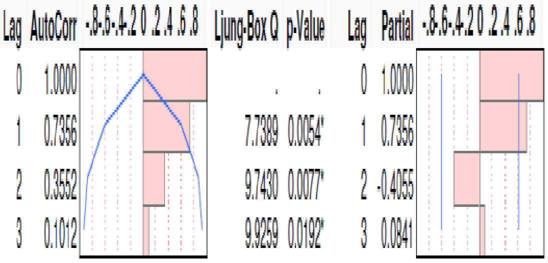


Figure 13: Behaviour of the annual profit of rice crop in India



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.7356	7.7389	0.0054*	1	0.7356
2	0.3552	9.7430	0.0077*	2	-0.4055
3	0.1012	9.9259	0.0192*	3	0.0841

Figure 14: Correlogram of annual profit of rice crop in India

The ADF showed the non-significant results *i.e.*, accept the null hypotheses of non-stationarity. The ACF and PACF plots depicts that the spikes are coming outside from the insignificant zone as shown in Figure 14 and fails to follow the assumption of randomness of the data. After differencing of order 1, stationarity has been achieved with mean (0.04), standard deviation (0.05) and trend ADF (-2.34). We may say that series is stationary at 1st order difference.

Table 10: Different models for annual profit of rice crop in India

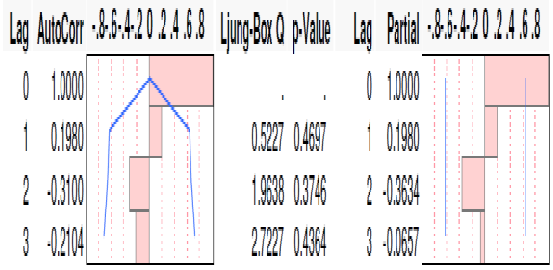
Model	Intercept	Significance of parameters/model	AIC	SBIC	R^2
ARIMA (0 1 1)	Yes	Significant	–25.6164	–25.0112	0.886
ARIMA (1 1 1)	Yes	Non-significant	–24.495	–23.5873	0.895
ARIMA(1 1 0)	Yes	Non-significant	–24.0353	–23.4301	0.866
ARIMA(1 0 1)	Yes	Non-significant	–21.1082	–19.9145	0.732
ARIMA(1 0 0)	Yes	Significant	–19.137	–18.3412	0.652
ARIMA(0 0 1)	Yes	Significant	–12.7685	–11.9727	0.532

Among the several models, the best five models have been proposed based on minimum AIC, SBIC and the value of R^2 as shown in Table10. The model ARIMA (0 1 1) had lowest AIC (–25.61) and SBIC (–25.01) values with the value of R^2 (0.88).

Table 11: Parameter estimates of ARIMA (0 1 1) for annual profit of rice crop in India

Term	Lag	Estimate	Std-error	*t* - ratio	*p*-value	MAPE	–2loglikelihood	Constant estimate
MA (1)	1	0.9999	0.3700	2.70	0.027	0.0460	–29.6163	0.0546
Intercept	0	0.0546	0.0046	11.74	<0.0001			

The estimates of the parameter MA (1) as 0.97 and intercept as 0.05 which were positively significant as shown in Table 11 and the ARIMA (0 1 1) model has also been selected because of minimum values of MAPE (0.04) and – 2 log likelihood (–29.61) which usually indicate a best fitted model.



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.1980	0.5227	0.4697	1	0.1980
2	-0.3100	1.9638	0.3746	2	-0.3634
3	-0.2104	2.7227	0.4364	3	-0.0657

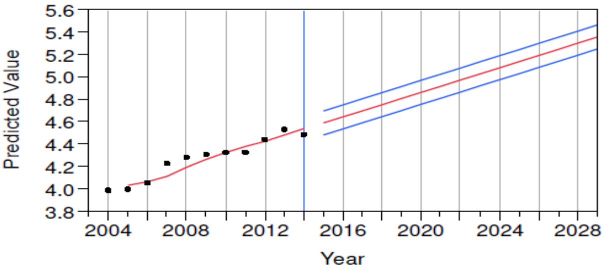


Figure 15: ACF and PACF plots of ARIMA (0 1 1) for the profit of rice crop in

Figure 16: Forecasting graph of ARIMA (0 1 1) for the profit of rice crop in India

The ACF and PACF plots of the residual and p-values of Ljung-Box Q test indicated that the residuals are independent, and the model meets the assumption of randomness (see Figure 15) whereas Figure 16 showed the graphical representation for forecasting of ARIMA (0 1 1) model.

Table 12: Prediction table of ARIMA (0 1 1) for the cost of cultivation and Profit of rice crop in India

Year	(Cost) Forecasted (Actual value) (Rs/hectare) [L-95, U-95]	(Profit) Forecasted (Actual) (Rs/hectare) [L-95, U-95]	Gain in %age (profit w.r.t cost)
2017-18	44814.14 (44810.32) [39799.24, 50460.94]	49805.55 (49800.26) [38931.56, 63716.74]	111.13
2018-19	49528.14 [43985.72, 55768.93]	56483.78 [44151.75, 72260.28]	114.04
2020-21	60495.90 [53726.14, 68118.68]	72646.71 [56785.85, 92937.67]	120.08
2022-23	73892.42 [65623.53, 83203.23]	93434.68 [73035.21, 119531.90]	126.45
2025-26	99749.51 [88587.09, 112318.40]	136284.5 [106529.70, 174350.10]	136.62

The gain in percentage for the profit of rice crop has been evaluated with respect to cost of cultivation as shown in Table 12. The gain in percentage for rice crop for the forecasted year 2025-26 will be 136.62 percent which represents that if farmer will spend Rs 100 for rice crop, then he will get a profit of 136.62 percent more than the actual investment.

4. Conclusions

1. In case of production, the variables area under rice (AUR), fertilizer consumption (FC), quality seed rate (QSR), electricity consumption (EC) and sale of power tillers (SPT) whereas in case of profit the variables cost of seed (CS), cost of fertilizer and manure (CFAM), cost of insecticide (COI) and irrigation charges (IC) will maximize the production and profit of rice at National level.
2. In case of production, the variable area under rice (AUR) whereas in case of profit the variable cost of seed (COS) will maximize the production and profit of rice for Jammu region.
3. The problem(s) faced by the farmers' during production of rice in the Jammu region was scarcity of labour during peak season followed by non-availability of power machines at appropriate time whereas in case of profit, the maximum problem faced by farmer was rate fluctuation followed by lack of storage facility.
4. The forecasted value of the production of rice based on estimated model for the year 2022 shall be 116.13 MT same as per the projection in annual report 2020-21(www.agricoop.nic.in). The forecasted value of the profit of rice on the basis of estimated model ARIMA (0 1 1) for the year 2025 shall be Rs. 136284.5 per hectare.

The gain in profit percentage with respect to cost has also been calculated for the year 2025 and shall be 136.62 percent which means government is still making the policies and taking initiatives to achieve the goal of doubling the farmers’ income at National level but the situation is different and poor in case of Jammu division of J&K UT. So, in order to enhance the farmers’ income in Jammu division as well, the studied variables FC, QSR, EC, SPT, CS, CFAM, COI and IC apart from the problems faced by farmers in case of production and profit may help the government to take appropriate initiatives and policies.

References

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, **66,** 237-242.

Benhin, J. K. A. (2008). South African crop farming and climate change: An economic assessment of impacts. *Global Environmental Change*, **18**(**4**), 666-678.

Box, G. E. P. and Jenkin, G. M. (1968). Discrete models for forecasting and control. *In Encyclopedia of Linguistics, Information, and Control*, pages 1-6, Oxford, Pergamon Press.

Box, G. E. P. and Jenkin, G. M. (1976). *Time Series of Analysis, Forecasting and Control*. San Franscico, Holden-Day, California, USA.

Cameron, A. C. and Trivedi, P. K. (2009). *Microeconometrics Using Stata*, StataCorp LP, Texas.

Cobb, C. W. and Douglas, P. H. (1928). A theory of production. *American Economic Review***,** **18**, 139-165.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association,* **19**(**1**), 15-18.

Galton, F. (1885). Regression towards Mediocrity in Heredity Stature. *Journal of the Anthropological Institute*, **15,** 246-263.

Garrett, H. E. and Woodworth, R. S. (1969). *Statistics in Psychology and Education*. Bombay: Vakils, Feffer & Simons Pvt. Ltd. P.329.

Koenker, R. and Bassett J. G. (1978). Regression quantiles. *Econometrica*, **46,** 33–50.

Marroquin, J. B. (2008). *Examination of North Dakota's Production, Cost, and Profit Functions: A Quantile Regression Approach*. Unpublished M.Sc. Thesis, North Dakota State University of Agriculture and Applied Science, Fargo, North Dakota.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality Quantity, **41**(**5**), 673-690.

Prajneshu (2008). Fitting of Cobb-Douglas production functions: Revisited. *Agricultural Economics Research Review*, **21**, 289-292.

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

Watson J., Watson G. S. (1950). Testing for serial correlation in least squares regression. *Biometrika*, **37**, 409–428.

https://www.jmp.com

http://eands.dacnet.nic.in

www.agricoop.nic.in

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference of SSCA, 23-27 February 2022; pp 107-113

An Application of Agrawal-Panda Type Post Stratified Estimates in Cluster Sampling with Application in Estimating Average Student Enrolment in a Class

Manish Trivedi[1] **and Purnendukisor Bandyopadhyay**[2]

[1]*School of Sciences, Indira Gandhi National Open University*

[2] *Department of Higher Education, Ministry of Education*

Received: 20 February 2022; Revised: 23 July 2022; Accepted: 23 July 2022

Abstract

The school education system in India cater to the students. Estimating the number of students in a typical class is essential for all types of policy planning, which target the students as recipient of a service. Due to a variety of reasons, it might not be possible to undertake a census of all the schools to have complete information about number of students. Therefore, resorting to sample surveys of schools in target areas can be a viable solution before launching of any programme. Due to paucity of data or incomplete frame information, it may not always be possible to stratify the schools in target areas before the survey. However, once the sample is selected, this information can be collected from the selected schools and post-stratified estimates can be developed from them. In this paper, an attempt has been made to develop an estimate using the Agrawal-Panda type methods of post stratified estimates of population mean on cluster sample using data from the student enrolments in 3 districts of one State for which recent data is available and it helps to derive the population parameters, including the variance and MSE of the proposed estimators.

Key words: Cluster sampling; Post stratification; Mean square error.

1. Introduction

Using stratified sampling for improving estimates of population parameters, where heterogeneity of population parameters is already known or anticipated, is a time-tested method. However, many-a-times, at the time of sample selection, the frame does not have enough data to divide the frame in appropriate strata. Sometimes, due to the nature of the variables under study, it becomes very difficult to use stratification before drawing of the samples, particularly, when the variables under study might be dependent on gender differentials, where the study variables might require stratification within nearly each household to be covered in the survey.

Post stratified sampling strategies are used in sample surveys in similar cases. Methodological developments on post-stratification in case of incomplete frame information has been studied since long (Holt and Smith,1979). This was followed by examination of the feasibility of ratio estimators in a post stratified setup (Jager *et. al.*, 1985). Critical thoughts on application of post stratification was presented subsequently (Smith, 1991). In a seminal paper on post stratification, weight structure which has made use of both the population information and the sampling fractions from different strata to arrive at the final estimate was examined

Corresponding Author: Purnendukisor Bandyopadhyay
Email: purnendukb@yahoo.com

(Agrawal and Panda, 1993, 1995). This has resulted in further study in the area of post stratification. Estimation methodologies for population mean under stratified population using prior information with grouping strategy was developed (Shukla *et. al.* 2001, 2002). This paper has essentially been motivated by the weight structure given in the Agrawal-Panda post stratified estimate. The structure has been first applied in a stratified sampling set up. A few alternate estimates using this structure has been proposed. Expectation and variance/Mean Square Error (MSE) of these alternate estimates have then been derived. After deriving these estimates, the variance / MSE of the estimates have been compared with that of the usual post stratified estimates using real life data.

Concluding remarks have been provided on performance of the proposed estimators and scopes for further work in this area.

2. Methodology

Let *U* be a finite population having *N* clusters. Let the clusters be of unequal sizes. The population can be divided into *k* strata such that *i*th stratum contains N_i clusters, $\sum_{i=1}^{k} N_i = N$. A random sample of *n* clusters (*n* < *N*) is drawn from the N clusters by simple random sampling without replacement (SRSWOR). The sample is post stratified such that n_i clusters are from the *i*th stratum $\sum_{i=1}^{k} n_i = n$.

2.1. Notations used for the population parameters

The notations used throughout are

Y Variable under study

Y_{ijl} *l*th value of the variable *Y* in *i*th stratum and *j*th cluster

$\bar{Y}_{ij}$ Mean value of Y in *i*th stratum and *j*th cluster

W_i Population proportion of clusters in *i*th stratum = $\frac{N_i}{N}$

M_{ij} Size of *j*th cluster in *i*th stratum

M_i Total elements in *i*th stratum = $\sum_{j=1}^{N_i} M_{ij}$

$\bar{M}_i$ Average size of clusters in *i*th stratum = $\frac{M_i}{N_i}$

u_{ij} ${M_{ij}}\big/{M_i}$

$\bar{Y}_i$ Population mean of *i*th stratum = $\sum_{i-1}^{k} \sum_{j=1}^{N_i} {Y_{ijl}}\big/{M_i} = \bar{Y}_{N_i} = {\sum_{j=1}^{N_i} M_{ij}\, \bar{Y}_{ij}}\big/{M_i} =$ $\frac{1}{N_i} \sum_{j=1}^{N_i} u_{ij} \bar{Y}_{ij}$

$\bar{Y}_{..}$ $= \bar{Y} = \sum_{i=1}^{k} W_i \bar{Y}_i$

$\bar{\bar{Y}}_N$ = mean of *N* cluster means $= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_{ij}} \bar{Y}_{ij}$

$\bar{\bar{Y}}_{N_i}$ Mean of $N_i clusters\ of\ i^{th} stratum = \frac{1}{N_i} \sum_{j=1}^{N_i} \bar{Y}_{ij}$

$S_{b_i}^2 = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$

$S_{b_i}'^2 = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left(u_{ij}\, \bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$

$$S'_{bu_i} = \sqrt{\sum_{J=1}^{N_I} \frac{\left(u_{ij} - 1\right)^2}{(N_i - 1)}}$$

$$S'_{byu_i} = \sum_{J=1}^{N_I} \frac{\left(u_{ij}\,\bar{Y}_{ij} - \bar{Y}_i\right)\ \left(u_{ij} - 1\right)}{(N_i - 1)}$$

$$S_{iM\bar{Y}} = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} \left(M_{ij} - \bar{M}_i\right) \left(\bar{Y}_{ij} - \bar{\bar{Y}}_{N_i}\right)$$

$$S''^{\,2}_{b_i} = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} u^2{}_{ij}\left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$$

$$S'''_{b_i} = \frac{1}{(N_i - 1)} \sum_{J=1}^{N_i} \left[u_{ij}\left(\bar{Y}_{ij} - \bar{Y}_i\right) - \bar{\bar{Y}}_N\left(u_{ij} - 1\right)\right]$$

2.2. Notations used for the sample estimates

The notations used for the sample estimates are

$\bar{y}_{ij}$ Mean of *j*th cluster in *i*th stratum included in the sample, *i* = 1, 2, …, *k*; *j* = 1, 2, …, n_i

$\bar{y}_i = \dfrac{\sum_{j=1}^{n_i} M_{ij}\bar{y}_{ij}}{\sum_{j=1}^{n_i} M_{ij}}$, sample mean of *i*th stratum

$P_i = \dfrac{n_i}{n}$: sample proportion of clusters from *i*th stratum

$f = \dfrac{n}{N} : sampling\ fraction$

$$\bar{u}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} u_{ij}$$

2.3. Usual estimators

Some of the usual estimators in cluster sampling scheme are:

$$\bar{\bar{y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \bar{y}_{ij}$$

$$\bar{y}'_i = \frac{1}{n_i}\sum_{j=1}^{n_i} u_{ij}\,\bar{y}_{ij}$$

$$\bar{y}''_i = {\sum_{j=1}^{n_i} u_{ij}\,\bar{y}_{ij}} \Big/ {\sum_{j=1}^{n_i} u_{ij}}$$

$\bar{y}_{ps} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{y}_i$ is the usual post stratified estimator

Variance in stratified sampling is $Var\left(\bar{y}_{stratified}\right) = \sum_{i=1}^{k}\left(\frac{N_i}{N}\right)^2 \left(\frac{N_i - n_i}{N_i}\right) \frac{S_i^2}{n_i}$

The variance of usual post stratified estimator is

$$Var\left(\bar{y}_{ps}\right) = Var\left(\bar{y}_{stratified}\right) + \frac{N - n}{n^2\,N}\sum_{i=1}^{k}(1 - W_i)\,S_i^2$$

3. Proposed Estimators

For developing the proposed estimates, first let us recall the weight structure used in the seminal paper of Agrawal-Panda (1993), which is $W_{i\,\alpha}^{*} = \left[\alpha\ \frac{n_i}{n} + (1-\ \alpha)\ \frac{N_i}{N}\right]$, α being a suitably chosen constant.

Taking inspiration from this weight structure, some proposed post stratified estimates, which we have examined in this paper are

$$\bar{y}_a^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\, \bar{\bar{y}}_i$$

$$\bar{y}_b^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\, \bar{y}_i'$$

$$\bar{y}_c^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\, \bar{y}_i''$$

$$\bar{y}_d^{*} = \sum_{i=1}^{k} W_{i\,\alpha}^{*}\, \bar{u}_i\, \bar{\bar{y}}_i$$

3.1. Deriving properties of the proposed estimators

Theorem 1: The estimator $\bar{y}_a^{*}$ is biased for $\bar{Y}$ and the MSE is

$$MSE\ (\bar{y}_a^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right) \textstyle\sum_{i=1}^{k}\ W_i S_{b_i}^2 +$$
$$\frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2} \textstyle\sum_{i=1}^{k}(1-\ W_i)\, S_{b_i}^2 + \frac{(N-n)}{Nn}\alpha^2 \left[S_a^2\ -\ \sum_{i=1}^{k}\ W_i S_{b_i}^2\ \right] + \left[\sum_{i=1}^{k} \frac{(N_i-1)\, W_i}{N_i \bar{M}_i}\quad S_{iM\bar{Y}}\right]$$

where $S_a^2 = \frac{1}{(N-1)}\ \sum_{i=1}^{k} \sum_{j-1}^{N_i}\left[\bar{Y}_{ij}\ -\ \bar{\bar{Y}}_N\ \right]^2$

Theorem 2: The estimator $\bar{y}_b^{*}$ is unbiased for $\bar{Y}$ and the variance is

$$Var\ (\bar{y}_b^{*}) = \left(\frac{1}{n}\ -\ \frac{1}{N}\right) \textstyle\sum_{i=1}^{k}\ W_i S_{b_i}'^{\,2} +$$
$$\frac{(N\ -n)\,(1\ -\ \alpha)^2}{(N\ -1)\,n^2} \sum_{i=1}^{k}(1-\ W_i)\, S_{b_i}'^{\,2} + \frac{(N\ -n)}{Nn}\alpha^2 \left[S_b^2\ -\ \sum_{i=1}^{k}\ W_i S_{b_i}'^2\ \right]$$

where $S_{b_i}'^{\,2} = \frac{1}{(N_i-1)}\ \sum_{J=1}^{N_i}\left(u_{ij}\ \bar{Y}_{ij}\ -\ \bar{Y}_{N_i}\right)^{\,2}$ and $S_b^2 =\ \sum_{i=1}^{k} \frac{1}{(N_i-1)} \sum_{j-1}^{N_i}\left[u_{ij}\bar{Y}_i\ -\ \bar{Y}\ \right]^2$

Theorem 3: The estimator $\bar{y}_c^*$ is unbiased for $\bar{Y}$ and the variance is

$$Var\ (\bar{y}_c^*) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_{b_i}^{\prime\prime 2} +$$

$$\frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2} \sum_{i=1}^{k} (1- W_i)\, S_{b_i}^{\prime\prime 2} + \frac{(N-n)}{Nn}\alpha^2 \left[S_c^2 - \sum_{i=1}^{k} W_i S_{b_i}^{\prime\prime 2}\right]$$

where $S_{b_i}^{\prime\prime 2} = \frac{1}{(N_i-1)} \sum_{J=1}^{N_i} u^2{}_{ij} \left(\bar{Y}_{ij} - \bar{Y}_{N_i}\right)^2$ and $S_c^2 = \frac{1}{(N-1)} \sum_{i=1}^{k} \sum_{j-1}^{N_i} u_{ij}^2 \left[\bar{Y}_{ij} - \bar{Y}\right]^2$

Theorem 4: The estimator $\bar{y}_d^*$ is biased for $\bar{Y}$ and the MSE is

$$MSE\ (\bar{y}_d^*) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_{b_i}^2 +$$

$$\frac{(N-n)\,(1-\alpha)^2}{(N-1)\,n^2} \sum_{i=1}^{k} (1- W_i)\, S_{b_i}^2 + \frac{(N-n)}{Nn}\alpha^2 \left[S_a^2 - \sum_{i=1}^{k} W_i S_{b_i}^2\right] +$$

$$\left[\sum_{i=1}^{k} \left\{\left(W_i - \frac{1}{n}\right) - \frac{(1-\alpha)\,(N-n)\,(1-W_i)}{(N-1)\,n^2\,W_i}\right\} \frac{S_{iM\bar{Y}}}{\bar{M}_i}\right]^2$$

4. Numerical Illustrations

The above proposed estimators were compared with usual post stratified estimators using 3 different values of α, so that it can show the performance of the proposed estimators for varying levels of weights distributed between the sample and population proportion of clusters.

The Department of School Education and Literacy (DoSEL), Government of India conducts the annual survey of schools. This data collection and dissemination system is called the Unified District Information System for Education Plus (UDISE+), the details of which can be accessed at https://udiseplus.gov.in/#/page/about. This is an annual census conducted in more than 15 lakh schools of India. The DoSEL disseminates anonymised unit level data. The numerical illustration has been done using class-wise enrolment of students in different schools at 3 districts of one State. Here, each district was considered as a separate stratum, each school as a cluster and variable under study was number of students in a class. The data set provides information on all the schools of a district, i.e., the entire frame was available for drawal of samples and computation of variance, MSE, etc. of the proposed estimators. In the present paper, samples were drawn by SRSWOR from the frame of schools of a specific State and then post-stratified among the 3 strata (i.e., 3 districts of this State). The results are given below:

Variable	Stratum			overall
	I	II	III	
N_i	176	184	58	418
n_i	23	20	7	50
$\bar{Y}$				24.2
$\bar{\bar{Y}}_N$				20.7
$\bar{y}_i$	29.47	14.51	21.14	
$\sum_{j=1}^{n_i} u_{ij}$	22.79	19.97	6.06	

Variable	Stratum			overall
	I	II	III	
$\bar{y}_{ps}$				21.73
$\bar{\bar{y}}_i$	22.78	13.30	17.61	
$\bar{y}_i'$	29.20	14.49	18.30	
$\bar{y}_i''$	29.47	14.51	21.14	
$\bar{y}_a^*$, α = 0.1				17.93
$\bar{y}_a^*$, α = 0.5				18.08
$\bar{y}_a^*$, α = 0.9				18.23
$\bar{y}_b^*$, α = 0.1				21.27
$\bar{y}_b^*$, α = 0.5				21.50
$\bar{y}_b^*$, α = 0.9				21.74
$\bar{y}_c^*$, α = 0.1				21.79
$\bar{y}_c^*$, α = 0.5				22.03
$\bar{y}_c^*$, α = 0.9				22.26
$\bar{y}_d^*$, α = 0.1				17.50
$\bar{y}_d^*$, α = 0.5				17.65
$\bar{y}_d^*$, α = 0.9				17.79
$Var\left(\bar{y}_{stratified}\right)$				16.88
$Var\left(\bar{y}_{ps}\right)$				**19.47**
$MSE\ (\bar{y}_a^*)$, α = 0.1				**16.346**
$MSE\ (\bar{y}_a^*)$, α = 0.5				**16.361**
$MSE\ (\bar{y}_a^*)$, α = 0.9				**16.814**
$Var\ (\bar{y}_b^*)$, α = 0.1				23.229
$Var\ (\bar{y}_b^*)$, α = 0.5				19.710
$Var\ (\bar{y}_b^*)$, α = 0.9				**12.234**
$Var\ (\bar{y}_c^*)$, α = 0.1				**16.942**
$Var\ (\bar{y}_c^*)$, α = 0.5				**17.325**
$Var\ (\bar{y}_c^*)$, α = 0.9				**18.756**
$MSE\ (\bar{y}_d^*)$, α = 0.1				**13.74**
$MSE\ (\bar{y}_d^*)$, α = 0.5				**13.76**
$MSE\ (\bar{y}_d^*)$, α = 0.9				**14.21**

As can be seen from the above results, the variance / MSE of the proposed estimators were less than the usual post stratified estimator $\bar{y}_{ps}$. Moreover, for a few specific choices of α, the variance/ MSE of proposed estimators were less than the variance of usual stratified SRSWOR estimator. In the given data set, the proposed estimator $\bar{y}_b^*$ has shown least variance with a choice of α = 0.9.

5. Concluding Remarks

The four estimates namely $\bar{y}_a^*, \bar{y}_b^*, \bar{y}_c^*$ *and* $\bar{y}_d^*$ have been developed using the weight structure of Agrawal and Panda (1993) after applying this on a post stratified cluster sampling scheme. The properties of these estimators have been derived and comparison among these estimators' variances/ Mean Square Errors have been made to find out whether the proposed estimator is behaving well or not. The methodology has been applied on a data set which is newly available, and which contain population level information. This has enabled to compute the theoretical variances and MSEs from real life recent data. The variances/ MSEs of the proposed post stratified estimators are lower than the traditional post stratified estimators for some of the values of α. This shows that the proposed post stratified estimator can be more efficient over the usual PS estimator for these values of α.

Acknowledgement

The authors are grateful to the editors for their useful suggestions that led to considerable improvement in the presentation of the contents. The authors are grateful, particularly to Prof. V. K. Gupta, who gave numerous suggestions for improvement of the paper and Dr. D. Roy Choudhury who was a constant source of inspiration for the authors. The authors humbly and respectfully acknowledge the constant guidance of Prof. V. K. Singh and the encouragement and support from the faculty of the School of Sciences, IGNOU. Last, but not the least, the authors are grateful to the Ministry of Education for their encouragement in taking up research work in addition to usual duties.

References

Agarwal, M. C. and Panda, K. B. (1993). An efficient estimator in poststratification. *Metron*, **5**(**3-4**), 179-187.

Bryant, E. C., Hartley, H. O. and Jessen, R. J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, **55**, 105-124.

Holt, D. and Smith, T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical* Society, **A, 142**, 33-36.

Shukla, D. and Trivedi, M. (2001). Mean estimation in deeply stratified population under post-stratification. *Journal of* the *Indian Society of Agricultural Statistics,* **54(2)**, 221-235.

Smith, T. M. F. (1991). Post-stratification. *The Statistician,* **40**, 315-323.

Unified District Information System for Education Plus – data sharing portal https://src.udiseplus.gov.in/udise-share/ accessed various times during 2021.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 115-127

Time Series Analysis of Major Cotton Production States in India using Box−Jenkins Approach

Prema Borkar
Gokhale Institute of Politics and Economics（Deemed University, Pune, India

Received: 19 July 2022; Revised: 25 July 2022; Accepted: 28 July 2022

Abstract

Cotton is one of the most important fibers and cash crops of India and plays a dominant role in the industrial and agricultural economy of the country. It provides the basic raw material (cotton fiber) to the cotton textile industry. In India, there are ten major cotton – growing states which are divided into three zones, viz., north zone, central zone, and south zone. The north zone consists of Punjab, Haryana, and Rajasthan. The central zone includes Madhya Pradesh, Maharashtra, and Gujarat. The South zone comprises Andhra Pradesh, Telangana, Karnataka, and Tamil Nadu.

In this study, the data on cotton production in major cotton-producing states in India were collected from the website of Cotton Corporation of India for the period from 1964-65 to 2021-22 and were used to fit the ARIMA model and to predict future production. The Box Jenkins (1970) ARIMA methodology has been used for forecasting. ARIMA forecasting model is the most popular and widely used forecasting model for time series data. Autocorrelations and partial autocorrelation functions were calculated for the data. Model parameters were estimated using R programming software. The performance of the fitted model was examined by computing various measures of goodness of fit *viz.*, AIC, BIC, and MAPE. Empirical results showed that ARIMA (0,1,0) model was most suitable to forecast the future production of cotton in India. Similarly, the ARIMA model was fitted separately for major cotton-producing states in India. The forecasts from 2022-23 to 2029-30 are calculated based on the selected model. Overall cotton production is expected to be 362.18 million tons by 2029-30. The forecasting power of the autoregressive integrated moving average model was used to forecast cotton production for eight leading years. The results of major cotton-growing states are presented numerically and graphically.

Key words: ACF - autocorrelation function; ARIMA - autoregressive integrated moving average; Cotton production; PACF - partial autocorrelation function; Residual analysis.

1. Introduction

Cotton is often referred to as the "birth place of India" and the industry plays an important role in the country's economy. It is important in both developed and developing countries as a cash crop for millions of farmers, including small and marginal farmers, and as a strategic raw material for the textile industry. Although cotton is farmed in almost 100 nations, only six countries—China, India, the United States, Brazil, Pakistan, and Uzbekistan contribute to nearly 80% of global production (19.84 million tons) (FAO, 2022). With a yearly production

Corresponding Author: Prema Borkar
Email: premaborkar@rediffmail.com

volume of 6.16 million tonnes, India is the world's second largest cotton producer (FAO, 2022). However, for a variety of well-known reasons, there are variances between nations in terms of the fundamental crop/commodity performance metrics, such as area, production, productivity, trade, etc. Cotton production has increased significantly since the introduction of genetically modified crops, especially in the US, China, India, Australia, Argentina, and South Africa.

The cotton sector in India directly supports approximately 5 million farmers spread across 10 states, and it plays an important role in the domestic economy as a strategic industrial raw material for the textile industry. India ranks first in world cotton area cultivation which is about 37 % of the world area under cotton cultivation between 12.0 million hectares to 13.5 million hectares (www.cotcorp.org.in). It is the second largest producer of cotton in the world accounting for about 22% of the world cotton production. Despite the fact that India has the most cotton land, its productivity is among the lowest in the world. Among the main factors cited for India's low cotton productivity are the prevalence of small and marginal holdings, insufficient transfer of production technologies, and insufficient financial resources. The yield per kgs hectare which is presently 469 kgs/ha is still lower against the world average yield of about 787 kgs /ha (www.cotcorp.org.in).

Millions of farmers and those working in industries related to cotton, such as transportation and processing, are employed by the crop. In terms of acres planted to cotton and cotton production, India leads the globe. Currently, ten major cotton-growing states account for the majority of the nation's cotton production. These states can be divided into three regions: The Northern Zone, which includes Punjab, Haryana, and Rajasthan; the Central Zone, which includes Gujarat, Maharashtra and Madhya Pradesh; and the Southern Zone, which includes Andhra Pradesh, Telangana, Karnataka, and Tamil Nadu.

Cotton cultivation, marketing, processing, and exports provide a living for nearly 60 million people today (www.ibef.org). India is also the only country in the world that commercially grows not only the four cultivated cotton species, but also their intra- and inter-specific hybrids. The textile industry, which uses cotton as its primary raw material, contributes about 4% of GDP and is the country's largest foreign exchange earner. As a result, the growth and development of the cotton and cotton-based textile industries is critical to the overall development of the Indian economy. Thus, it becomes important to study the cotton production in major states of India and to forecast cotton production in India. The main objective of this study is to develop an ARIMA model for forecasting the cotton production in major states of India using Box-Jenkins approach.

2. Material and Methods

In this study, the data on cotton production in major cotton−producing states in India were collected from the website of Cotton Corporation of India for the period from 1964-65 to 2021-22 and were used to fit the ARIMA model and to predict future production using Box-Jenkins approach.

2.1. Autoregressive integrated moving average

ARIMA stands for auto-regressive integrated moving average and is defined by three order parameters: (*p*, *d*, *q*). The Box-Jenkins technique is another name for the procedure of fitting an ARIMA model. ARIMA model is a technique for prediction the future values or events of the variable. This method is suitable for any time series with any pattern of change. It requires a long time series data for analysis (Biswas *et al.*, 2014). When past values are used in the regression equation for the series Y, this is referred to as an auto regressive AR(*p*) component. The auto-regressive parameter *p* provides the model's number of lags. The general model AR(*p*) is represented as:

$$\boldsymbol{Y_t} = \boldsymbol{\mu} + \sum_{i=1}^{p} \phi_i \boldsymbol{Y_{t-i}} - \boldsymbol{i} + \boldsymbol{e_t}$$

where, *p* is the order of the AR model and $\phi_1, \phi_2, \ldots, \phi_p$ are *p* partial autocorrelation parameters for the AR(*p*) model. The AR(*p*) model contains only *p* statistically significant partial autocorrelations. The AR(*p*) model's autocorrelation coefficient approaches zero and is constrained between –1 and 1. The degree of differentiation in the integrated I(*d*) component is represented by *d*. To differentiate a series, just subtract its current and prior values *d* times. When the stationarity assumption is violated, differencing is frequently utilized to stabilize the series. Hossain *et. al.* (2006) forecasted three different varieties of pulse price in Bangladesh using ARIMA model. Mandal (2005) forecasted sugarcane production in India. Assis *et al.* (2010) forecasted cocoa bean prices in Malaysia along with other competing models. Cooray (2006) forecasted Sri Lanka's monthly total production of tea and paddy monthly data from January 1988 to September 2004.

2.2. Moving average

The error of the model is represented as a linear combination of past error terms by a moving average MA(*q*) component. The number of terms to include in the model is determined by the order.

$$\boldsymbol{Y_t} = \mu + \boldsymbol{\theta_1 e_{t-1}} + \boldsymbol{\theta_2 e_{t-2}} + \cdots + \boldsymbol{\theta_q e_{t-q}} + \mathbf{e_t}$$

The moving average is a linear combination of past forecast errors. The general MA(q) model is represented as:

$$\boldsymbol{Y_t} = \mu + \sum_{i=1}^{q} \Theta_t \, \boldsymbol{e_{t-i}} + \boldsymbol{e_t}$$

where, *q* is the order of the model and $\Theta_1, \Theta_2, \ldots\ldots, \Theta_q$ are parameters of the model. The $e_t, e_{t-1}, \ldots\ldots, e_{t-q}$ are the white noise error terms. This can be equivalently written in terms of the backshift operator B as:

$$\boldsymbol{Y_t} = \mu + \left(\sum_{i=1}^{q} \Theta_i B^i + \mathbf{1} \right) \boldsymbol{e_t}$$

The capability of inverting an MA model to obtain an AR model with infinite order is made possible by the invertibility condition of an MA(*q*) process. A non-seasonal ARIMA model is made up of varying, autoregressive, and moving average components and can be expressed as a linear equation

$$\boldsymbol{Y_t} = \mu + \sum_{i=1}^{p} \phi_i \boldsymbol{Y_{di}} + \sum_{i=1}^{q} \Theta_t \, \boldsymbol{e_{t-i}} + \boldsymbol{e_t}$$

where, y_d is *Y* differenced *d* times and μ is a constant mean.

The Box-Jenkins methodology for estimating a time series model consists of four iterative steps: model identification, estimation of model parameters, diagnostic checking, and forecasting. The tentative model parameters are identified first using ACF and PACF, and then the coefficients of the most likely model are determined. The next steps involve forecasting, validating, and checking the model performance by observing the residuals using the Ljung Box test and ACF plot of residuals.

3. Model Identification

Theoretically, ARIMA models are the most general class of models for forecasting a time series that may be made "stationary" by differencing (if necessary), sometimes in conjunction with nonlinear transformations such as logging or deflating (if necessary). A stationary random variable is one whose statistical features remain consistent across time. A stationary series has no trend, constant amplitude variations around its mean, and wiggles in a consistent manner, i.e., its short-term random temporal patterns always look the same statistically. The ARIMA forecasting equation for a stationary time series is a linear (regression-type) equation in which the predictors are dependent variable lags and/or forecast error lags.

To find the best ARIMA model for *Y*, first determine the order of differencing (*d*) required to stationeries the series and eliminate the gross seasonal features, maybe in conjunction with a variance stabilizing operation like logging or deflating. If you stop here and forecast that the differenced series is constant, you've just fitted a random walk or random trend model. However, the stationeries series may still have auto correlated errors, implying that several AR terms (*p* 1) and/or few of MA terms (*q* 1) are required in the forecasting equation. Statistical tests are used to determine if a time series is stationary. To determine if the time series were stationary, Augmented Dickey Fuller (ADF) were used in this study.

3.1. Estimating the parameters

After tentatively identifying the suitable model, next step is to obtain least square estimates of the parameters, such as R^2, Root mean square error (RMSE), Mean absolute percentage error (MAPE), Mean absolute error (MAE) and normalized Bayesian Information Criterion (BIC) to check the accuracy of the model. In this study, three different parameters are considered for the evaluation of the forecasting models *i.e.*, MAPE, AIC and BIC.

The Mean Absolute Percent Error (MAPE)

The mean absolute percent error was used as a measure of accuracy of the models. It is also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where, A_t is the actual value and F_t is the forecast value. Their difference is divided by the actual value $A_{t.}$

Low Akaike information criteria (AIC)

AIC is estimated by AIC = $(-2\log L + 2m)$,

where, $m = p + q$ and L is the likelihood function.

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Low Bayesian Information Criteria (BIC)

Bayesian Information Criteria (BIC) is also used and estimated by

$$\text{BIC} = \log \sigma 2 + (m\log n)/n.$$

The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

3.2. Diagnostic checking

After estimating the parameters of a tentatively recognized ARIMA model, diagnostic testing is required to ensure that the model is adequate. Examining the ACF and PACF of residuals may reveal the model's adequacy or insufficiency. If it has random residuals, it means that the model that was tentatively selected is adequate. When an inadequacy is found, the checks should indicate how the model should be updated, followed by more fitting and checking. When all of their ACF were under the limitations, the residuals of ACF and PACF were considered random (Burark and Sharma, 2012).

3.3. Forecasting

Future values of the time series are forecasted. R programming software was used for time series analysis and developing ARIMA models.

4. Results and Discussion

The time series data of production in major cotton growing states were subjected to a stationary check, which demonstrated the non-stationarity of cotton production except Punjab which was found to be stationary. Nonstationary time series data were made stationary using first order differencing and best fit ARIMA models were developed and used to forecast production of cotton in major cotton growing states in India during 2022-23 to 2029-30. The initial values for the orders of the non-seasonal parameters "p" and "q" were used to identify ARIMA models. They were discovered by plotting autocorrelation and partial autocorrelation functions for significant spikes. During the identification stage, one or more models that appear to provide statistically adequate representations of the available data were tentatively chosen.

Initially, all cotton growing state production ACF and PACF plots were plotted and then it was examined. In the plots, the continuous line above and below the x-axis represents the confidence limits. It was observed that the spikes were above the confidence limits. It means the data was not stationary. No spikes exceeded confidence limits after differencing. This indicated that the series had reached its stationary point. The plot of ACF and PACF with differencing is shown in Figure 1 except Punjab.

The order of *p* and *q* were determined based on the ACF and PACF plots for developing the preliminary ARIMA model for major cotton producing states in India. Ten tentative ARIMA models were chosen with different *p*, *d*, and *q* values that were within a reasonable range. The model's parameters were then precisely estimated using least squares. After fitting the model, accuracy of the model was tested based on diagnostics statistics *i.e.*, MAPE, AIC and BIC. The model which had lowest value of these parameters was selected for validation. The fitted ARIMA model is presented in Table 1.

Table 1: Model fit statistics of the fitted ARIMA model

State	Best Fitted ARIMA Model	MAPE	AIC	BIC
Maharashtra	(2,1,2)	8.37	390.36	402.62
Andhra Pradesh and Telangana	(0,1,1)	7.94	397.28	403.41
Gujarat	(0,1,0)	8.53	434.10	436.14
Rajasthan	(0,1,0)	8.34	271.65	273.69
Karnataka	(0,1,2)	9.54	306.04	312.17
Haryana	(1,1,1)	7.32	281.63	287.76
Madhya Pradesh	(2,1,2)	7.34	276.64	286.86
Punjab	(1,0,0)	8.36	309.26	35.44
Tamil Nadu	(2,1,1)	7.45	168.74	176.91
India	(0,1,0)	8.41	509.36	511.40

The auto-correlation function (ACF) and partial auto-correlation function (PACF) of residuals were further examined to see if the selected models contained any systemic pattern that could be removed to improve predictability. The ACF and PACF of these models' residuals for major cotton producing states were plotted. This figure shows that the ACF and PACF of residuals are within the confidence interval and are not significantly different from zero. This indicated that the models were chosen correctly.

To see if the forecast errors are normally distributed with mean zero, plot a histogram of the forecast errors with an overlaid normal curve with mean zero and the same standard deviation as the forecast error distribution. Figure 2 histogram plots demonstrates the histogram of forecast errors of residuals of major cotton production states in India.

The forecast error time plots of major cotton producing states demonstrates that the variation of the forecast errors is almost consistent across time. The forecast error histogram indicates that it is likely that the errors are normally distributed, with mean zero and variance constant. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance. Since successive forecast errors do not appear to be connected and the forecast errors appear to be normally distributed with mean zero and constant variance, the ARIMA model (2,1,2), (0,1,1), (0,1,0), (0,1,0), (0,1,2), (1,1,1), (2,1,2), (1,0,0), (2,1,1) and (0,1,0) appears to be appropriate for predicting Maharashtra, Andhra Pradesh and Telangana, Gujarat, Rajasthan, Karnataka, Haryana, Madhya Pradesh, Punjab, Tamil Nadu and India.

Figure 3 shows the actual and forecasted plots of major cotton production states in India. Eight years ahead forecast was done for major cotton production states in India using the fitted ARIMA models *i.e.*, ARIMA (2,1,2), (0,1,1), (0,1,0), (0,1,0), (0,1,2), (1,1,1), (2,1,2), (1,0,0), (2,1,2) and (0,1,0) at the 95 per cent confidence interval.

From Figure 3, it is observed that Maharashtra, Karnataka, Haryana and Tamil Nadu is showing a slightly increasing trend. The forecasted values for cotton production in Maharashtra, Karnataka, Haryana and Tamil Nadu in 2029-30 was found to be 95.49, 18.62, 20.88 and 4.31 million tonnes, respectively. While the production of cotton in Andhra Pradesh, Gujarat and Rajasthan will remain constant throughout the study period. The production of cotton in Madhya Pradesh is showing a declining trend. The overall cotton production in India will remain constant throughout the study period i.e., 362.18 million tonnes. The fitted models accurately forecast 91.63 percent for Maharashtra, 92.06 percent for Andhra Pradesh, 91.47 percent for Gujarat, 91.66 percent for Rajasthan, 90.46 percent for Karnataka, 92.68 percent for Haryana, 92.66 percent for Madhya Pradesh, 91.64 percent for Punjab and 92.55 percent for Tamil Nadu, according to the mean absolute percentage error (MAPE).

5. Conclusion

The Box-Jenkins method as an ARIMA model was used to forecast future values based on the historical movement patterns of a variable. In this paper, a model for forecasting of cotton production in major states of India was developed. The forecasted cotton production in Maharashtra, Karnataka, Haryana and Tamil Nadu exhibited a slightly increasing trend and Madhya Pradesh exhibited a declining trend. Whereas, Andhra Pradesh, Gujarat, Rajasthan exhibited no trend *i.e.*, it will remain constant throughout the study period. Based on the forecasting and validation results, it is possible to conclude that the ARIMA model might be used successfully forecast cotton production in major states of India in the coming years. The current study's findings provided direct support for the potential use of accurate forecasts in decision-making, assist the government in formulating policies, production, import, and/or export and cotton production management in India.

References

Assis, K., Amran, A., Remali, Y. and Affendy, H. (2010). A comparison of univariate time series methods for forecasting cocoa bean prices. *Trends in Agricultural Economics*, **3**, 207–215.

Biswas, B., Dhaliwal, L. K., Singh, S. P and Sandhu, S. K. (2014). Forecasting wheat production using ARIMA model in Punjab. *International Journal of Agricultural Sciences,* **10(1)**, 58 -161

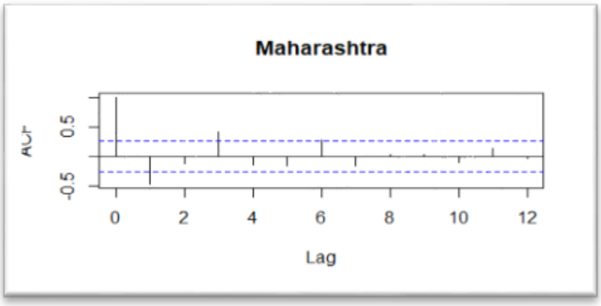
Box, G. E. P. and Jenkins, J. M. (1970). *Time Series Analysis – Forecasting and Control*. Holden-Day Inc., San Francisco.

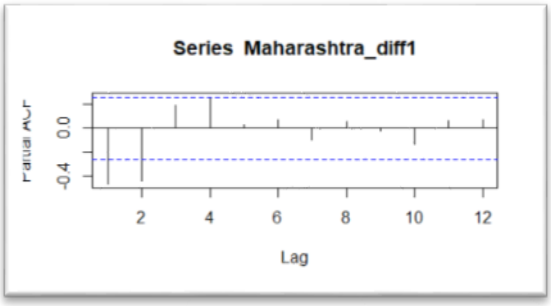
Cooray, T. M. (2006). Statistical analysis and forecasting of main agriculture output of Sri Lanka: rule-based approach. Appeared in 10th International Symposium, **221**: 1–9. Sabaragamuwa University of Sri Lanka

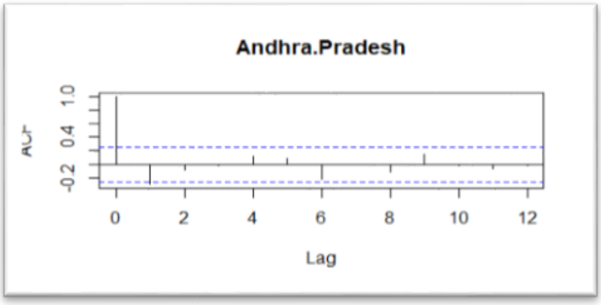
Hossain, M. Z., Samad, Q. A. and Ali, M. Z. (2006). ARIMA model and forecasting with three types of pulse prices in Bangladesh: A case study. *International Journal of Social Economics*, **33**, 344–353.

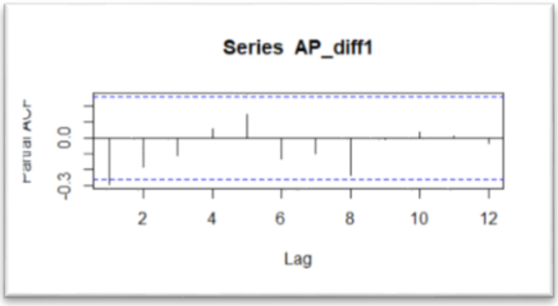
Mandal, B. N. (2005). Forecasting Sugarcane Productions in India with ARIMA Model. *Inter Stat*, October, 2005.

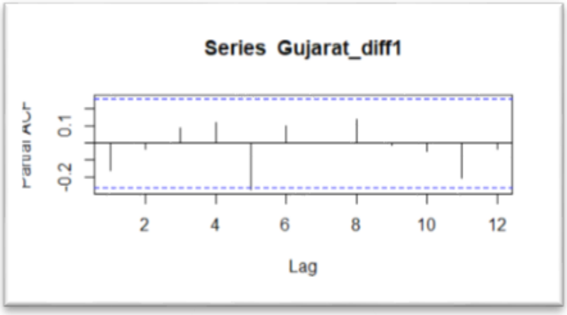
Appendix

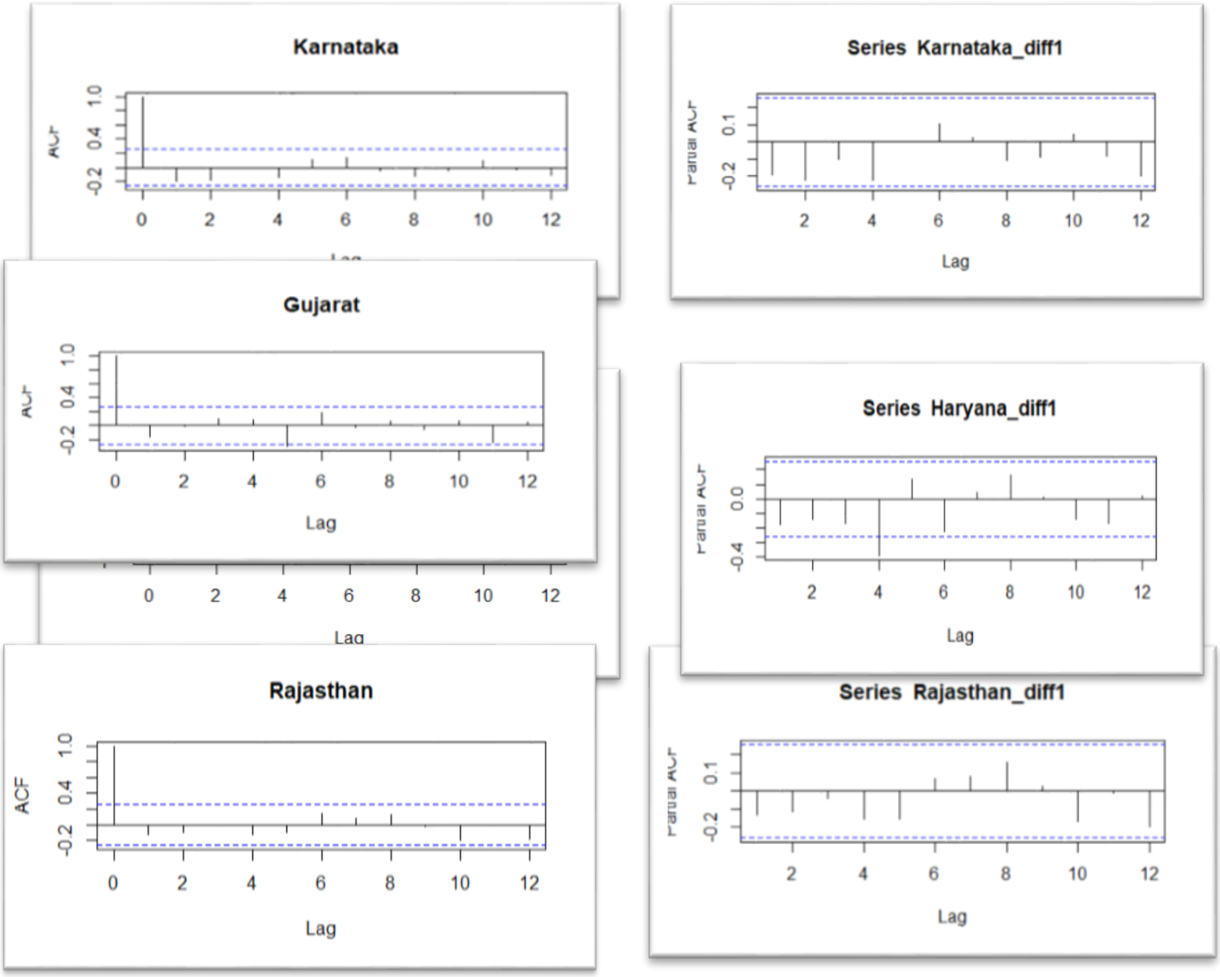


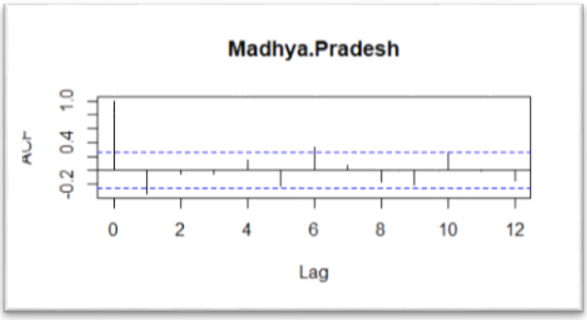


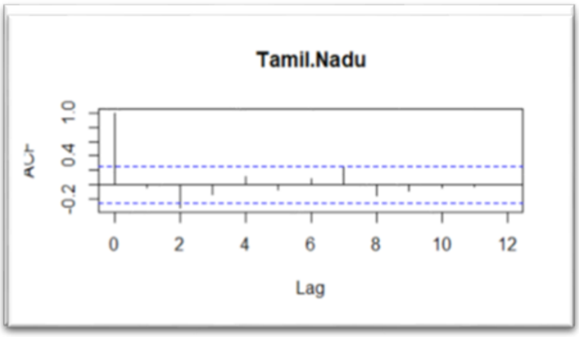


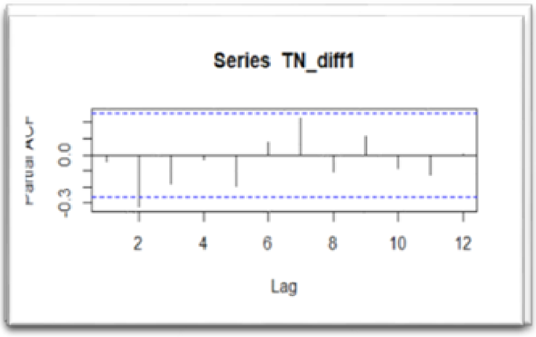





Karnataka
ACF
-0.2 0.4 1.0
0 2 4 6 8 10 12
Lag
Series Karnataka_diff1
Partial ACF
-0.2 0.1
2 4 6 8 10 12
Lag
Gujarat
ACF
-0.2 0.4 1.0
0 2 4 6 8 10 12
Lag
0 2 4 6 8 10 12
Lag
Series Haryana_diff1
Partial ACF
-0.4 0.0
2 4 6 8 10 12
Lag
Rajasthan
ACF
-0.2 0.4 1.0
0 2 4 6 8 10 12
Lag
Series Rajasthan_diff1
Partial ACF
-0.2 0.1
2 4 6 8 10 12
Lag


Madhya.Pradesh
ACF
-0.2 0.4 1.0
0 2 4 6 8 10 12
Lag


Tamil.Nadu
ACF
-0.2 0.4 1.0
0 2 4 6 8 10 12
Lag


Series TN_diff1
Partial ACF
-0.3 0.0
2 4 6 8 10 12
Lag

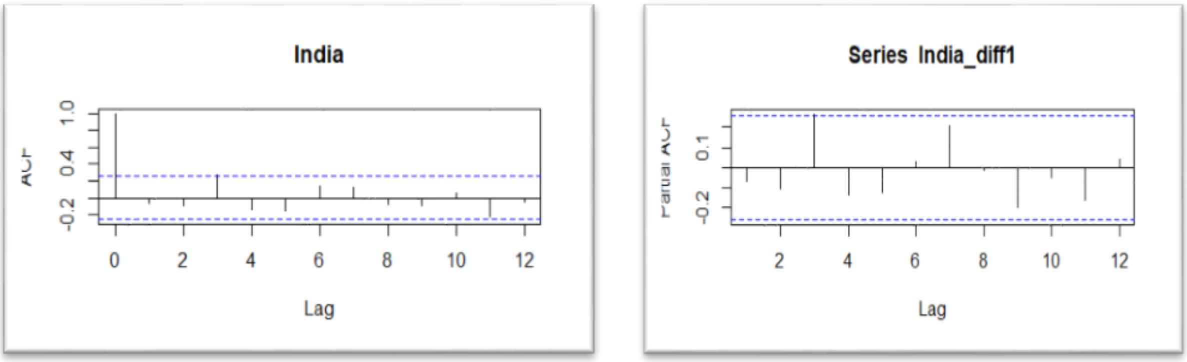
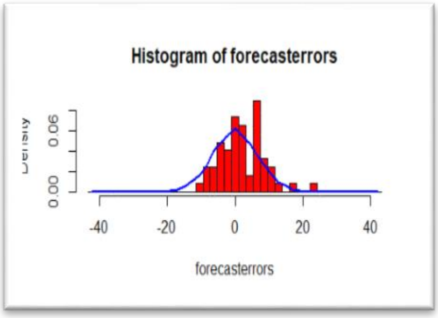
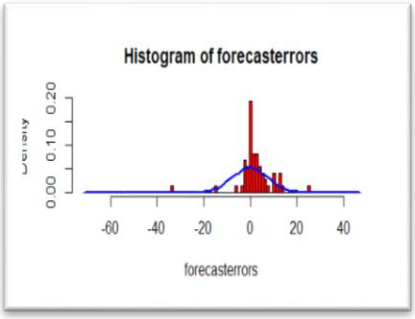


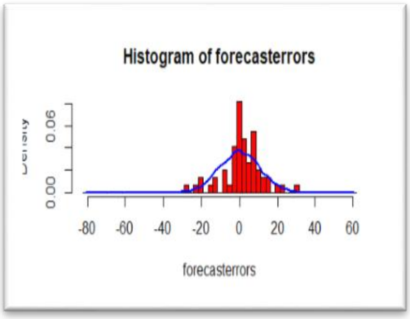
Figure 1: Auto-correlation function (ACF) and Partial auto-correlation function (PACF) of fitted ARIMA models for major cotton producing states



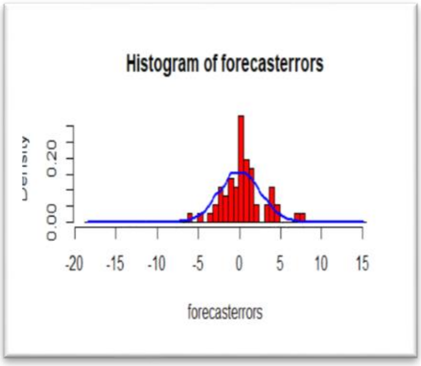
Maharashtra



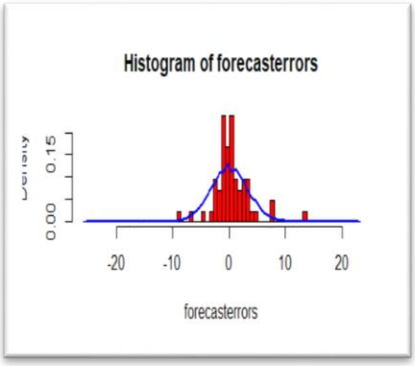
Andhra Pradesh



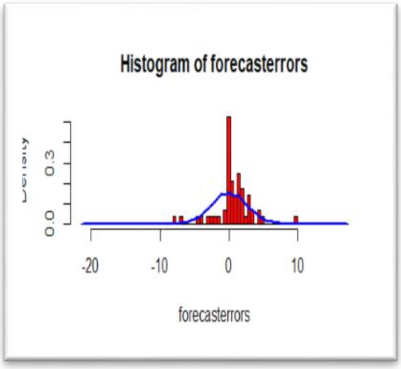
Gujarat



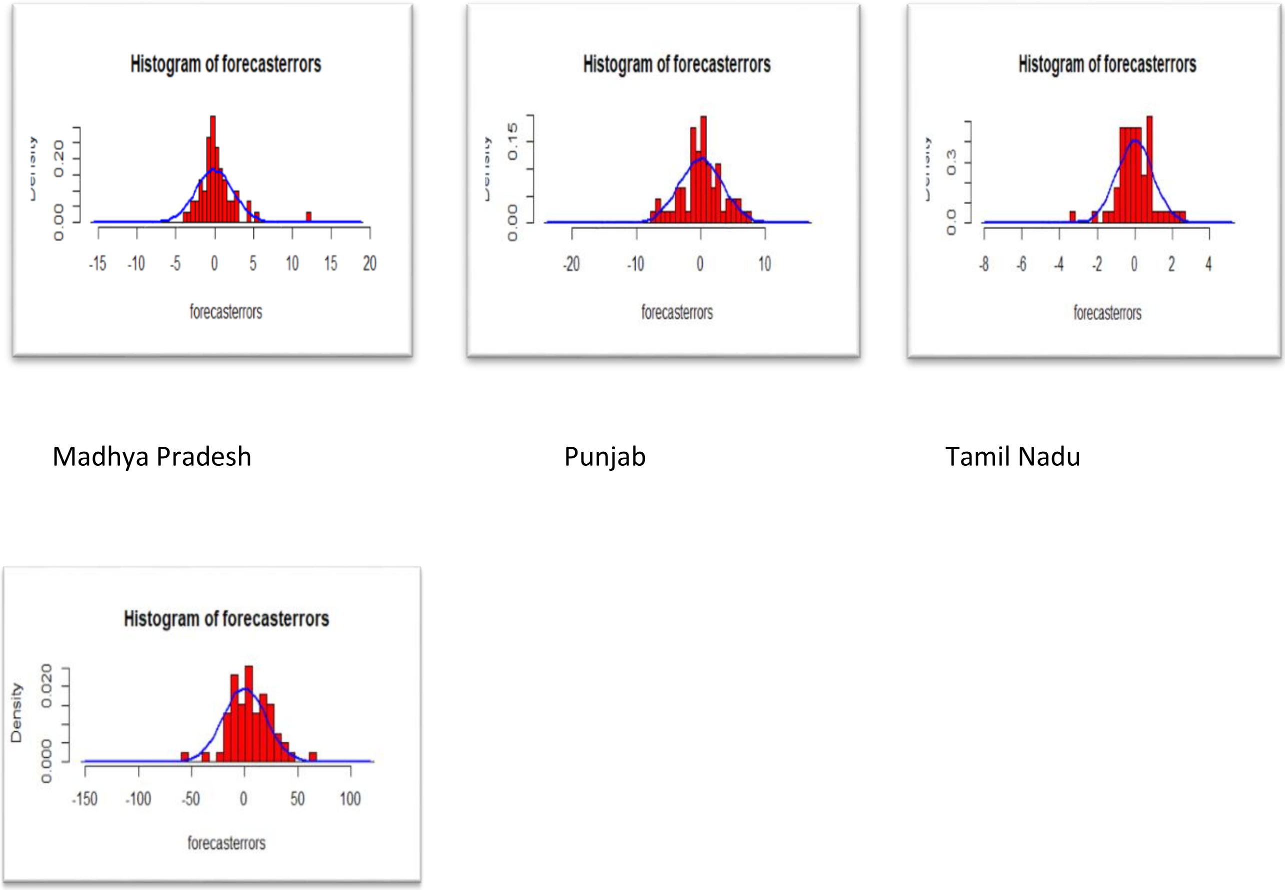
Rajasthan



Karnataka



Haryana



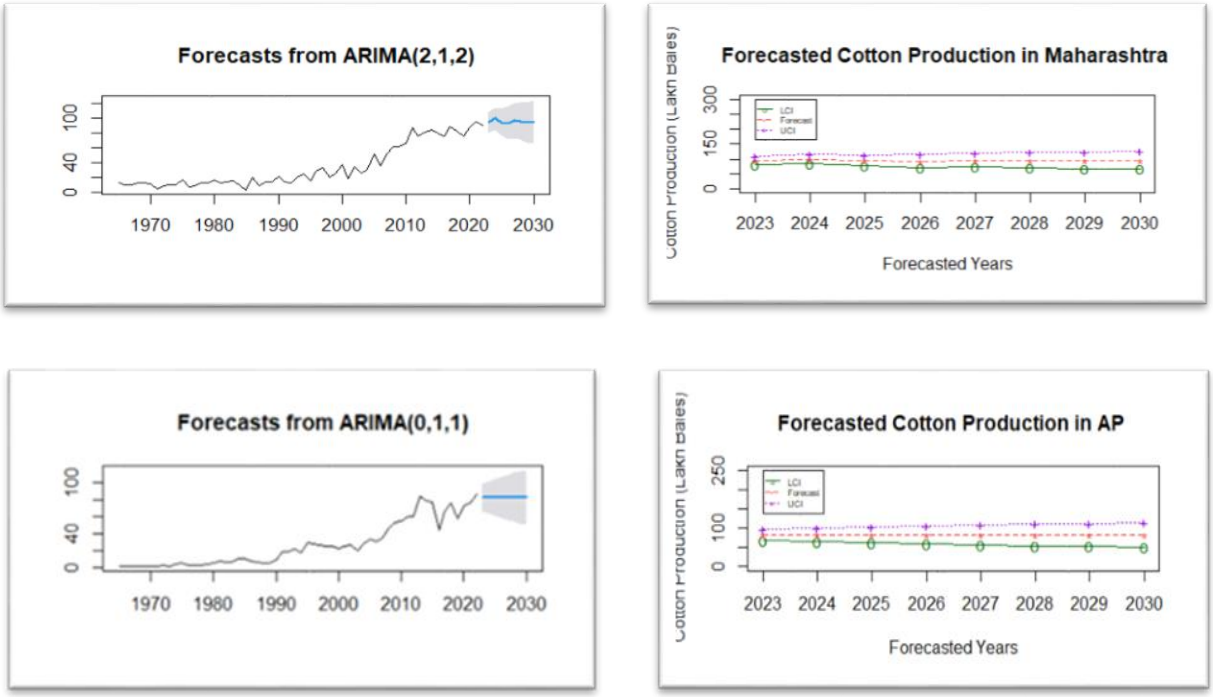
Madhya Pradesh

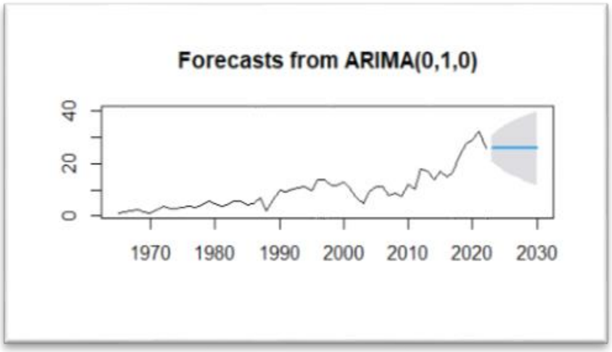
Punjab

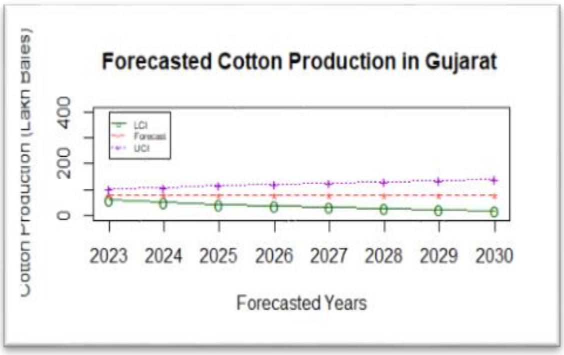
Tamil Nadu

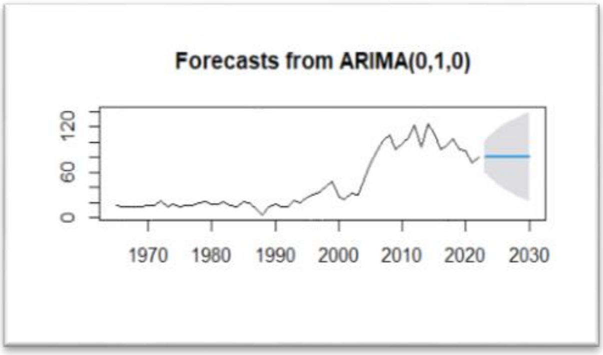
India

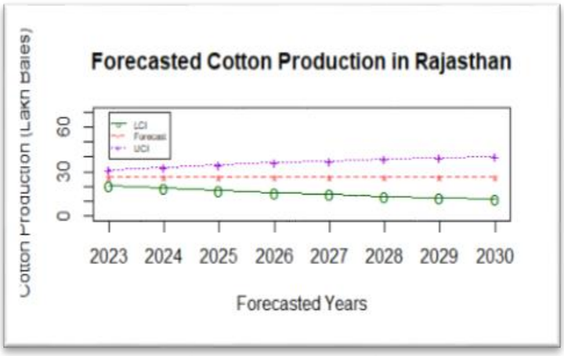
Figure 2: Histogram of forecast errors of residuals of major cotton production states in India

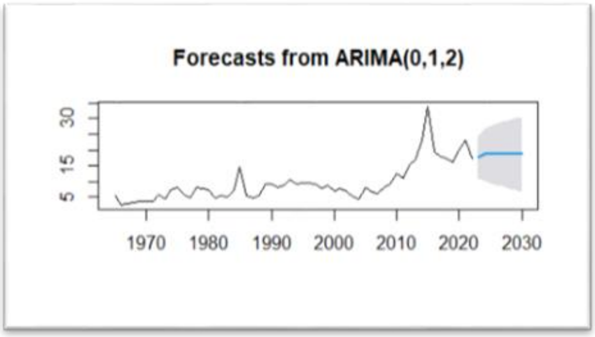


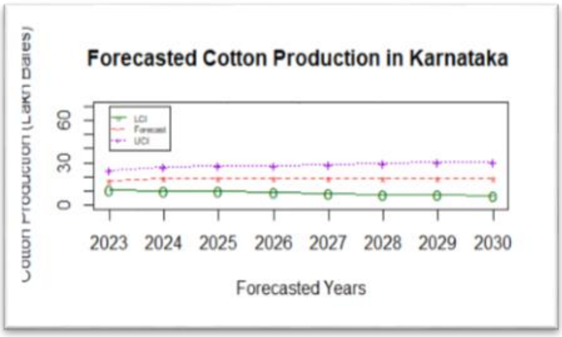

Forecasts from ARIMA(0,1,0)
0
20
40
1970
1980
1990
2000
2010
2020
2030

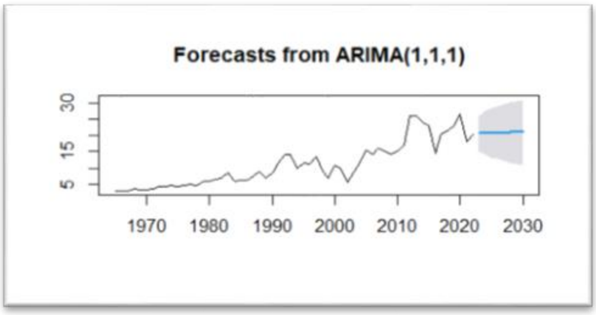

Forecasted Cotton Production in Gujarat
Cotton Production (Lakh Bales)
0
200
400
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

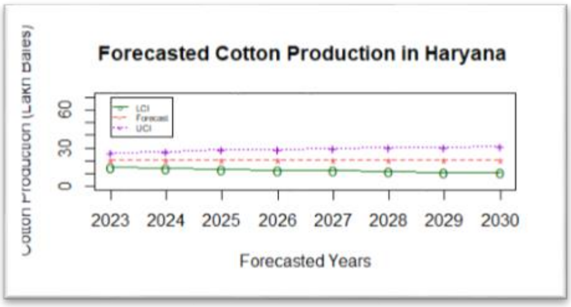

Forecasts from ARIMA(0,1,0)
0
60
120
1970
1980
1990
2000
2010
2020
2030

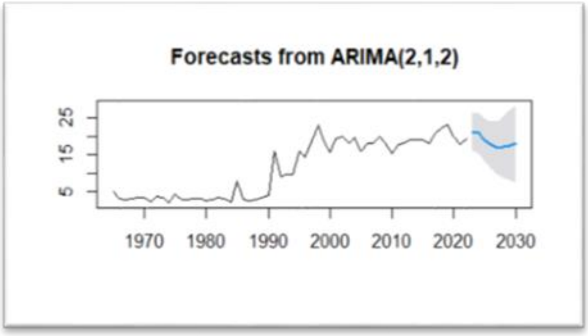

Forecasted Cotton Production in Rajasthan
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

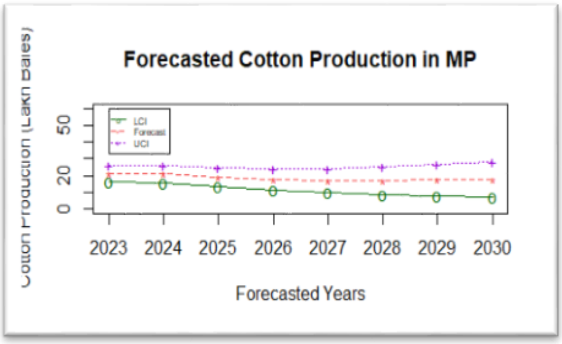

Forecasts from ARIMA(0,1,2)
5
15
30
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in Karnataka
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years


Forecasts from ARIMA(1,1,1)
5
15
30
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in Haryana
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years


Forecasts from ARIMA(2,1,2)
5
15
25
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in MP
Cotton Production (Lakh Bales)
0
20
50
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

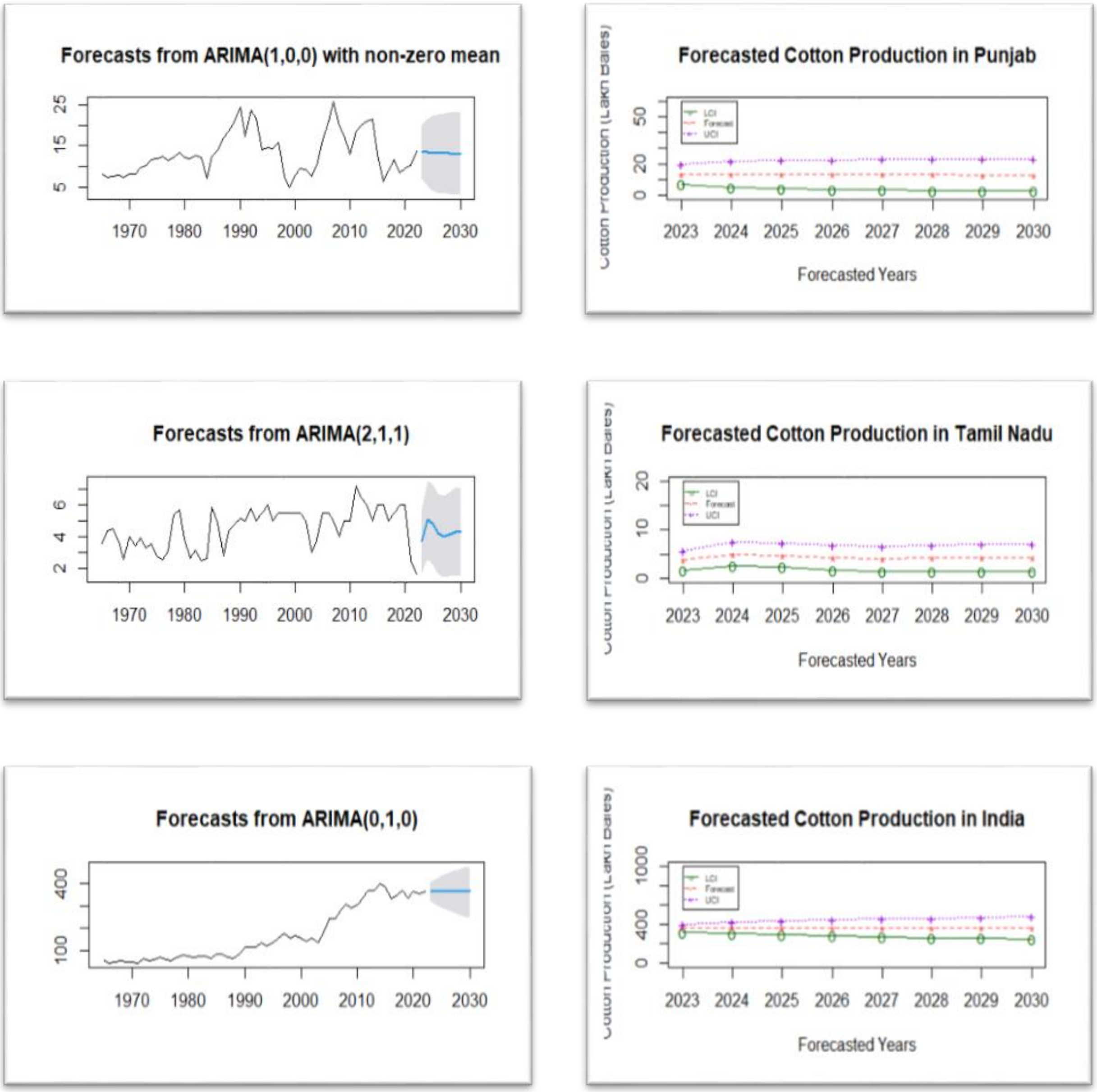


Figure 3: Observed and forecasted plots of major cotton production states in India

Disclaimer: The views expressed in this paper are of the author only and that the Gokhale Institute is not responsible for it.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 129-148

Inequalities among Agricultural Households: An Exploration Through Various Agricultural Surveys

K.J. Satyasai and Vinay Jadhav

Department of Economic Analysis and Research, National Bank for Agriculture and Rural Development, Mumbai-400013, Maharashtra

Received: 15 June 2022; Revised: 27 July 2022; Accepted: 30 July 2022

Abstract

While agriculture has grown impressively over the years, inequalities in access to resources and incomes remained high. The present paper attempted to capture the inequality across states, social groups, size class of land possessed with respect to various economic parameters concerning agricultural households using the data from the 70th and 77th round of NSSO's Situational Assessment Survey of agricultural household pertaining to agricultural year (AY) 2012-13 and 2018-19 respectively and NABARD All India Rural Financial Inclusion Survey (NAFIS) conducted for the reference period AY 2015-16. The paper has examined the changes with respect to parameters such as income, debt, access to credit, purpose of debt, etc. Our analysis in this paper shows that the outstanding debt as percentage of annual income increases as size class of land possessed increases and the same ratio as significantly increased from AY 2012-13 to AY 2018-19 across all size class of land except for HH possessing land less than 1 hectare. The level of indebtedness among agricultural HH has increased over the years across all size class of land but a huge variation in the level of indebtedness is seen across the states, with levels varying from 93.2% in Andhra Pradesh to 6% in Nagaland. The southern states lead the states where the average amount of loan per agriculture household is high. Increase in indebtedness can be attributed to increase in the reach of formal credit sources, whose share touched 69.6% in 2018-19 from 59.8 % in 2012-13. However, significant imbalance in the distribution of credit across size class of land is still evident and relatively higher dependence on informal source of credit by agricultural household possessing small land size can be seen. In terms of purpose for which this credit is used, a clear distinction can be seen among agricultural households belonging to different size class of land. Analysis reveals that agricultural household possessing large size of land, have high proportion of outstanding loan for combined expenditure in farm business (i.e capital and revenue expenditure) whereas agricultural household possessing smaller land have more than half (50%) of their loan for the purpose other than that for agriculture. On the income front, the average monthly income per agricultural household increased from Rs 6426 in the AY 2012-13 to Rs. 10,218 in AY 2018-19 registering a growth of 59%. But this rise in income was not uniform across the country, with smaller states like Bihar, Meghalaya, Mizoram, and Uttarakhand witnessing almost double the average monthly income since AY 2012-13. SC, ST and OBC agricultural households earn less compared to households under 'others'. Analysis of NAFIS data shows a wide range in monthly surplus (income-expenditure) per rural household across states. Punjab and Kerela topped the list with roughly Rs 4000 surplus, compared to states like Andhra Pradesh, Bihar, Jharkhand, Sikkim, and Uttar Pradesh with monthly surplus less than Rs 350. With an average monthly surplus as low as Rs 1413 at all India level reflects rural and agri household's high vulnerability to any unforeseen situations.

Name of the Corresponding Author: K.J. Satyasai
Email: **satyasaik@outlook.com**

1. Introduction

Inequality refers to the phenomenon of unequal and/or unjust distribution of resources and opportunities among members of a given society. Inequality in access to resources results in inequality in incomes too. Indian agriculture is characterized by small land holdings. Income inequality is very high across farm size categories. According to Chakravarty 1987, “No sustainable improvement in the distribution of incomes is possible without reducing the ‘effective’ scarcity of land”. Continuing fragmentation of landholding has resulted in Income from wages and not the cultivation as one of the important sources for small and marginal farmers.

Apart from land distribution, access to credit can determine input use across farm size classes and thereby the income. Credit plays pivotal role in the agricultural production. Of all the sources of credit, institutional sources offer cheaper credit compared to the informal sources such as private money lenders, the difference in cost of credit between the two being more than two to three times. Thus, access to institutional sources of credit means lower costs and hence, higher net income. The empirical studies have highlighted that share of institutional credit has been rising over the years across states and size class of land. However, the persistence of money lenders in the rural credit market is still a major concern with small holders depending more on informal sources. Satyasai et al. (2017), have highlighted that small landholders and SC/ST households face disadvantages in terms of access to credit and the degree of institutionalization is lower for ST, SC and OBC households.

Thus, study of inequalities in resource access and incomes in the light of fresh evidence becomes important to understand the issue better. This paper seeks to examine inequality in credit access and incomes based on 77th round of NSO Situation Assessment Survey of agricultural households.

2. Data And Methodology

The data on income and credit distribution for the paper has been culled from NSO 77th round Situation Assessment Survey of Agricultural Households and All India Debt & Investment Survey. Inequalities across farm size categories and states are measured using Gini Coefficient and also heatmaps.

3. Indebtedness Level Across Farm Sizes & States

Low scale and low productivity characterise Indian agriculture. Around 86 % of the country's operational landholdings are less than 5 acres, while 68 % of farm households live on less than one acre. Furthermore, irrigation is unavailable to more than half of the land under agriculture. Surplus from unprofitable crop farming is insufficient to invest in modern agriculture, which necessitates the acquisition of farm machinery and the usage of purchased inputs such as seed, fertiliser, agri-chemicals, diesel, and hired labour. Hence, farmers avail loans to meet cultivation expenses (working capital), invest on farms and meet their consumption requirements. According to the NSSO's Situational Assessment Survey (SAS) 2019, indebted agricultural households decreased from 52 % in 2012-13 to 50.2 % in 2018-19. Despite a drop in the percentage of indebted agricultural households, the average outstanding loan among agricultural households climbed by 58 %, from Rs 47,000 to Rs 74,121. Among indebted agricultural households, 82.9% were landless, marginal and small farmers. NABARD's All India Rural Financial Inclusion Survey (NAFIS 2016-17) found that 52.5 % of agricultural households and 42.8 % of non-agricultural households were in debt at the time of the survey. As of the date of the survey, each agricultural household had an average outstanding loan of Rs. 1,04,602. Farmers' indebtedness is rising for a variety of reasons, including increased access to institutional finance, agricultural mechanisation, and high-value agriculture. The cost of health care, education, social gatherings, and non-food items has increased, putting further financial pressure on farming families.

3.1. Indebtedness across farm size class categories

In 2018-19, the average farming household in India owed 60% of their annual income as debt. The ratio has not changed significantly during 2012-13 to 2018-19, but a closer examination of its distribution reveals that it has increased for all land sizes greater than 1 hectare, while decreasing for land sizes less than 1 hectare (Table 1). For example, for the size class (10.00+ hectare), the ratio nearly doubled (from 58.45 % in 2012-13 to 108.51 % in 2018-19), and for the size class (<0.01 hectare), the ratio nearly halved (from 56.82 % in 2012-13 to 20 % in 2018-19). This shows that large farmers witnessed a massive rise in debt, which is much more than small farmers. The indebtedness of large farmers rose significantly in comparison to their income. On the other hand, small and marginal farmers appear to be in a better situation regarding the level of debt. The average amount of outstanding loans per agricultural household increased with the rise in the possessed land's size class. Among small and marginal farmers owning less than 1 hectares, slightly less than 50% of the households were in debt.

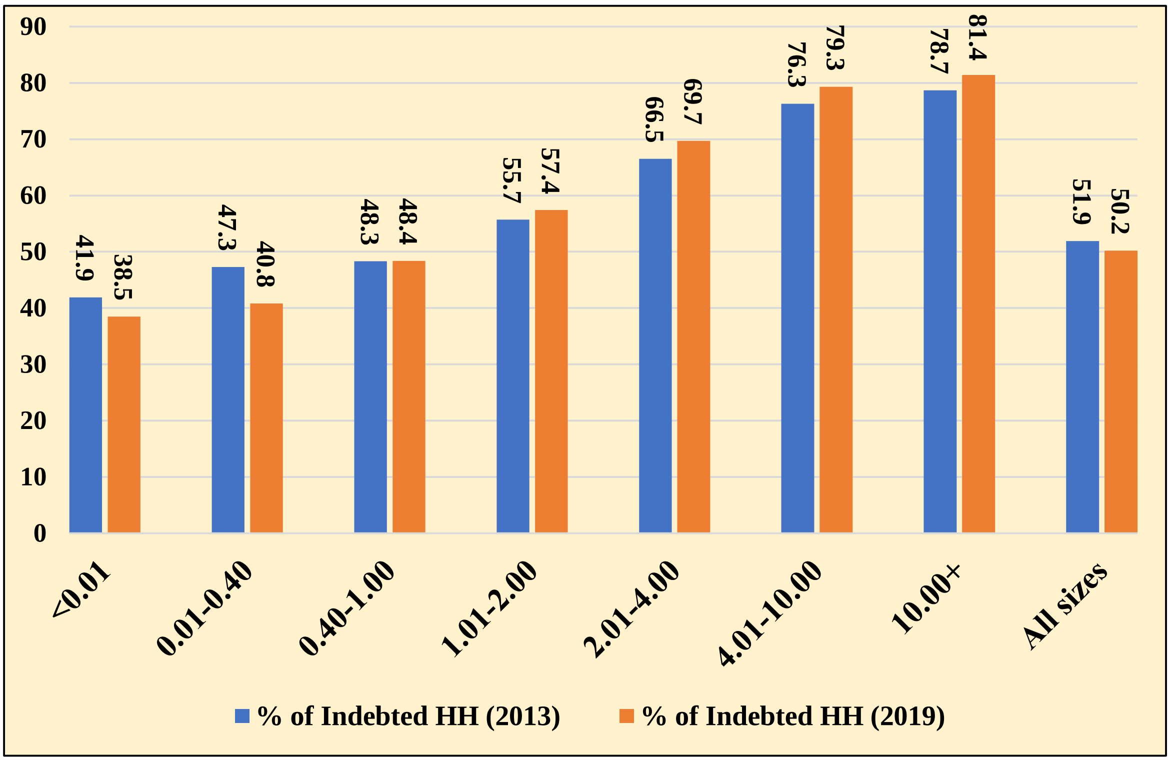
Table 1: Average debt as percentage of annual income across size class of land for the period AY 2012-13 and AY 2018-19

Size class of land possessed (ha.)	2012-13			2018-19			Increase in Debt (%)	Increase in Income (%)	Borrowing Propensity
	Average Debt	Average Annual Income	Ratio (%)	Average Debt	Average Annual Income	Ratio (%)	(8)	(9)	(8/9)
<0.01	31100	54732	56.82	26883	134448	20.00	–13.56	145.65	–0.09
0.01- 0.40	23900	49824	47.97	33220	90264	36.80	39.00	81.17	0.48
0.41-1.00	35400	62964	56.22	51933	102852	50.49	46.70	63.35	0.74
1.01-2.00	54800	88176	62.15	94498	137388	68.78	72.44	55.81	1.30
2.01-4.00	94900	128760	73.70	175009	197220	88.74	84.41	53.17	1.59
4.01-10.00	182700	235644	77.53	326766	339384	96.28	78.85	44.02	1.79
10.00+	290300	496656	58.45	791132	729096	108.51	172.52	46.80	3.69
all sizes	47000	77112	60.95	74121	122616	60.45	57.70	59.01	0.98
Gini Coefficient	0.4655	0.4524		0.5722	0.4049				

Source: Authors calculation on 70th and 77th round data of SAS.

During the period 2012-13 to 2018-19, the percentage rise in income and debt has stayed relatively consistent with each other when looked for "All sizes" category (57.7 % - for debt and 59 % for income). However, the distribution changed across farm size classes. For the first size class (<0.01 hectare) debt has declined between two time points. Hence, propensity to borrow (ratio of % change in debt to % change in income) was negative. For the next two classes (0.01-0.40 and 0.40 – 1.00 hectare) the growth in income outweighed the growth in debt during the period. Thus, the propensity to borrow is less than unity. Agri Households in the category of land size class > 1 hectare are clearly more leveraged than those in the other categories, with a borrowing propensity of more than one (Table 1). Large farmers (>10 hectares) added 173% to their credit burden between 2012-13 and 2018-19 than they could add to their income. Their borrowing propensity being 3.7. This has implications for debt servicing ability in case of agricultural losses.

The proportion of household indebted as also seen an increase over the years. For those with land size classes 'less than 0.01 hectare', '0.01-0.40 hectare', '0.40-1.00 hectare', '1.01-2.00 hectares', '2.01-4.00 hectares', '4.01-10.00 hectares', and more than 10.00 hectares', respectively, the proportion of indebted farm households in 2018-19 was 38.5 %, 40.8 %, 48.4 %, 57.4 %, 69.7 %, 79.3 % and 81.4 %. Except for the size classes < 0.01 and 0.01-0.40, which experienced drop of 3.4 and 6.5 percentage points respectively, there was a marginal upward movement in the proportion of agri household indebted as compared to 2012-13 in other size classes (Graph 1). Furthermore, the percentage of indebted agricultural HH increases as land size increases.



Source: - NSO's 70th and 77th rounds of SAS

Graph 1: Percentage of indebted agricultural household across size class of land

When it comes to the frequency with which agricultural households took out loans, those with more land were clearly more likely to have multiple loans. This could be due to the fact that these economically better-off households are more likely to take out loans since they have sufficient assets to serve as collateral for the loans. According to NABARD All India Rural Financial Inclusion Survey (NAFIS), there is higher proportion of households of land size (> 2.00 ha) in the category Two loans and 3-5 loans in the reference period (July 2015- June 2016). For Example, 15.4 and 7.5 % of households having more than 2.0 ha land took two loans and 3-5 loans respectively compared to 10.8 and 2.2 % of household belonging to land holding category (1.01-2.0 ha). The data in Table 2 on the distribution of agricultural households reporting multiple loans according to farm size classes show that the appetite for taking more loans is higher among above 2 hectare farm size classes with 23% households taking more than 2 loans.

Table 2: Distribution of agricultural households reporting any loan by number of loans taken by size class of land possessed (in %)

Category	No. of Loans taken during July,15 to June, 16			Total
	One Loan	Two Loans	3-5 Loans	
1	2	3	4	5
<0.01 ha	85.7	11.7	2.6	100.0
0.01-0.4 ha	83.0	14.3	2.7	100.0
1.01-2.0 ha	87.0	10.8	2.2	100.0
>2.0 ha	77.1	15.4	7.5	100.0
All Size Classes	83.2	13.4	3.4	100.0

Source: NAFIS, 2015-16

3.2. Indebtedness across states

The level of indebtedness also varied across the states, 93.2% in Andhra Pradesh and 91.0% % in Telangana to 25.3 % in Jharkhand and 6 % in Nagaland. Andhra Pradesh had the highest average outstanding loan (Rs. 2,45,554), followed by Kerala (Rs. 2,42,482) and Punjab (Rs. 2,03,249). Agricultural households in 11 of the 28 states reported borrowing more than the national average, with at least eight having an average outstanding loan of more than Rs 1 lakh. All southern states on an average reported more than Rs 1 lakh in outstanding loans per household.

The distribution of indebtedness in 2018-19 have not changed much compared to that in 2012-13 at both the state and national level. The proportion of indebtedness has decreased by just 1.7 percentage points over 6 years at national level, while not much change in terms of proportion of households indebted was seen at the state level either. It can be noted that, in both time periods, southern states (*viz*. Kerela, Tamil Nadu, Andhra Pradesh, Telangana and Karnataka) remained at the top. The proportion of household indebted in Andhra Pradesh, Telangana, Kerela, Karnataka, Tamil Nadu are 93.2, 91.7, 69.9, 67.6, 65.1 respectively. Agri-household indebtedness is quite low in the NE States and Jharkhand. The same pattern is emerged in NSO's All-India Debt and Investment Survey (AIDIS) for 2018-19.

4. Access to Credit and it’s Utilisation

Agricultural credit by providing necessary capital for meeting the ever-increasing demand for productivity and efficiency has played an important role in the development of the farm sector. Agriculture credit, though not a direct input for production, can help to raise farmers from low productivity trap by removing financial constraints and accelerating the adoption of new technologies. Over the years, the government of India's policies and interventions have yielded appreciable results in the field of agricultural credit. However, many reports have highlighted that dependence of farmers especially small and marginal farmers, tenant farmers, landless labourers and sharecroppers on non-institutional sources of credit is high even though the credit from these institutions is available at significantly higher rate of interest.

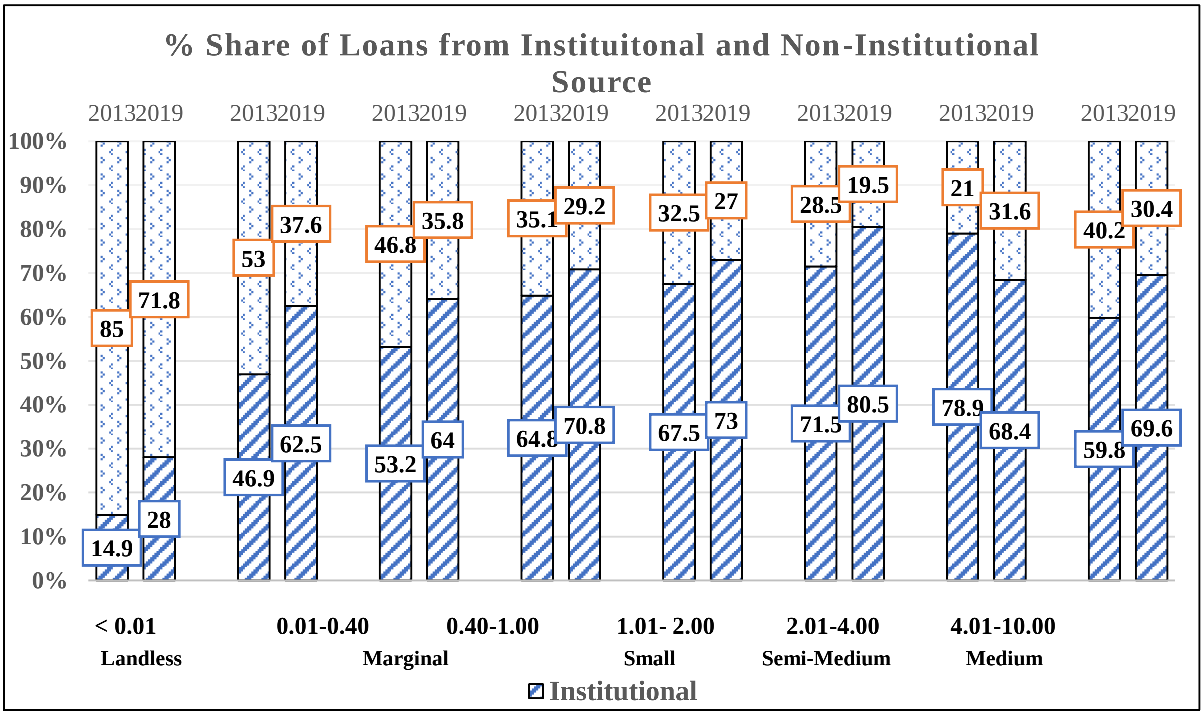
According to the NAFIS (2015-16), 30.3% of Agricultural Household borrowed only from Non-Institutional Sources while it was 9.2% who borrowed from both Institutional and Non-Institutional sources. The report further highlights that 28% (*i.e.*, Rs. 29,611) of loan taken by Agricultural Household comes from non-institutional sources, thus indicating a sizeable proportion of loan requirement met by non-institutional source. Lengthy application process, excessive collateral requirement and short loan term (maturity) were some of the reasons cited for not taking loan from institutional sources.

4.1. Inequality in access to credit across size class of land

Even though formal credit sources have expanded their reach and their proportion of agri credit has increased considerably year on year, the 77th round survey results show a major disparity in institutional loan distribution across the land size classes. It is clear that agri households with small land sizes have a larger reliance on informal sources of financing. Except in the case of the largest farms (>10.00+ hectare), SAS data demonstrate a link between farm size and access to institutional finance, with reliance on non-institutional loan sources such as money lenders and relatives growing as land holding decreases (Graph 2).

Institutional sources (SCBs, RRBs, Co-operative societies, co-operative banks, SHGs, and other institutional agencies) provided Rs 64 of the Rs 100 taken by agricultural households with land between 0.40 and 1.00 hectares, while institutional sources provided Rs 81 of the Rs 100 taken by the agricultural households with land between 4.01 and 10.00 hectares. The percentage of credit from institutional sources was 28 %, 62.5 %, 64 %, 70.8 %, 73 %, 80.5 %, and 68.4 %, respectively, for possessed land size classes 'less than 0.01 hectare', '0.01-0.40 hectare', '0.40-1.00 hectare', '1.01-2.00 hectares', '2.01-4.00 hectares', '4.01-10.00 hectares', and more than 10.00 hectares. There was a significant increase in percent increase of institutional credit among all size classes, with the exception of the size class greater than 10.00 hectares, which saw a 10.5 percentage point drop. (Graph 2: shows the percentage distribution of amount of outstanding loans by sources in 2018-19 compared to 2012-13 across size class of land).

Source: NSO's 70th and 77th rounds of SAS



Graph 2: Percentage distribution of Outstanding loan by source for each land size category

The KCC scheme, which was launched in 1998, has emerged as an innovative credit delivery mechanism for meeting farmers' credit needs at various stages in a timely and hassle-free manner. The scheme has become one of the major tools of government to bring more farmers under to the gamut of institutional credit. Over the year its reach has improved and has become successful in providing institutional credit to farmers at concessional rate of interest. The situational assessment survey conducted in 2018-19, have collected the information on percentage of agricultural household possessing KCC, throws relevant light on unequal access to credit. The data shows unequal access of KCC across size class of land and the proportion of households reporting KCC (penetration) increased significantly with increase in land sizes. For Example, 48.7% agri household possessing land (>10.00+ hectare) have access to KCC, while it is only 19.4% and 9.7% for Agri Household possessing land 0.41-1.00 and 0.01-0.40 ha. The lower proportion at the bottom end of the spectrum hints that these households may not be pursuing cultivation on a scale and hence their need for KCC and eligibility may be less. Still the difference in the proportion between small, semi-medium and large farmers is a matter of further study.

According to the All-India Debt and Rural Investment Survey (AIDIS), institutional sources alone are unable to meet the credit needs of cultivator households (All the households having area of land operated 0.002 hectares or more were considered as ‘cultivator household’), and a considerable percentage of cultivator households rely on non-institutional sources for loans. In addition, the survey reveals a pattern of borrowing from institutional and non-institutional sources, depending on the purpose of loan. The majority of loans (64%) obtained from institutional sources by cultivator households were used for farm business and non-farm business (Table 3), although the possibility that these loans were not diverted for consumption purposes cannot be fully ruled out completely. However, data clearly demonstrates that the majority of loans (56%) obtained from non-institutional sources was spent on household expenditure, housing, others, etc. This clearly shows that farmer households have to rely substantially on non-institutional sources to carry out their daily activities.

Table 3: Rs. 1,000 breakup of amount of cash loan outstanding by purpose of loan for cultivator households

State/UT/ All India	Credit Agency	Purpose of loan	cultivator	
			per 1000 no. of households reporting cash loan outstanding	cash loan (Rs.) per Rs. 1000 of total cash loan outstandin g
All-India	Institutional	capital expenditure in farm business	78	257
		revenue expenditure in farm business	122	309
		expenditure in farm business	197	566
		capital expenditure in non-farm business	9	51
		revenue expenditure in non-farm business	4	19
		expenditure in non-farm business	13	70
		expenditure on litigation	0	0

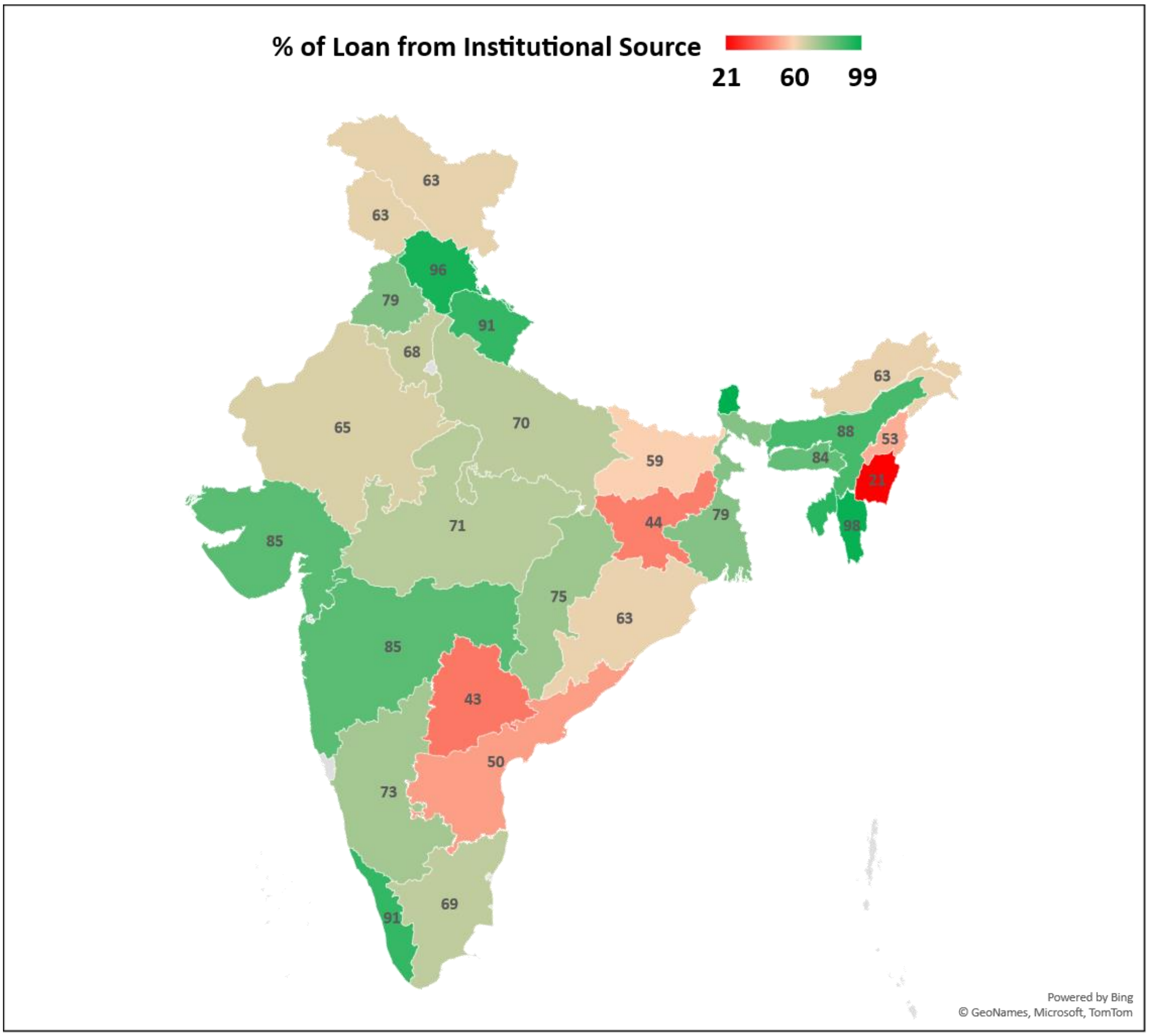
State/UT/ All India	Credit Agency	Purpose of loan	cultivator	
			per 1000 no. of households reporting cash loan outstanding	cash loan (Rs.) per Rs. 1000 of total cash loan outstanding
		repayment of debt	4	7
		financial investment expenditure	0	1
		for education	4	16
		for medical treatment	10	13
		for housing	23	177
		for other household expenditure	47	82
		Others	23	68
		All (incl. n.r.)	299	1,000
	Non-Institutional	capital expenditure in farm business	18	106
		revenue expenditure in farm business	27	162
		expenditure in farm business	43	267
		capital expenditure in non-farm business	3	32
		revenue expenditure in non-farm business	2	11
		expenditure in non-farm business	4	42
		expenditure on litigation	0	4
		repayment of debt	3	22
		financial investment expenditure	0	1
		for education	4	25
		for medical treatment	23	81
		for housing	24	155
		for other household expenditure	76	296
		others	22	106
		All (incl. n.r.)	190	1,000

Source: AIDIS, NSO's 77th round

4.2. Inequality in access to credit across states

The SAS 2018-19 found significant heterogeneity in the percentage share of formal/institutional credit sources in rural credit across states. More than 80% of rural credit supply comes from formal/institutional sources in states like Kerela, Uttarakhand, Himachal Pradesh, and Maharashtra. Non-institutional sources (agricultural money lenders, professional money lenders, relatives and friends, and other non-institutional sources) accounts for 57 %, 50 %, and 56 % of rural credit in states such as Telangana, Andhra Pradesh, and Jharkhand, respectively.

Institutional credit sources account for 70% of all agri-credit in 2018-19, up from 59.8 % in 2012-13. Though the percentage of institutional credit has been increasing, the 77th round findings reveal some alarming facts: agricultural states such as Bihar, Jharkhand, Uttar Pradesh, Telangana and Andhra Pradesh have a share of institutional credit that is less than the national average of 70%. These states have 59%, 44%, 70%, 43%, and 50% share respectively of institutional credit in total rural credit of state. This clearly shows that non-institutional sources of credit are still relevant in the agri-credit sector of major parts of the country. (Figure 1 shows the variation across the states wrt % share of formal/institutional credit in total rural credit).



Source: NSO's 77th round of SAS

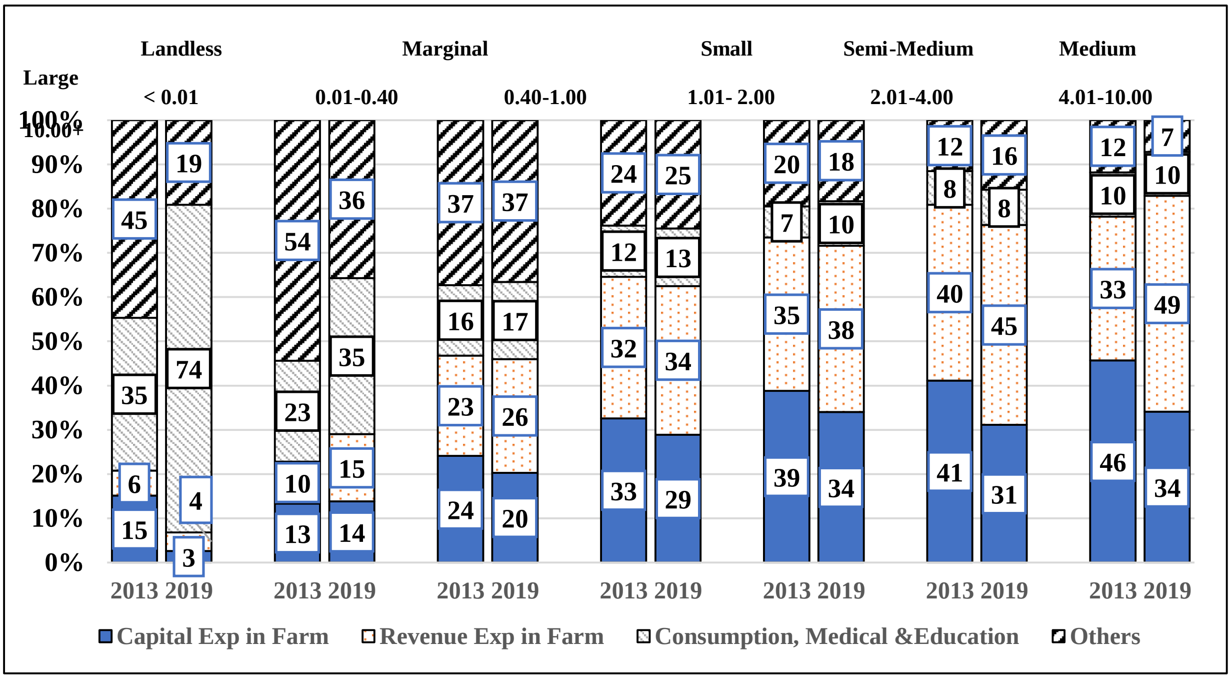
Figure 1: Percentage share of formal/institutional credit in total rural credit state wise (2018-19)

According to the Report of Internal Working Group to Review Agricultural Credit (RBI-2019). Some states receive substantially higher credit against their input cost requirements such as Andhra Pradesh (7.5 times), Kerala (6 times), Goa (5 times), Telangana, Tamil Nadu, and Uttarakhand (4 times), and Punjab (3 times). Jharkhand, NE states, West Bengal, Chhattisgarh, Bihar, Odisha, Maharashtra, Uttar Pradesh, and Rajasthan, on the other hand, are not receiving credit even to satisfy their input requirements. This illustrates the uneven distribution towards a few states and calls into question if the credit is being used for its intended purpose.

4.3. Purpose of the credit

The purpose of loan is defined as the event that prompted the households to take the loan. The purpose of loan taken by agricultural households belonging to different land size classes shows a notable contrast. Furthermore, across size classes of land owned, there is a clear downward trend in loans acquired for non-farm businesses. For example, it is 5.7 % for land size class (0.01- 0.40) and just 1.6 % for class (10.00 and above). According to the data, agricultural households with land sizes of (0.01-0.40) and (0.40-1.00) have more than half of their loans (71 % and 54 % in 2018-19, respectively) for purposes other than farm business. (*viz*. non-farm business, for housing, marriages and ceremonies, education and medical, other consumption expenditure, others). As the size class of land possessed increases, the percentage of loans obtained for purposes other than farm business decreases (Refer Graph 3). Data shows that agri-households with large landholdings had a higher share of outstanding loans for combined farm expenditures, for example, it is 76 % for Agri HH with land between 4.00 and 10.00 ha and 83 % for Agri HH with land over 10 ha compared to just 47% and 63% for household possessing land between 0.40-1.00 ha and 1.00-2.00 ha respectively. This clearly shows that as the size class of land possessed increased from 'less than 0.01 hectare' to 'more than 10.00 hectares,' a higher proportion of outstanding loan was taken for agricultural purposes and a lower proportion for non-agricultural purposes.

Graph 3 clearly indicates how purpose of loan taken by agricultural households over the years have changed towards Revenue expenditure on Farm, Consumption, Medical and Education expenditure. This pattern of variation is seen not only across the size class of land but also across the length and breadth of the nation. In Kerala, 33% of loans were for housing while only 27% were for agricultural purpose. The expenditure on agriculture is less than the national average (57.5%) for states such as Bihar, Uttar Pradesh, Telangana, West Bengal, Odisha *etc*. Few states such as Gujarat (41.6%), Maharashtra (26.7%), Madhya Pradesh (33.4%), Punjab (35.2%) have large share in capital expenditure in farm business. In NE states, significant proportion of loan is taken for housing, non-farm business and for other consumption expenditure.



Source: NSO's 70th and 77th round SAS

Graph 3: Purpose of loan taken by agricultural households across size class of land in 2012-13 and 2018-19

5. Income Level and Composition

The average monthly income of an Indian farmer after deducting paid-out expenses reached Rs 10,218 in 2018-19, showing an increase of 59% since the previous SAS survey conducted in 2012-13. Nominal income has grown at an annual compounded growth rate (CAGR) of 8%. Monthly average income increased by 16% after adjusting for inflation, at a CAGR of 2.5% (Table 4). When net receipts are calculated after deducting both paid out and imputed expenses, the average monthly income fell to Rs. 8,337. According to NSS 70th round the average monthly income of agricultural households was Rs 6,426 during the period July 2012 to June 2013. Between 2002-03 and 2012-13, the compounded annual growth rate of average monthly income(nominal) of agricultural household was 11.8%, but it slowed to 8 % between 2012-13 and 2018-19(Table 5). In the 77th round of the survey, household income included rent from leasing out land, which was not included in 2012-13. In 2018-19, the average household income is Rs. 10,084 without such rent.

Under its surveys, the NSSO has worked to improve its assessment methodology over time. Only landowner farmers were evaluated in 2002-03, but this requirement was removed in the 2012-13 evaluation, making the two figures not comparable. Many changes were made in the 2018-19 survey, including (i) the addition of a new source of farmer income, namely "revenue from leasing out land," and (ii) "an assessment of pensions and remittances received by the farmer household." The former is included in the monthly income calculation, while the latter is not. In other words, 'revenue from leasing out land' was included in the estimates of agricultural household income for 2018-19 (NSSO's 77th round), while this head of income was not recorded in the 2012-13 SAS (NSSO's 70th round), making the two estimates non-comparable. The 'revenue earned from leasing of land' must be removed from the 2018-19 estimate to make the two estimates comparable.

Table 4: CAGR (Nominal & Real) of average monthly income of agricultural household for the period between 2012-13 and 2018-19 (CPI-AL: Base 2012-13)

Size class of land possessed (ha.)	Total income (2012-13)	Total income (2018-19) *	CAGR % (Nominal)	CAGR % (Real)
<0.01	4,561	10,950	15.72	10.1
0.01- 0.40	4,152	7,333	9.94	4.6
0.41-1.00	5,247	8,495	8.36	3.1
1.01-2.00	7,348	11,375	7.55	2.3
2.01-4.00	10,730	16,289	7.21	2.0
4.01-10.00	19,637	27,841	5.99	0.8
10.00+	41,388	60,177	6.44	1.2
all sizes	6,426	10,084	7.80	2.5

Source: Authors calculation on 70th and 77th rounds of SAS (* Income from rent is excluded)

5.1. Diversification of income

Income from wages or non-farm businesses may be earned by an agricultural household in addition to income from agriculture. In 2018-19, wages, cultivation, animal farming, and non-farm business had a share of 40%, 38%, 16%, and 6%, respectively. In 2012-13, these percentages were 32%, 48%, 12%, and 8%, respectively. This suggests that farming or crop production is contributing to total income of a household to lesser extent, relatively.

Between 2012-13 and 2018-19, the growth of income realised through crop cultivation slowed dramatically. Between 2012-13 and 2018-19, the annual growth in crop cultivation income was negative -1.5 %, compared to 4.2 % annual growth from 2002-03 to 2012-13 (CPI (AL)- Base 2012-13)). In absolute terms, nominal crop production or cultivation income per agricultural household was Rs 3,798 in 2018-19, up 23% from 2012-13. In real terms, however, it has fallen by 8.7%.

Table 5: CAGR (Nominal & Real) of sources of income for time period 2002-03 to 2012-13 and 2012-13 to 2018-19 (Real: CPI (AL)- Base 2012-13)

Particulars	AY 2002-03	AY 2012-13	AY 2018-19	CAGR (2002-03 to 2012-13) Nominal	CAGR (2002-03 to 2012-13) Real	CAGR (2012-13 to 2018-19) Nominal	CAGR (2012-13 to 2018-19) Real
Income from wages	819	2071	4063	9.72	1.83	11.89	6.42
Net receipt from Crop production	969	3081	3798	12.26	4.19	3.55	-1.51
Net receipt from Farming of Animals	91	763	1582	23.69	14.79	12.92	7.41
Net Receipt from Non-Farm Business	236	512	641	8.05	0.28	3.82	-1.25
Total	2115	6426	10084	11.75	3.72	7.79	2.53

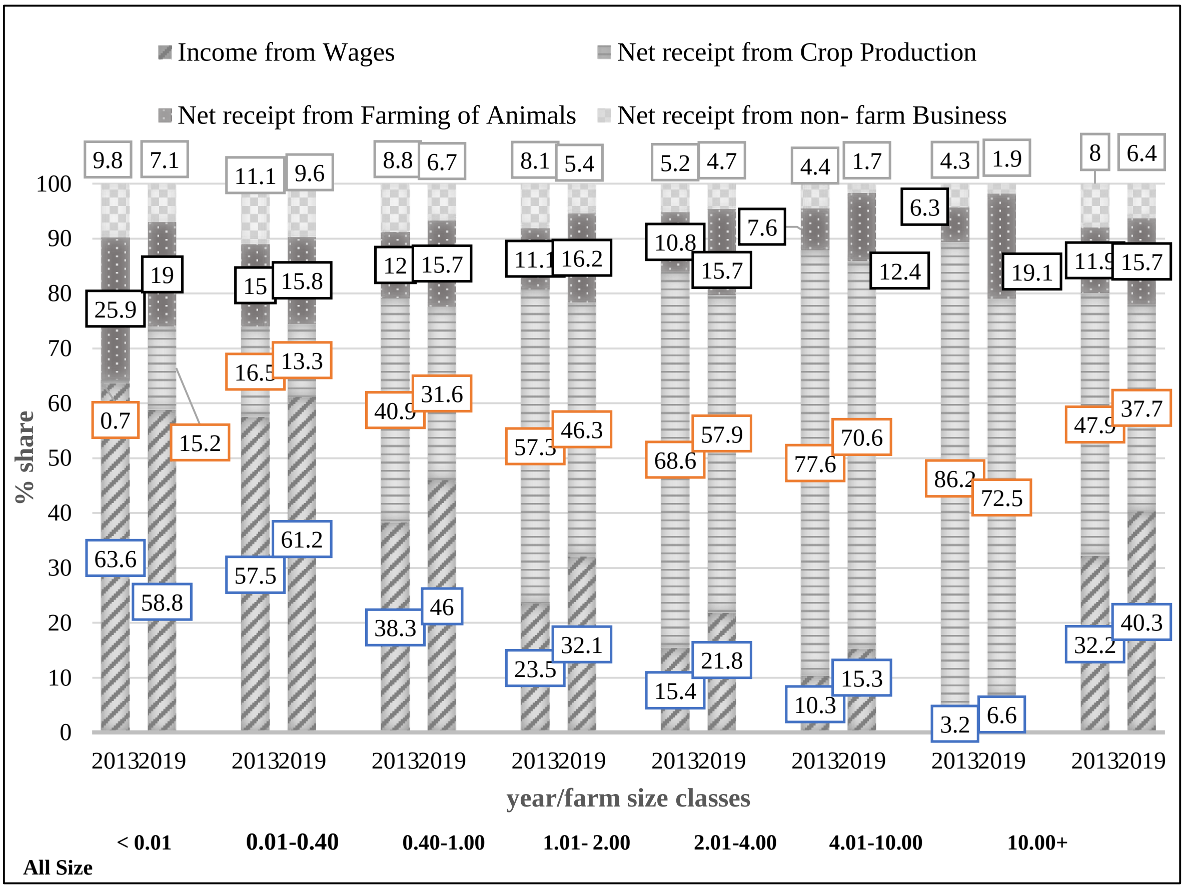
Source: Authors calculation on 70th and 77th rounds of SAS

The period between AY 2012-13 and AY 2018-19, also saw erosion of income from non-farm business. In real terms, non-farm income has declined from Rs 512 per month to Rs 475 per month in 2018-19. The 59th,70th and 77th round data clearly shows that revenue is mostly derived from wages and animal farming. Farmer's family sustained throughout the year from income primarily from livestock, as well as work on others' farms, MGNREGA, and other similar activities. Income from wages and net receipts from livestock witnessed a compounded annual growth rate of 6.4 % and 7.4 %, respectively, in real terms.

5.2. Income across farm size classes

NSO data from the 77th round points out huge disparity and variation in income and its composition across farm size classes (Graph 4). The distribution of income from different components varies significantly across the land size. Agricultural households owning land between 0.01 - 0.40 hectare earned more than half (60%) of their income from wages, it is 46% from wages for households owning land between (0.40-1.00 ha) compared to 6% of income from wages in case of agricultural household possessing land 10 hectares and above. Comparing with AY 2012-13, in AY 2018-19 the share of wages in income of all household size classes except for landless households have increased, showing the increasing dependence of agri households on wage labour to meet their financial needs.

Data would make it clear that when land size increases, the income share from net receipts of agricultural operations (crop production and animal farming) per agricultural household increases. It is 91 % for agri-households with land of 10 hectares and above, and 28 % for those with land of 0.01-0.40 hectares. The income disparity between agricultural households with 0.40 to 1.00 hectares and agricultural households with 10 hectares & more is significant, with the latter's average monthly income being eight times that of the former compared to 10 times during 2012-13, the income gap seems to have narrowed (Table 4).



AY 2012-13 and 2018-19

Source: NSO’s 70th and 77th round SAS

Graph 4: Percentage share in income activity wise across size class of land for period

The NSSO classifies farmer household income into seven categories based on land size holdings in hectares, namely (i) < 0.01, (ii) 0.01-0.4, (iii) 0.41-1, (iv) 1.01-2, (v) 2.01-4., (vi) 4.01—10, and (vi) 10 and above. For the landholding categories of <0.01, 0.01-0.4, 0.41-1, 1.01-2, 2.01-4, 4.01—10, and 10 hectares and above, the CAGRs of real incomes (deflated by CPI-AL are 10.1 %, 4.6 %, 3.1 %, 2.3 %, 2.0 %, 0.8 %, and 1.2 %, respectively. According to the recent Agricultural Census, India has 14.65 crore agricultural households, of whom 10.03 crore belong to the first three categories—this equates to around 68 % of the farmer population. The first three groups have an average CAGR of 6 % in real terms (Base: CPI-AL 2012-13).

The SAS 2018-19 data show a decline in agricultural profitability overall, as well as a need to augment farm income with income from other sources. Given the apparent positive association between farm size and profitability per acre, it hints to a catastrophe for these farmers. Households in marginal farming earned up to Rs. 8,571, whereas large farms earned more than Rs. 60,000 per month.

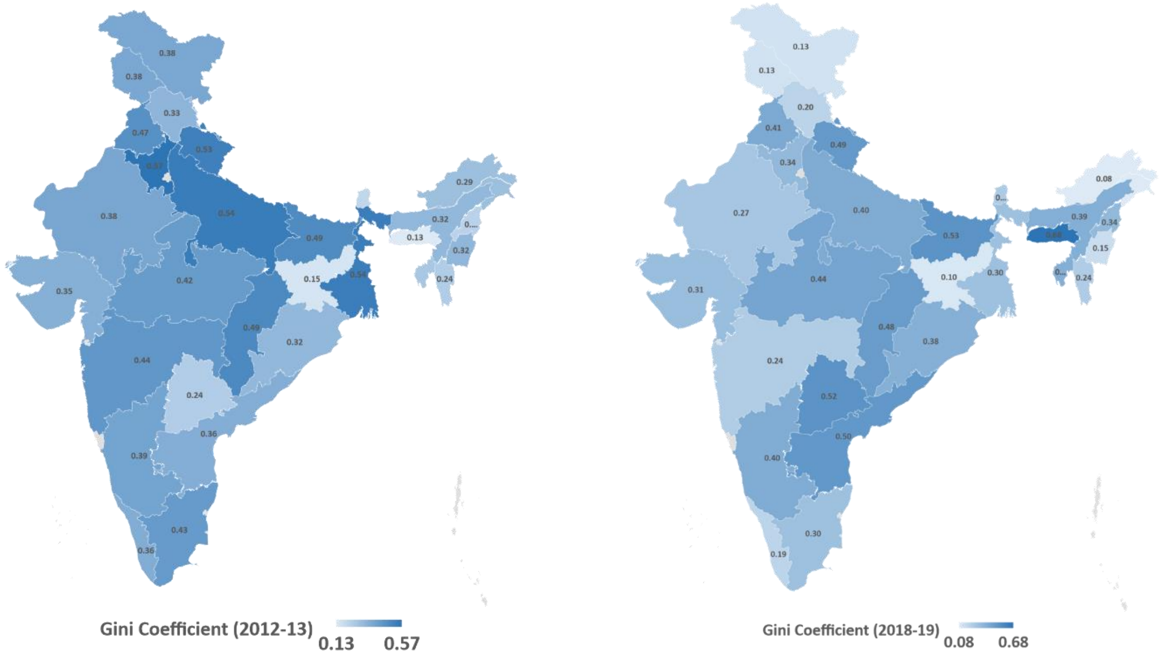
In order to assess the income gap or income inequality, we calculated Gini coefficient of income across size class of land at state and at national level. At national level, it was found that the Gini Coefficient decreased from 0.4523 in 2012-13 to 0.4049 in 2018-19, showing that the inequality in income across size class of land at national level decreased. The decrease in inequality can be substantiated from the fact that the real income growth has been higher for agricultural households possessing smaller lands compared to those possessing larger lands between the 70th and 77th round of SAS. Between 2012-13 to 2018-19, majority of the states witnessed decreased in inequality resembling the national level picture. But for states like Andhra Pradesh, Assam, Telangana, Bihar, etc the Gini coefficient saw an increase, showing the rise in income inequality within the states across size class of land (See Table 6 and Figure 2).

Table 6: State wise value of Gini coefficient of income across size class of land for AY 2012-13 and AY 2018-19

	AY 2012-13	AY 2018-19
Andhra Pradesh	0.36	0.50
Arunachal Pradesh	0.29	0.08
Assam	0.32	0.39
Bihar	0.49	0.53
Chhattisgarh	0.49	0.48
Gujarat	0.35	0.31
Haryana	0.57	0.34
Himachal Pradesh	0.33	0.20
Jammu & Kashmir	0.38	0.13
Jharkhand	0.15	0.10
Karnataka	0.39	0.40
Kerala	0.36	0.19
Madhya Pradesh	0.42	0.44
Maharashtra	0.44	0.24
Manipur	0.32	0.15
Meghalaya	0.13	0.68
Mizoram	0.24	0.24
Nagaland	0.22	0.34
Odisha	0.32	0.38
Punjab	0.47	0.41
Rajasthan	0.38	0.27
Sikkim	0.20	0.24
Tamil Nadu	0.43	0.30
Telangana	0.24	0.52
Tripura	0.26	0.39
Uttarakhand	0.53	0.49
Uttar Pradesh	0.54	0.40
West Bengal	0.54	0.30
All India	0.45	0.40

Source: Authors calculation from 70th and 77th rounds of SAS

Figure 2: - State-wise Gini Coefficient of Income across size class of land for AY 2012-13 and AY 2018-19 (Source: - NSO 70th and 77th round SAS)



All India(Gini Coeff)

2012-13: 0.4523

2018-19: 0.4049

Gini coefficient was also calculated across all states for different size class of land. It was found that Gini coefficient decreased for size class of land <0.01 ha, 0.01-0.40 ha, 0.41-1.00 ha, 1.00-2.00 ha and 2.01-4.00 ha, showing that the inequality among the members of these class between states decreased. For size class of land 4.01-10.00 ha and > 10.00+ ha gini coefficient reached 0.5078 and 0.5754 from 0.2761 and 0.3117 respectively. Table 7 shows that inequality in income between states have decreased for small landholder as compared to large landholders which also enables us to say that the situation of small landholders is more or less same across the country compared to large land holders.

Table 7: Gini Coefficient across size class of land at all India level

Size Class of Land	Gini Coefficient (2012-13)	Gini Coefficient (2018-19)
<0.01	0.2970	0.2910
0.01- 0.40	0.3370	0.2604
0.41-1.00	0.3806	0.2196
1.01-2.00	0.2136	0.2012
2.01-4.00	0.2669	0.2546
4.01-10.00	0.2761	0.5078
10.00+	0.3117	0.5754

Source: Authors calculation from NSO’s 77th round of SAS

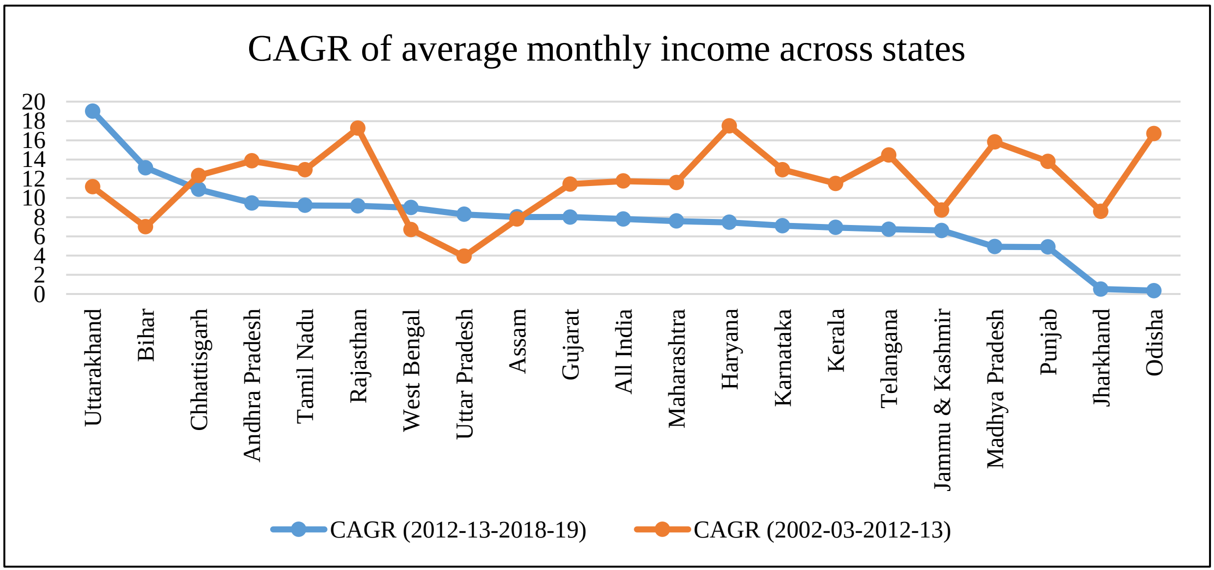
5.3. Income across states

The period between AY 2012-13 to AY 2018-19, witnessed non-uniform growth in average monthly income across the states. Between the 70th and 77th round, CAGR of average monthly income of Agricultural Households has slowed in most Indian states. Uttarakhand, Bihar, West Bengal, Uttar Pradesh and Assam were only exceptions, showing annual growth rates between 2012-13 and 2018-19 significantly higher than in 2002-03 and 2012-13.

Farmers' incomes in Odisha and Jharkhand grew at an impressive rate between 2002-03 and 2012-13, but after that, they have registered at slowest growth rate (Graph 5). Despite significant procurement of food grains at the minimum support price (MSP), incomes in Punjab and Madhya Pradesh grew at a slower rate between 2012-13 and 2018-19.

In terms of absolute value of income huge variation was seen across states, agricultural states like Bihar, Uttar Pradesh, Madhya Pradesh, Telangana, West Bengal and Chhattisgarh, saw incomes lagging behind the national average in 2018-19. Jharkhand and Odisha reported the lowest at Rs 4,895 and Rs 5,112 per month, respectively. Punjab and Haryana top in terms of average monthly farmer incomes among states, with income at Rs 26,701 and Rs 22,841, respectively. This came even though the growth rate of income of both the states slowed substantially during the period 2012-13 to 2018-19.

Only 12 states, three of which are from the North East, have an average monthly income (considering both the paid-out expenses and imputed expenses) of more over Rs 10,000, according to SAS estimates for 2018-19. The incomes of the remaining 16 states range between Rs. 4,013 and Rs. 9,995. Bihar, Jharkhand, Madhya Pradesh, Odisha, Telangana, Uttar Pradesh, and West Bengal, all of which being key agricultural states, have lower incomes than the national average.



Source: Authors calculation from 59th, 70th and 77th round of SAS

Graph 5: CAGR of average monthly income (nominal) of agricultural households statewise

While analysing the composition of state average monthly income during 2012-13 and 2018-19, income from agricultural activities (net receipts from cultivation and farming of animals) in total income have decreased significantly over the year. This pattern can be seen in most of the states. For Example, Crop income share more than 50% in the monthly income can be seen in only five states. Majority of the states (16 out of 28) have share of crop income less than 40% of total monthly income. The seriousness of the problem can be gauged from the fact that there are 9 states which have crop income less than 25% of monthly income. When compared with previous round, Madhya Pradesh which accounted for highest (76.5%) of income from agricultural activities in 2012-13 saw a dip to 67.5% in 2018-19, similarly Assam, Telangana, Haryana, Punjab and Uttar Pradesh saw a dip from 74.8%, 72.9%, 72.8%, 69.3%, and 69% in 2012-13 to 41%, 59.8%, 57.4%, 63.8% and 57.7% respectively in 2018-19. This clearly suggests that sustainability of crop cultivation is a serious problem across most of the states and immediate attention at both ground and policy level is the need of hour.

Income from leasing-out land the information of which was collected in this round was found not as significant for many states but for agricultural households in Punjab and Haryana, where monthly lease rent contribution equalled to Rs. 2,652 and about Rs. 621, respectively. The Agricultural Households in Arunachal Pradesh showed zero earnings from this source.

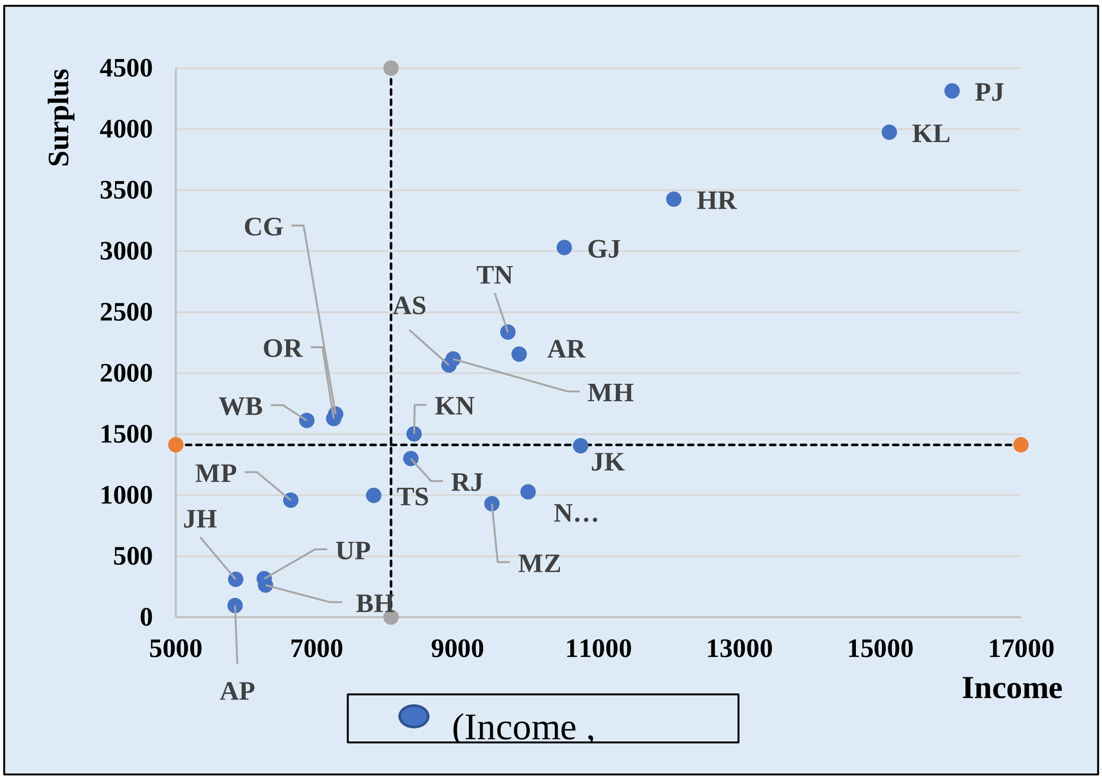
NAFIS provided a different dimension by providing information on consumption expenditure of states which helped us to better understand the economic status of the rural households. Based on the analysis of this data a quadrant graph (Graph 6) is drawn on Income vs Surplus where surplus is calculated as Income minus (–) Consumption expenditure. The first quadrant which is in the top right-hand side of the figure, shows states having income and surplus above national average (average national monthly income of all Rural Households: Rs. 8,059 and average national surplus: Rs. 1,413) while the third quadrant which is at the bottom left position of the graph shows states having both income and surplus below national average.

The graph shows, Punjab and Kerala at the positive extreme of the hierarchy with maximum reported surplus of roughly Rs. 4,000 per month as compared to states like Andhra Pradesh, Uttar Pradesh and Bihar showing monthly surplus less than Rs. 350 per household per month. The states such as Haryana, Himachal Pradesh and Gujarat have a sizeable amount of monthly surplus. NAFIS data also puts light on the average monthly income and consumption expenditure by size class of land possessed. Analysis of data suggests a positive correlation between size class of land possessed and surplus remaining after the monthly consumption expenditure has been subtracted from the income.

5.4. Income across social groups

Variation in income and its composition across various social group was seen during both the rounds of NSO's SAS. The Average Monthly Income (Rs) of SC, ST and OBC HH are less than that of HH belonging to category "Others" across all size class of land during both the survey round. In India, many surveys and studies have pointed out the unequal distribution of resources (i.e., land) across social groups. Though these reasons have remained at a core and is one of the driving factors leading to huge variation in income and its composition, but many recent studies have high lightened that the crop yield rates and agricultural land productivity differs across social groups when other things are kept constant. For both the rounds the average income of SC, ST and OBC were below national average and stood on an average 30% less than the income of category Others in both rounds (Table 8).

Close look in income composition shows SC, ST, OBC household more dependent on income from wages than Household belonging to category Others (Table 9). Households belonging to category Others earns 42% of their income from crop cultivation which is significantly higher from other categories 37.6(OBC), 25.2(SC), 34.4(ST). In terms of growth rate of income between the period 2012-13 to 2018-19, SC (10.22%) household growth rate was the highest followed by Others (8.02%), OBC (7.74%) and ST (7.36%). Livestock has income source, doesn't show much variation across social groups, in fact its share in total income for OBC households is more than category Others, while that for SC and ST households, its share is on similar line when compared to category Others.



Source: NAFIS 2015-16

Graph 6: - State-wise income *vs* surplus of rural households for the period 2015-16

Table 8: Income (in Rs) social group across different size class of land for 2012-13 and 2018-19 (Source: - NSO's 70th and 77th round of SAS)

	ST		SC		OBC		Others		Overall	
Farm-size, HA	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19
Landless (< 0.01)	6467	9451	4177	7840	4582	10611	3786	15865	4561	11204
Lower marginal (0.01 - 0.40)	4815	7487	3649	7177	4170	7127	4339	8675	4152	7522
Upper marginal (0.41 - 1.00)	4957	8030	4390	7559	5249	8573	6028	9704	5247	8571
Small (1.01 - 2.00)	6375	9336	6138	10182	7211	11338	8761	13706	7348	11449
Semi-medium (2.01 - 4.00)	8153	12214	7874	13307	10654	16733	12677	18573	10730	16435
Medium (4.01 - 10.00)	14270	23451	13074	23768	18904	22426	22384	38675	19637	28292
Large (>10.00)	100792	145517	24961	17763	35214	56205	46030	57700	41388	60758
All sizes	5864	8979	4539	8142	6378	9977	8059	12806	6426	10218
Income (CAGR)	7.35		10.22		7.74		8.02		8.03	

Table 9: Income composition of social groups for 2012-13 and 2018-19

Social Category	Wages/salaries		Crop cultivation		Livestock		Non-farm		Total	
	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19	2012-13	2018-19
ST	38.98	50.6	43.72	34.4	14.34	11.7	2.97	3.0	100	100
SC	50.89	53.0	32.52	25.2	10.13	14.0	6.43	7.1	100	100
OBC	29.54	36.9	48.82	37.7	12.78	17.8	8.86	6.7	100	100
OC	26.52	33.8	54.05	42.6	10.24	14.2	9.19	6.6	100	100
All	32.23	39.8	47.95	37.2	11.87	15.5	7.97	6.3	100	100

Source: - NSO's 70th and 77th round of SAS

6. Conclusions

This paper studied patterns and inequalities in credit access and farmers' income based on NSO 70th and 77th round surveys and the following broad conclusions emerge there from:

- Outstanding debt as percentage of annual income increases as size class of land possessed increases. Agri Household in the top category of size class of land are more leveraged than those in the bottom category.
- Proportion of loan for Consumption, Medical and Educational purposes have seen an increase across all the land size category especially at the lower level of landholders/land size. The proportion of loan for agricultural purposes increases as land size increases.
- The growth in average monthly income have not been uniform across the states in the period between AY 2012-13 to AY 2018-19. While inequality at All India level seems to have declined, certain surprises at state level are there. Some states have reduced inequality against our expectations. And certain others shocked us with increased inequality.

References

Government of India (GoI). (2016). *State of Indian Agriculture 2015-16*. Ministry of Agriculture and Farmers Welfare, New Delhi.

Gulati, A., Saini, S. and Roy, R. (2021). Going beyond agricultural GDP to farmers' incomes. In: Gulati, A., Roy, R., Saini, S. (eds) *Revitalizing Indian Agriculture and Boosting Farmer Incomes*. India Studies in Business and Economics. Springer, Singapore. https://doi.org/10.1007/978-981-15-9335-2_10.

Hussain, S. and Saini, S. (2021). Most Indian farmers faced erosion of their real incomes since 2012-13 but survey can't tell, https://theprint.in/opinion/most-indian-farmers-faced-erosion-of-their-real-incomes-since-2012-13-but-survey-cant-tell/749637/.

Hussain, Siraj and Saini, Shweta (2021). Most Indian farmers faced erosion of their real incomes since 2012-13 but survey can't tell, theprint.in.

Paliath, S. (2021). In 6 Years Pre-Covid, Average Farm Incomes Rose 59%, Debt 58%, IndiaSpend, https://www.indiaspend.com/agriculture/pre-covid-average-farm-incomes-rose-debt-777585.

Reserve Bank of India (RBI). (2019). Report of the *Internal Working Group to Review Agricultural Credit (2019).*

Saini, S., Gulati, A., von Braun, Joachim and Kornher, Lukas (2020). Indian farm wages: Trends, growth drivers and linkages with food prices, *ZEF Discussion Papers on Development Policy* **No. 301.**

Satyasai, K. J, Kumar, V., and Balanarayana, M. (2017). Do farm size and social group affiliation determine credit access and income of agricultural households? *Agricultural Economics Research Review*, **30** (Conference Number) 2017, 143-152**.**

Sharma, S. and Aggarwal, P. (2021). Income and debt account of India's farmers explained. https://www.indiatoday.in/diu/story/indian-agriculture-debt-data-msp-farmers-protest-1878975-2021-11-20.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 149-157

Perishable Stochastic Inventory Models for Two and Multiple Suppliers

Khimya Tinani[1] and Deepa Kandpal[2],
[1]*P.G. Department of Statistics, Faculty of Science,*
Sardar Patel University, Vallabh Vidyanagar, India

[2]*Department of Statistics, Faculty of Science,*
The Maharaja Sayajirao University of Baroda, Vadodara, India

Received: 26 May 2022; Revised: 27 July 2022; Accepted:01 August 2022

Abstract

In order to avoid the relevant losses due to deterioration of perishable items, we need an efficient and effective inventory management. This research work develops a perishable stochastic inventory models for two suppliers and multiple suppliers to determine an optimal ordering policy for allowable shortages. In case of two suppliers, spectral theory is used to derive explicit expression for the transition probabilities of a four-state continuous time Markov chain representing the status of the systems. These probabilities are used to compute the exact form of the average cost expression. We use concepts from renewal reward processes to develop average cost objective function. Optimal solution is obtained using Newton Rapson method in R programming. Finally, sensitivity analysis of the varying parameter on the optimal solution is done. We have extended the case of two suppliers to multiple suppliers and for the multiple suppliers problem, assuming that all the suppliers have similar availability characteristics; we develop a simple model and show that as the suppliers become large, the model reduces to classical EOQ model.

Key words: Future supply uncertainty; Two suppliers; Deteriorating items; EOQ model; Multiple suppliers, Sensitivity analysis.

1. Introduction

Inventory can be defined as the goods or stock hold by a person or a firm in order to use it in future for consumption, production or sale. Inventory management is used to minimize cost required to hold the inventory effectively in such way that there is no gap between demand and supply. The main and foremost reason for maintaining inventory level is to shorten the gap between demand and supply for the commodity under consideration. Any inventory system consists of an input process and output process. The input process refers to supply either by means of production or purchase while the output process refers to demand due to which depletion of inventory occurs. Thus, supply is a replenishment process, whereas demand is a depletion process. Though the inventories are essential and provide an alternative to production or purchase in future, they also mean lock up capital of an enterprise. Maintenance of inventories also costs money by way of expenses on stores, equipment, personnel, insurance *etc*. Thus, excess of inventories is undesirable. This calls for controlling the inventories in the most profitable way. Hence inventory theory deals with the determination of the optimal level of such ideal resources. Some products lose value faster than others, these are known as

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

perishable products. Perishable inventory forms a large portion of total inventory and include virtually all foodstuffs, pharmaceuticals, fashion goods, electronic items, digital goods (computer software, video games, DVD), periodicals (magazines/Newspapers), and many more as they lose value with time due to deterioration or obsolescence. Perishable goods can be broadly classified into two main categories based on: (i) Deterioration (ii) Obsolescence. Deterioration refers to damage, spoilage, vaporization, depletion, decay (*e.g.* radioactive substances), degradation (*e.g.* electronic components) and loss of potency (*e.g.* chemicals and pharmaceuticals) of goods. Obsolescence is loss of value of a product due to arrival of new and better product. Perishable goods have continuous or discrete loss of utility and therefore can have either fixed life or random life. Fixed life perishable products have a deterministic, known and definite shelf life and examples of such goods are pharmaceuticals, consumer packed goods and photographic films. On the other hand, random life perishable products have a shelf life that is not known in advance and variable depending on variety factors including storage atmosphere. Items are discarded when they spoil and the time to spoilage is uncertain. For example, fruits, vegetables, dairy products, bakery products *etc.*, have random life.

2. Review of Literature

An excellent survey on research in inventory management in a single product, single location inventory environment is provided by Lee and Nahmias (1993). A large number of researchers developed the models in the area of deteriorating inventories. At first Whitin (1957) considered an inventory model for fashion goods deteriorating at the end of a prescribed storage period. Various types of inventory models for items deteriorating at a constant rate were discussed by Chowdhury and Choudhuri (1983). A complete survey of the published literature in mathematical modelling of deteriorating inventory systems is given by Raafat (1991). Goyal and Giri (2001) developed recent trends of inventory models for deteriorating items. Teng and Chang(2005) determined economic production quantity in an inventory model for deteriorating items. Supply uncertainty can have a drastic impact on firms who fail to protect against it. Supply uncertainty has become a major topic in the field of inventory management in recent years. Supply disruptions can be caused by factors other than major catastrophes. More common incidents such as snow storms, customs delays, fires, strikes, slow shipments, *etc.* can halt production or transportation capability, causing lead time delays that disrupt material flow. Silver (1981) appears to be first author to discuss the need for models that deal with supplier uncertainty. Articles by Parlar and Berkin (1991) consider the supply uncertainty problem, for a class of EOQ model with a single supplier where the availability and unavailability periods constitute an alternating Poisson process. Parlar and Parry (1996) generalized the formulation of Parlar and Berkin (1991) by first assuming that the reorder point r is a non-negative decision variable instead of being equal to zero. Kandpal and Tinani (2009) developed inventory model for deteriorating items with future supply uncertainty under inflation and permissible delay in payment for single supplier. According to Yavari *et al.* (2020), one of the challenges when managing inventories is the inherent perishability of many items, which means their freshness and quality decrease over time and these cannot be sold after their expiration date.
Tirkolaee *et al.* (2017) noted that the inherent perishability widely occurs in food goods organisms and ornamental flowers. These authors also stated that the time window between preparation and sales of perishable items is very significant for producers and purchasers.

In this paper it is assumed that the inventory manager may place his order with any one of two suppliers who are randomly available. Here we assume that the decision maker deals with two suppliers who may be ON or OFF. Here there are three states that correspond to the availability of at least one supplier that is states 0, 1 and 2 whereas state 3 denotes the non-availability of either of them. State 0 indicates that supplier 1 and supplier 2 both are available. Here it is assumed that one may place order to either one of the two suppliers or partly to both. State 1 represents that supplier 1 is available but supplier 2 is not available. State 2 represents that supplier 1 is not available but supplier 2 is available.

3. Notations, Assumptions and Model

The inventory model here is developed on the basis of following assumptions.

a) Demand rate d is deterministic and it is d >1.
b) We define Xi and Yi be the random variables corresponding to the length of ON and OFF period respectively for i^{th} supplier where i =1, 2. We specifically assume that $X_i \sim exp\ (\lambda_i)$ and $Y_i \sim exp\ (\mu_i)$. Further Xi and Yi are independently distributed.
c) Ordering cost is Rs. k/order.
d) Holding cost is Rs. h/unit/unit time.
e) Shortage cost is Rs. π/unit.
f) θ is the rate of deterioration which is constant fraction of on hand inventory.
g) q_i = order up to level i =0, 1, 2.
h) r = reorder up to level; qi and r are decision variables.
i) Time dependent part of the backorder cost is Rs. $\hat{\pi}$/unit/time.
j) Purchase cost is Rs. c/unit.
k) T_{00} is the expected cycle time. T_{00} is a decision variable.

The policy we have chosen is denoted by (q_0, q_1, q_2, r). An order is placed for q_i units i = 0, 1, 2, whenever inventory drops to the reorder point r and the state found is i = 0, 1, 2. When both suppliers are available, q_0 is the total ordered from either one or both suppliers. If the process is found in state 3 that is both the suppliers are not available nothing can be ordered in which case the buffer stock of r units is reduced. If the process stays in state 3 for longer time, then the shortages start accumulating at rate of d units/time. When the process leaves state 3 and supplier becomes available, enough units are ordered to increase the inventory to q_i $+r$ units where i = 0, 1, 2. The cycle of this process start when the inventory goes up to a level of q_0+r units. Once the cycle is identified, we construct the average cost objective function as a ratio of the expected cost per cycle to the expected cycle length. *i.e.* $AC\ (q_0, q_1, q_2, r) = \frac{C_{00}}{T_{00}}$ where, C_{00} = E(cost per cycle) and T_{00} = E(length of a cycle). Analysis of the average cost function requires the exact determination of the transition probabilities $P_{ij}(t)$, i, j=0, 1, 2, 3 for the four state CTMC. The solution is provided in the lemma. (Refer Parlar and Perry [1996]). $A(q_i, r, \theta)$ = cost of ordering+ cost of holding inventory+ cost of items that deteriorate during a single interval that starts with an inventory of q_i units and ends with r units.

$$A(q_i\,,r,\theta) = k + \frac{1}{2}\frac{hq_i^2}{(d+\theta)} + \frac{hrq_i}{(d+\theta)} + \frac{\theta c q_i}{(d+\theta)} \qquad i = 0,\ 1,\ 2.$$

Lemma 3.1: Define C_{i0} = E(cost incurred to the beginning of the next cycle from the time when inventory drops to r at state i = 0, 1, 2, 3 and q_i units are ordered if i = 0, 1 or 2). Then,

C_{i0} is given by

$$C_{i0} = P_{i0}\left(\frac{q_i}{d+\theta}\right)A(q_i,r,\theta) + \sum_{j=1}^{3} P_{ij}\left(\frac{q_i}{d+\theta}\right)\left[A(q_i,r,\theta) + C_{j0}\right] \quad i=0,\ 1,\ 2. \tag{1}$$

$$C_{30} = \bar{C} + \sum_{i=1}^{2} \rho_i C_{i0} \qquad \text{where } \rho_i = \frac{\mu_i}{\delta} \text{ with } \delta = \mu_1 + \mu_2 \tag{2}$$

$$\bar{C} = \frac{e^{\frac{-\delta r}{(d+\theta)}}}{\delta^2}\left[he^{\frac{\delta r}{(d+\theta)}}(\delta r - (d+\theta)) + (\pi\delta d + h(d+\theta) + \hat{\pi}) - \theta c\delta\right] + \frac{\theta c}{\delta} \tag{3}$$

Proof: First consider *i* = 0. Conditioning on the state of the supplier availability process when inventory drops to r, we obtain

$$C_{00} = P_{00}\left(\frac{q_0}{d+\theta}\right)A(q_0, r, \theta) + \sum_{j=1}^{3} P_{0j}\left(\frac{q_0}{d+\theta}\right)\left[A(q_0, r, \theta) + C_{j0}\right] \tag{4}$$

The equation follows because $q_0 + r$ being the initial inventory, when q_0 units are used up we either observe state 0, 1, 2 or 3 with probabilities $P_{00}\left(\frac{q_0}{d+\theta}\right), P_{01}\left(\frac{q_0}{d+\theta}\right)$, $P_{02}\left(\frac{q_0}{d+\theta}\right)$ and $P_{03}\left(\frac{q_0}{d+\theta}\right)$ respectively. If we are in state 0 when r is reached, we must have incurred a cost of $A(q_0, r, \theta)$. On the other hand, if state *j* = 1, 2, 3 is observed when inventory drops to r, then the expected cost will be $A(q_0, r, \theta) + C_{j0}$ with probability $P_{0j}\left(\frac{q_0}{d+\theta}\right)$. The equation relating C_{10} and C_{20} are very similar but C_{30} is obtained as

$$C_{30} = [\bar{C} + C_{10}]\frac{\mu_1}{\mu_1+\mu_2} + [\bar{C} + C_{20}]\frac{\mu_2}{\mu_1+\mu_2} \tag{5}$$

Here, $\bar{C}$ is defined as the expected cost from the time inventory drops to r until either of the suppliers becomes available and it is computed as follows:
Now, note that the cost incurred from the time when inventory drops to r and the state is OFF to the beginning of next cycle is equal to

$$\begin{array}{ll} \frac{1}{2}hy^2(d+\theta) + hy[r - y(d+\theta)] + \theta cy & y < \frac{r}{d+\theta} \\ \frac{1}{2}\frac{hr^2}{(d+\theta)} + \pi\left(y - \frac{r}{(d+\theta)}\right)d + \frac{\hat{\pi}}{2}\left(y - \frac{r}{(d+\theta)}\right)^2 + \frac{\theta cr}{(d+\theta)} & y \geq \frac{r}{d+\theta} \end{array}$$

Hence,

$$\bar{C} = \int_0^{r/(d+\theta)} \left\{\frac{1}{2}hy^2(d+\theta) + hy(r - y(d+\theta) + \theta cy\right\}\delta e^{-\delta y}$$

$$+ \int_{r/(d+\theta)}^{\infty} \left\{\frac{1}{2}\frac{hr^2}{(d+\theta)} + \pi\left[y - \frac{r}{(d+\theta)}\right]d + \frac{\hat{\pi}}{2}\left[y - \frac{r}{(d+\theta)}\right]^2 + \frac{\theta cr}{(d+\theta)}\right\}\delta e^{-\delta y}$$

$$\bar{C} = \frac{e^{\frac{-\delta r}{(d+\theta)}}}{\delta^2}\left[he^{\frac{\delta r}{(d+\theta)}}(\delta r - (d+\theta)) + (\pi\delta d + h(d+\theta) + \hat{\pi}) - \theta c\delta\right] + \frac{\theta c}{\delta}$$

with $\delta = \mu_1 + \mu_2$ as the rate of departure from state 3. This follows because if supplier availability process is in state 3 (OFF for both suppliers) when inventory drops to r, then the expected holding and backorder costs are equal to $\bar{C}$. If the process makes a transition to state 1, the total expected cost would then be $\bar{C} + C_{10}$. The probability of a transition from state 3 to state 1 is $P(Y_1 < Y_2) = \int_0^\infty P(Y_1 < Y_2/Y_2 = t)\,\mu_2 e^{-\mu_2 t}dt = \frac{\mu_1}{\mu_1+\mu_2}$.
Multiplying this probability with the expected cost term above gives the first term of (5). The second term is obtained in a similar manner. Combining the results proves the lemma.

The following lemma provides a simpler means of expressing C_{00} in an exact manner. To simplify the notation, we let $A_i = A(q_i, r, \theta)$, *i* = 0, 1, 2 and $P_{ij} = P_{ij}\left(\frac{q_i}{d+\theta}\right)$ *i*, *j* = 0, 1, 2, 3.

Lemma 3.2: The exact expression for *C00* is

$$C_{00} = A_0 + P_{01}C_{10} + P_{02}C_{20} + P_{03}(\bar{C} + \rho_1 C_{10} + \rho_2 C_{20}) \tag{6}$$

where the pair [*C10, C20*] solves the system

$$\begin{bmatrix} 1 - P_{11} - P_{13}\rho_1 & -(P_{12} + P_{13}\rho_2) \\ -(P_{21} + P_{23}\rho_1) & (1 - P_{22} - P_{23}\rho_2) \end{bmatrix} \begin{bmatrix} C_{10} \\ C_{20} \end{bmatrix} = \begin{bmatrix} A_1 + P_{13}\bar{C} \\ A_2 + P_{23}\bar{C} \end{bmatrix} \tag{7}$$

Proof: Rearranging the linear system of four equations in lemma (3.1) in matrix form gives

$$\begin{bmatrix} 1 & -P_{01} & -P_{02} & -P_{03} \\ 0 & 1 - P_{11} & -P_{12} & -P_{13} \\ 0 & -P_{21} & 1 - P_{22} & -P_{23} \\ 0 & -\rho_1 & -\rho_2 & 1 \end{bmatrix} \begin{bmatrix} C_{00} \\ C_{10} \\ C_{20} \\ C_{30} \end{bmatrix} = \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \bar{C} \end{bmatrix} \tag{8}$$

We have $C_{30} = \bar{C} + \rho_1 C_{10} + \rho_2 C_{20}$ from the last row of the system. Substituting this result in rows two and three and rearranging gives the system in (7), with (*C10*, *C20*). From the first row of (8) we obtain $C_{00} = A_0 + \sum_{j=1}^{3} P_{0j}C_{j0}$.
Hence above lemma is proved.

Lemma 3.3: Define, *Ti0* = *E*[Time to the beginning of the next cycle from the time when inventory drops to *r* at state *i* = 0, 1, 2, 3 and *qi* units are ordered if *i* = 0, 1, 2]. Then, expected cycle length is given by

$$T_{i0} = P_{i0}\left(\frac{q_i}{d+\theta}\right)\frac{q_i}{d+\theta} + \sum_{j=1}^{3} P_{ij}\left(\frac{q_i}{d+\theta}\right)\left[\frac{q_i}{d+\theta} + T_{j0}\right] \qquad i = 0, 1, 2.$$

$$T_{30} = \bar{T} + \sum_{j=1}^{2} \rho_i T_{i0}$$

where $\bar{T} = \frac{1}{\mu_1 + \mu_2}$ is the expected time from the time inventory drops to r until either supplier 1 or 2 becomes available.

Lemma 3.4: The exact expression for *T 00* is

$$T_{00} = \frac{q_0}{d+\theta} + P_{01}T_{10} + P_{02}T_{20} + P_{03}(\bar{T} + \rho_1 T_{10} + \rho_2 T_{20})$$

where the pair *[T10, T20]* solves the system.

$$\begin{bmatrix} 1 - P_{11} - P_{13}\rho_1 & -(P_{12} + P_{13}\rho_2) \\ -(P_{21} + P_{23}\rho_1) & (1 - P_{22} - P_{23}\rho_2) \end{bmatrix} \begin{bmatrix} T_{10} \\ T_{20} \end{bmatrix} = \begin{bmatrix} q_1 + P_{13}\bar{T} \\ q_2 + P_{23}\bar{T} \end{bmatrix}$$

The proof of the above two lemmas (3.3) and (3.4) are very similar to lemma (3.1) and (3.2).

Theorem 3.5: The Average cost objective function for deteriorating items in case of two suppliers is given by $AC = \frac{C_{00}}{T_{00}}$,

$$AC = \frac{C_{00}}{T_{00}} = \frac{A(q_0\,, r, \theta) + P_{01}C_{10} + P_{02}C_{20} + P_{03}(\bar{C} + \rho_1 C_{10} + \rho_2 C_{20})}{\frac{q_0}{d+\theta} + P_{01}T_{10} + P_{02}T_{20} + P_{03}(\bar{T} + \rho_1 T_{10} + \rho_2 T_{20})}$$

Proof: Proof follows using Renewal reward theorem (RRT). The optimal solution for q_0, q_1, q_2 and r is obtained by using Newton Rapson method in R programming.

4. Numerical Analysis

Here, we assume that k = Rs. 5/order, c = Rs.5/unit, d = 20/units, θ = 5, h = Rs. 5/unit/time, π = Rs. 250/unit, $\hat{\pi}$ = Rs.25/unit/time, $\lambda_1 = 0.25$, $\lambda_2 = 1$, $\mu_1 = 2.5$, $\mu_2 = 0.5$.
With these parameters the long run probabilities are obtained as $p_0 = 0.303$, $p_1 = 0.606$, $p_2 = 0.030$ and $p_3 = 0.061$. The optimal solution is obtained as
$q_0 = 1.86448$, $q_1 = 10.490$, $q_2 = 15.44333$, $r = 20.4988$ and $AC = \frac{C_{00}}{T_{00}} = 197.81$.

5. Sensitivity Analysis

(i) To observe the effect of varying parameter values on the optimal solution, we have conducted sensitivity analysis by varying the value μ_1 and keeping other parameter values fixed. We resolve the problem to find optimal values of q_0, q_1, q_2, r and AC. The optimal values of q_0, q_1, q_2, r and AC are shown in Table1.

Table 1: Sensitivity Analysis by varying the parameter values of µ1

μ_1	q_0	q_1	q_2	r	AC
2.4	0.7827	10.6468	15.3941	20.4989	234.57
2.5	1.86448	10.4905	15.4433	20.4988	197.81
2.6	3.2545	10.3081	15.497	20.4987	186.73
2.7	5.03811	10.0839	15.5592	20.4986	181.83
2.8	7.3501	9.7962	15.6347	20.4984	179.52

We see that as μ_1 increases *i.e.*, expected length of OFF period for 1st supplier decreases the value of q_0, q_2 increases, q_1 decreases and r remain almost constant which result in decrease in average cost.

(ii) To observe the effect of varying parameter values on the optimal solution, we have conducted sensitivity analysis by varying the value λ_1 and keeping other parameter values fixed. We resolve the problem to find optimal values of q_0, q_1, q_2, r and AC. The optimal values of q_0, q_1, q_2, r and AC are shown in Table 2.

Table 2: Sensitivity Analysis by varying the parameter values of λ_1

λ_1	q_0	q_1	q_2	r	AC
0.25	1.86448	10.4905	15.4433	20.4988	197.81
0.28	2.8761	10.3761	15.6391	20.4987	199.95
0.3	3.5782	10.3275	15.7820	20.4987	205.17
0.35	4.1852	10.2792	15.8861	20.4986	213.63
0.37	5.8807	9.8391	15.9840	20.4986	224.87

We see that as λ_1 increase *i.e.*, expected length of ON period for 1st supplier decreases the value of q_0, q_2 increases, q_1 decreases and r remain almost constant which result in increase in average cost.

(iii) To observe the effect of varying parameter values on the optimal solution, we have conducted sensitivity analysis by varying the value θ and keeping other parameter values fixed. We resolve the problem to find optimal values of q_0, q_1, q_2, r and AC. The optimal values of q_0, q_1, q_2, r and AC are shown in Table 3.

Table 3: Sensitivity Analysis by varying the parameter values of θ

θ	q_0	q_1	q_2	r	AC
3	16.7227	8.5525	16.0268	20.4979	169.89
3.5	9.4464	9.4878	15.7431	20.4982	170.74
4	5.5986	9.8872	15.5901	20.4985	175.58
4.5	3.3168	10.2917	15.5017	20.4986	183.73
5	1.86448	10.4905	15.4433	20.4988	197.81

We see that as θ increases it results in decrease in q_0, increase in q_1 but q_2 decreases and r remains almost constant. This results in increase in average cost.

(iv) To observe the effect of varying parameter values on the optimal solution, we have conducted sensitivity analysis by varying the value of holding cost h and keeping other parameter values fixed. We resolve the problem to find optimal values of q_0, q_1, q_2, r and AC. The optimal values of q_0, q_1, q_2, r and AC are shown in Table 4.

Table 4: Sensitivity analysis by varying the parameter values of *h*

h	q_0	q_1	q_2	r	AC
4	2.0571	10.469	15.4496	20.4985	170.55
5	1.86448	10.4905	15.4433	20.4988	197.81
6	1.67464	10.5106	15.4374	20.499	225.58
7	1.48785	10.5295	15.4318	20.4992	254.01
8	1.30409	10.5471	15.4267	20.4993	283.38

We see that as h increases it results in decrease in q_0, increase in q_1 but q_2 decreases and r remains almost constant. This results in increase in average cost.

6. Multiple Suppliers

We have generalized the model and consider the case where there are M suppliers, and at any time suppliers may be available or not available which we represent as ON or OFF state. The stochastic process representing the supplier availabilities would have 2^M states:
0, 1, 2, ...,2^M-1. State 0 would correspond to the situation where all the suppliers being ON, state 1 would correspond to only the M^{th} supplier being OFF *etc.* and finally state 2^M-1 would correspond to all being OFF. The transition probabilities $P_{ij}(t)$, $i, j = 0, 1, 2, ...,2^M$-1, decision variables q_i and costs C_{i0}, $i = 0, 1, 2, ...,2^M$-1 are defined in a manner similar to two suppliers. The system of equations for C_{i0} is obtained as

$$C_{i0} = P_{i0}\left(\frac{q_i}{d+\theta}\right)A(q_i,r,\theta) + \sum_{j=1}^{2^M-1} P_{ij}\left(\frac{q_i}{d+\theta}\right)\left[A(q_i,r,\theta) + C_{j0}\right], \quad i = 0,1,\ldots,2^M-2\,.$$

$C_{2^M-1,0} = \overline{C} + \sum_{i=1}^{M} \rho_i \, C_{i0}$ where, $\rho_i = \frac{\mu_i}{\sum_{j=1}^{M} \mu_j}$

$$\bar{C} = \frac{e^{\frac{-\delta r}{(d+\theta)}}}{\delta^2}\left[he^{\frac{\delta r}{(d+\theta)}}(\delta r - (d+\theta)) + (\pi\delta d + h(d+\theta) + \hat{\pi}) - \theta c\delta\right] + \frac{\theta c}{\delta}, \quad \delta = \sum_{j=i}^{M} \mu_j$$

Equations for T_{i0} are written in a similar way as in Lemma (3.3).

Solving the above equations require the exact solution for the transient probabilities $P_{ij}(t)$ of the CTMC with the 2^M states which appears to be a formidable task, because we would first need the exact solution for the transient probabilities $P_{ij}(t)$ of the CTMC with the 2^M states. It would also be necessary to solve explicitly for the quantities C_{00} and T_{00} using the system of 2^M equations in 2^M unknowns. As the number of suppliers is very large, that is we have a situation approximating a free market, we can develop a much simpler model by assuming that if an order needs to be placed and at least one of the suppliers is available, then the order quantity will be q units regardless of which supplier is available. We combine the first 2^M-1 states where at least one supplier is available and define a super state denoted by o. The last state denoted by I, is the state where all the suppliers are OFF. We also assume that for any supplier the ON and OFF periods are exponential with parameters λ and μ, respectively. With these assumptions the expected cost and the expected length of a cycle are obtained as

$$C_{00} = A(q, r, \theta) + P_{0\,\mathrm{I}}\left(\frac{q}{d+\theta}\right) C_{\mathrm{I}0}(r)$$

$$T_{00} = \frac{q}{(d+\theta)} + \frac{P_{0\mathrm{I}}\left(\frac{q}{d+\theta}\right)}{M_\mu}$$

Therefore, the average cost function is given by

$$AC = \frac{C_{00}}{T_{00}}$$

where, A(q, r, θ), $P_{0\,I}\left(\frac{q}{(d+\theta)}\right)$ and $C\,Io(r)$ have the same meaning as in single supplier case.

Thus, when the number of suppliers become large, the objective function of multiple suppliers problem reduces to that of classical EOQ model. This can be shown by arguing that as the length of stay in state I is exponential with parameter M_μ it becomes a degenerate random variable with mass at 0; that is the process never visits or stays in state I.

7. Discussion and Conclusions

In this paper, we have analysed order quantity and reorder point of perishable stochastic inventory models with two and multiple suppliers. We have assumed that suppliers may be ON available or not available (OFF) at a given time and duration of these periods are exponential with specified parameters. Using concepts from renewal reward processes we have constructed the average cost objective function for the case of two and multiple suppliers. When the number of suppliers become large, the objective function of multiple suppliers problem reduces to that of classical EOQ model.

Acknowledgements

The authors express their gratefulness to the anonymous reviewer for giving some useful suggestions that led to considerable improvement in the presentation of the results. The authors also express their thanks to the Chief Editor for his suggestions on restructuring the contents for better expression.

References

Kandpal, D. H. and Tinani, K. S. (2009). Future supply uncertainty model for deteriorating items under inflation and permissible delay in payment for single supplier. *Journal of Probability and Statistical Science*, **7**(**2**), 245-259.

Lee, H. L. and Nahmias, S. (1993). Single product, single location model. In S.C. Graves. *Handbooks in Operations Research and Management Science,* **4**, 3-55.

Parlar, M. and Berkin, D. (1991). Future supply uncertainty in EOQ models. *Naval Research Logistics*, **38**, 295-303.

Parlar, M. and Perry, D. (1996). Inventory models of future supply uncertainty with single and multiple suppliers. *Naval Research Logistic*, **43**, 191-210.

Raafat, F. (1991). Survey of literature on continuously deteriorating inventory model. *Journal of Operational Research Society,* **42**, 27-37.

Roy Chowdhury, M. and Chaudhuri, K. S. (1983). An order level inventory mode for deteriorating items with finite rate of replenishment. *Opsearch*, **20**, 99-106.

Goyal, S. K. and Giri, B. C. (2001). Recent trends in modelling of deteriorating inventory. *European Journal of Operational Research,* **134**, 1-16.

Silver, E. A. (1981). Operations research inventory management: A review and critique, *Operations Research*, **29**, 628-64.

Teng, J. T. and Chang, C. T. (2005). Economic production quantity models for deteriorating items with price and stock dependent demand. *Computers and Operational Research,* **32**, 297-308.

Tirkolaee, E. B., Goli, A., Bakhsi, M. and Mahdavi, I. (2017). A robust multi-trip vehicle routing problem of perishable products with intermediate depots and time windows. *Numerical Algebra, Control and Optimization,* **7**(**4**), 417–433.

Whitin, T. M. (1957). Theory of inventory management. *Princeton University Press, Princeton.*

Yavari, M., Enjavi, H. and Geraeli, M. (2020). Demand management to cope with routes disruptions in location-inventory-routing problem for perishable products. *Research in Transportation Business and Management,* **37**, 1–14.

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp:159-169

Change Point Detection using Kumaraswamy Power Lomax Distribution

R. Vishnu Vardhan[1] **and Vasili. B. V. Nagarjuna**[2]
[1]*Department of Statistics, Pondicherry University, Puducherry.*
[2]*Department of Mathematics, Vellore Institute of Technology, Andhra Pradesh.*

Received: 25 July 2022 ; Revised: 04 August 2022 ; Accepted: 07 August 2022

Abstract

The Kumaraswamy Power Lomax distribution is an extension of Power Lomax distribution which can be applied many fields engineering, finance and medical research. In this paper, we study a change point problem of this distribution. A procedure based on Modified Information Criterion (MIC) is proposed to detect change point(s) in parameters of this distribution through binary segmentation. The practical applications are provided to illustrate the detection of multiple change points.

Key words: Kumaraswamy power lomax; Change point; Information criterion; Binary segmentation.

AMS Subject Classifications: 92B15, 62P10

1. Introduction

Lomax distribution (Lomax, 1954) is one of the more versatile forms of the Pareto distributional forms. The extension works of Lomax distribution are carried out by many researchers like Kumaraswamy-generalized Lomax distribution (Shams, 2013), type II Topp-Leone Power Lomax (TIITLPL) distribution (Al-Marzouki, Jamal, Chesneau and Elgarhy, 2020), Marshall-Olkin exponential Lomax distribution (Nagarjuna VBV and Vishnu Vardhan, 2020), and Sine Power Lomax (Nagarjuna, Vardhan and Chesneau, 2021b) to utilizing generalized family of distributions by adding additional parameters to the model.

In generalized family of distributions, Kumaraswamy generalized family of distributions is a well know family and has been utilized by several researchers to come out with new functional forms. To mention a few, there are the Kumaraswamy-Weibull distribution (Cordeiro, Ortega and Nadarajah, 2010), Kumaraswamy-Burr XII (KBXII) distribution (Paranaiba, Ortega, Cordeiro and Pascoa, 2013) and Kumaraswamy generalized Power Lomax distribution(KPL)(Nagarjuna, Vardhan and Chesneau, 2021a).

Recently, the attractive properties of Power Lomax distributions and its mathematical tractability was presented by Nagarjuna et al. (2021a). The cumulative distribution

Correponding Author: R. Vishnu Vardhan
Email:vrstatsguru@gmail.com

function (cdf) and probability density function (pdf) of KPL distribution are

$$F_{KPL}(x;\xi) = 1 - \left\{1 - \left[1 - \left(\frac{\lambda}{\lambda + x^{\beta}}\right)^{\alpha}\right]^{a}\right\}^{b}, \quad x > 0, \tag{1}$$

where $\xi = (\alpha, \beta, \lambda, a, b) > 0$, and

$$f_{KPL}(x;\xi) = \frac{ab\alpha\beta}{\lambda} x^{\beta-1} \left(\frac{\lambda}{\lambda + x^{\beta}}\right)^{\alpha+1} \left[1 - \left(\frac{\lambda}{\lambda + x^{\beta}}\right)^{\alpha}\right]^{a-1} \left\{1 - \left[1 - \left(\frac{\lambda}{\lambda + x^{\beta}}\right)^{\alpha}\right]^{a}\right\}^{b-1}. \tag{2}$$

The different shapes of the KPL distribution has been observed at several parameters of the distribution and very nicely depicted in Figure (1). From this Figure (1), we can observe that the density curves of the KPL distributional are decreasing or uni-modal shapes with very flexible to the skewness too.

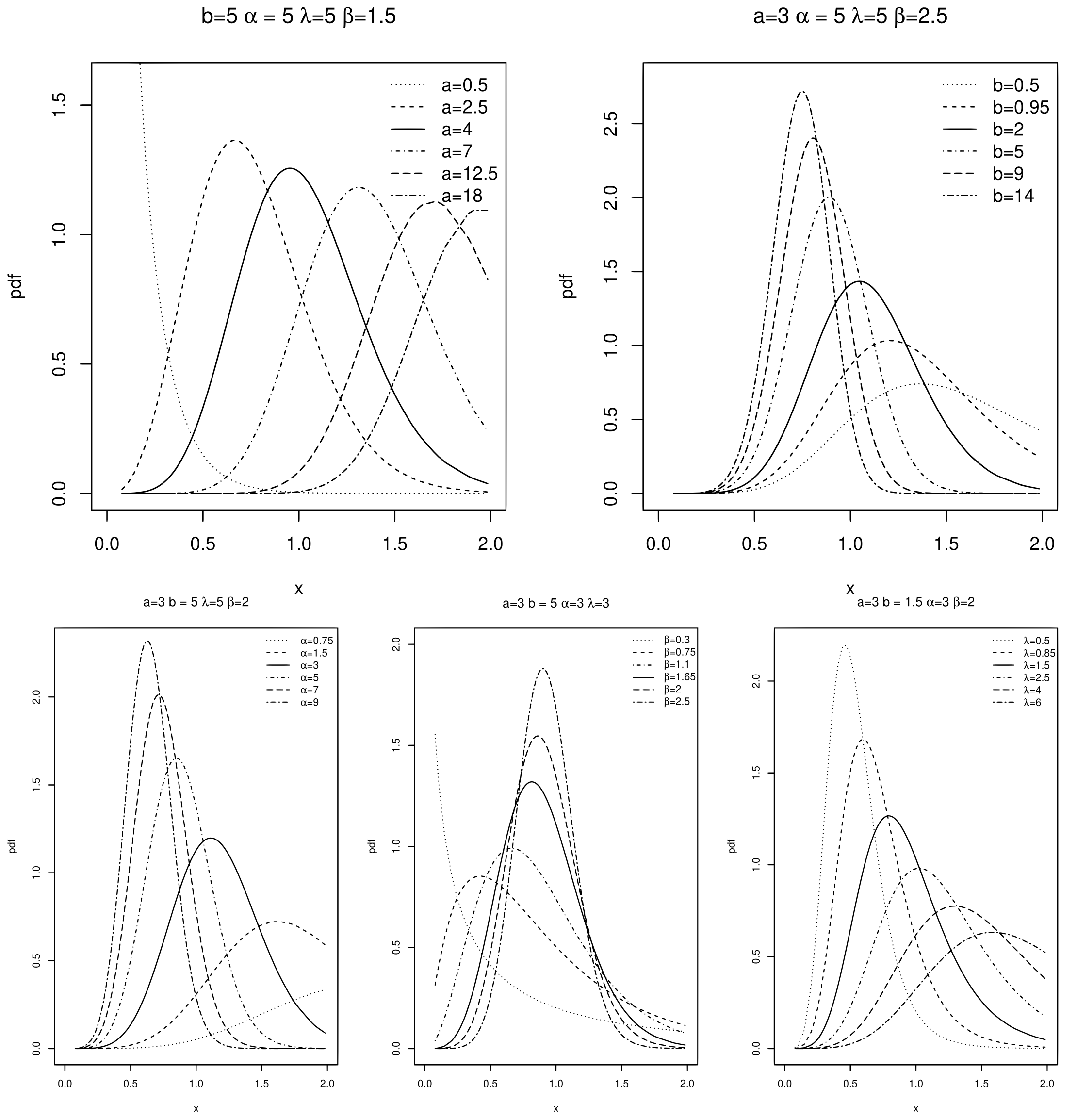


Figure 1: Density curves of KPL distribution at different parameter values

Now-a-days, the study of Change Point (CP) problem is has grabbed the attention of many researchers from industry, weather, quality control etc. Briefly, the CP problem is the problem to study a change or changes in data. A change in data represents the point before which the data follows a distribution and follows a different distribution after that point. Initial works in CP detection was by Page (1954, 1955), where the methods for the detection of single and multiple points were addressed.

In general, the change point problem involves two steps: estimation and hypothesis testing. In hypothesis testing step, we test the null hypothesis of no change versus the alternative of at least one change in data. If the null hypothesis is rejected, we move from the estimation step to estimate the location of change. Otherwise we stop and conclude there is no change in data. To know about the theoritical developments and applications of change point problems, readers can refer to the work of Chen and Gupta (2011). and applications in this field.

2. Change Point Methodology

Let $x_1, x_2, \ldots, x_n$ be a sequence of independent observations that follows a particular distribution. We would like to test the null hypothesis (H_0) versus the alternative hypothesis (H_1), which refers in testing the presence of atmost one change point in the data. For this problem, the binary segmentation procedure and MIC are used to search for all possible change points.
We define the H_0 and H_1 as

$$H_0 : \xi_1 = \xi_2 = \ldots = \xi_n \ vs \ H_1 : \xi_1 = \xi_2 = \ldots = \xi_k \neq \xi_{k+1} = \ldots = \xi_n$$

here $\xi_i; i = 1, 2, \ldots n$ is a parameter set of a particular distribution and 'k' is the position of the change point. If a change point is detected then we choose H_1, otherwise. Let us assume that 'k' is the change point location and at this point the data gets divided into segments, *first segment* will be from 1 to k^{th} point and the *second segment* will start from $(k+1)^{th}$ and n.

Generally, for finding a change point problem there are two popular methods such as, likelihood ratio test (LRT) and Bayesian procedures. Apart from these methods, researchers have also shown interest in the well known for model selection that is Akaike Information Criterion (AIC) and Bayes or Schwarz Information Criterion (BIC).

Chen and Gupta (1997) proposed a test to locate the change in variances of the normal distribution using BIC. Later, Chen et al. (2006) pointed out, that the BIC do not concentrate more on penalty term and it needs some modifications related to the concepts of change point problems. They proposed a new information criterion and named it as the Modified Information Criterion (MIC) which is the modification of the approach based on BIC. This is done for refining the model complexity as a function of the change location in the context of change point problem.

3. Binary Segmentation Procedure

Vostrikova (1981) developed the binary segmentation procedure which was shown to be consistent. Such a procedure transfers the detection of multiple changes to a sequence of consecutive steps of at most one change in each step. Before starting the procedure, at each iteration we need to test for the goodness of fit of data to a particular distribution. The steps involved in binary segmentation procedure are as follows:

1. Under the H_0 and H_1, the general form of log-likelihood functions are given by $\log L(X, \xi)$ and $\log L(X, \xi', \xi'')$. Here $X \sim f_x(.)$; ξ' is the parameter set of first segment and ξ'' is the parameter set of second segment.
2. Compute the MIC values under H_0 and H_1. The general expressions of MIC under H_0 and H_1 are
$$H_0 : \quad MIC_n = -2\,\log L_{H_0} + m\,\log n$$
$$H_1 : \quad MIC_k = -2\,\log L_{H_1} + \left(2m + \left(\frac{2m}{n} - 1\right)^2\right) \log n$$
here 'm' is the number of parameters; 'n' is the number of observations at each iteration.
3. If $MIC_n < \underset{1 \le k < n}{min}\, MIC_k$, then we accept H_0, i.e., there is no change point in the data, otherwise we accept H_1 meaning to that there exits a change point. At this change point location, the dataset divide into two segments.
4. This will continue until the condition given in step-3 is satisfied, i.e., H_0 is accepted.

As per the steps of the binary segmentation procedure, the log-likelihood functions and MIC are presented below.

Step-1: Let us define the H_0 and H_1 based on the parameter set of KPL distribution.

$$H_0 : \xi_1 = \xi_2 = \ldots = \xi_n \ vs \ H_1 : \xi_1 = \xi_2 = \ldots = \xi_k \neq \xi_{k+1} = \ldots = \xi_n$$

here $\xi_i = (a_i, \, b_i, \, \alpha_i, \beta_i, \lambda_i); i = 1, 2, \ldots n$ which are the parameters of KPL distribution.

Under H_0 and H_1, the log-likelihood functions of KPL distributions are

$$\log L_{H_0}(x;\xi) = n \log\left(\frac{ab\alpha\beta}{\lambda}\right) + (\beta - 1)\sum_{i=1}^{n} \log x_i + (\alpha + 1)\sum_{i=1}^{n} \log\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right) + (a - 1)\sum_{i=1}^{n} \log \Upsilon_i + (b - 1)\sum_{i=1}^{n} \log(1 - \Upsilon_i^{a})$$

$$\begin{aligned}\log L_{H_1}(x;\xi',\xi'') &= k \log\left(\frac{a'b'\alpha'\beta'}{\lambda'}\right) + (\beta' - 1)\sum_{i=1}^{n} \log x_i + (\alpha' + 1)\sum_{i=1}^{k} \log\left(\frac{\lambda'}{\lambda' + x_i^{\beta'}}\right) \\ &+ (a' - 1)\sum_{i=1}^{k} \log \Upsilon_i + (b' - 1)\sum_{i=1}^{k} \log(1 - \Upsilon_i^{a'}) \\ &+ (n - k) \log\left(\frac{a''b''\alpha''\beta''}{\lambda''}\right) + (\beta'' - 1)\sum_{i=k+1}^{n} \log x_i + (\alpha'' + 1)\sum_{i=k+1}^{n} \log\left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right) \\ &+ (a'' - 1)\sum_{i=k+1}^{n} \log \Upsilon_i + (b'' - 1)\sum_{i=k+1}^{n} \log(1 - \Upsilon_i^{a''})\end{aligned}$$

where it is set $\Upsilon_i = 1 - \left[\lambda/(\lambda + x_i^{\beta})\right]^{\alpha}$.

The maximum likelihood estimators (MLEs) of KPL distribution for H_0 and H_1 are

$$\frac{\partial}{\partial a} \log L_{H_0} = \frac{n}{a} + \sum_{i=1}^{n} \log \Upsilon_i + \sum_{i=1}^{n} (1 - b)\frac{\log \Upsilon_i}{(\Upsilon_i^{-a} - 1)} = 0$$

$$\frac{\partial}{\partial b}\log L_{H_0} = \frac{n}{b} + \sum_{i=1}^{n}\log(1-\Upsilon_i^{a}) = 0$$

$$\frac{\partial}{\partial \alpha}\log L_{H_0} = \frac{n}{\alpha} + \sum_{i=1}^{n}\log\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right) + \sum_{i=1}^{n}\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right)^{\alpha}\log\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right)$$
$$\left[\frac{a(b-1)}{\Upsilon_i(\Upsilon_i^{-a}-1)} - \frac{(a-1)}{\Upsilon_i}\right] = 0$$

$$\frac{\partial}{\partial \beta}\log L_{H_0} = \frac{n}{\beta} + \sum_{i=1}^{n}\log x_i - (\alpha+1)\sum_{i=1}^{n}\left(\frac{x_i^{\beta}}{\lambda + x_i^{\beta}}\right)\left[\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right)^{\alpha-1}\frac{\lambda x_i^{\beta}\log x_i}{(\lambda + x_i^{\beta})^2}\right.$$
$$\left[\frac{\alpha(a-1)}{\Upsilon_i} - \frac{a\alpha(b-1)}{\Upsilon_i(\Upsilon_i^{-a}-1)}\right]\log x_i = 0$$

$$\frac{\partial}{\partial \lambda}\log L_{H_0} = -\frac{n}{\lambda} + \left(\frac{\alpha+1}{\lambda}\right)\sum_{i=1}^{n}\left(\frac{x_i^{\beta}}{\lambda + x_i^{\beta}}\right) + \sum_{i=1}^{n}x_i^{\beta}\left(\frac{\lambda}{\lambda + x_i^{\beta}}\right)^{\alpha+1}$$
$$\left[\frac{a\alpha(b-1)}{\lambda^2\Upsilon_i(\Upsilon_i^{-a}-1)} - \frac{\alpha(a-1)}{\lambda^2\Upsilon_i}\right] = 0$$

$$\frac{\partial}{\partial a'}\log L_{H_1} = \frac{k}{a'} + \sum_{i=1}^{k}\log\Upsilon_i' + \sum_{i=1}^{k}(1-b')\frac{\log\Upsilon_i'}{(\Upsilon_i'^{-a'}-1)} = 0$$
$$\frac{\partial}{\partial b'}\log L_{H_1} = \frac{k}{b'} + \sum_{i=1}^{k}\log(1-\Upsilon_i'^{a'}) = 0,$$

$$\frac{\partial}{\partial \alpha'}\log L_{H_1} = \frac{k}{\alpha'} + \sum_{i=1}^{k}\log\left(\frac{\lambda'}{\lambda' + x_i^{\beta'}}\right) + \sum_{i=1}^{k}\left(\frac{\lambda'}{\lambda' + x_i^{\beta'}}\right)^{\alpha'}\log\left(\frac{\lambda'}{\lambda' + x_i^{\beta'}}\right)$$
$$\left[\frac{a'(b'-1)}{\Upsilon_i'(\Upsilon_i'^{-a'}-1)} - \frac{(a'-1)}{\Upsilon_i'}\right] = 0$$

$$\frac{\partial}{\partial \beta'}\log L_{H_1} = \frac{k}{\beta'} + \sum_{i=1}^{k}\log x_i - (\alpha'+1)\sum_{i=1}^{k}\left(\frac{x_i^{\beta'}}{\lambda' + x_i^{\beta'}}\right)\left[\left(\frac{\lambda'}{\lambda' + x_i^{\beta'}}\right)^{\alpha'-1}\frac{\lambda' x_i^{\beta'}\log x_i}{(\lambda' + x_i^{\beta'})^2}\right]$$
$$\left[\frac{\alpha'(a'-1)}{\Upsilon_i'} - \frac{a'\alpha'(b'-1)}{\Upsilon_i'(\Upsilon_i'^{-a'}-1)}\right]\log(x_i) = 0$$

$$\frac{\partial}{\partial \lambda'}\log L_{H_1} = -\frac{k}{\lambda'} + \left(\frac{\alpha'+1}{\lambda'}\right)\sum_{i=1}^{k}\left(\frac{x_i^{\beta'}}{\lambda' + x_i^{\beta'}}\right) + \sum_{i=1}^{k}x_i^{\beta'}\left(\frac{\lambda'}{\lambda' + x_i^{\beta}{}'}\right)^{\alpha'+1}$$
$$\left[\frac{a'\alpha'(b'-1)}{\lambda'^2\Upsilon_i'(\Upsilon_i'^{-a'}-1)} - \frac{\alpha(a'-1)}{\lambda'^2\Upsilon_i'}\right] = 0$$

where $\Upsilon_i' = 1 - \left[\lambda'/(\lambda' + x_i^{\beta'})\right]^{\alpha'}$ and

$$\frac{\partial}{\partial a''} \log L_{H_1} = \frac{(n-k)}{a''} + \sum_{i=k+1}^{n} \log \Upsilon_i'' + \sum_{i=k+1}^{n} (1-b'') \frac{\log \Upsilon_i''}{(\Upsilon_i''^{-a''} - 1)} = 0,$$

$$\frac{\partial}{\partial b''} \log L_{H_1} = \frac{(n-k)}{b''} + \sum_{i=k+1}^{n} \log(1 - \Upsilon_i''^{a''}) = 0$$

$$\frac{\partial}{\partial \alpha''} \log L_{H_1} = \frac{(n-k)}{\alpha''} + \sum_{i=k+1}^{n} \log\left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right) + \sum_{i=1}^{k} \left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right)^{\alpha''} \log\left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right) \left[\frac{a''(b''-1)}{\Upsilon_i''(\Upsilon_i''^{-a''} - 1)} - \frac{(a''-1)}{\Upsilon_i''}\right] = 0,$$

$$\frac{\partial}{\partial \beta''} \log L_{H_1} = \frac{(n-k)}{\beta''} + \sum_{i=k+1}^{n} \log x_i - (\alpha''+1) \sum_{i=k+1}^{n} \left(\frac{x_i^{\beta''}}{\lambda'' + x_i^{\beta''}}\right) \left[\left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right)^{\alpha''-1} \frac{\lambda'' x_i^{\beta''} \log x_i}{(\lambda'' + x_i^{\beta''})^2}\right] \left[\frac{\alpha''(a''-1)}{\Upsilon_i''} - \frac{a''\alpha''(b''-1)}{\Upsilon_i''(\Upsilon_i''^{-a''} - 1)}\right] \log x_i = 0$$

$$\frac{\partial}{\partial \lambda''} \log L_{H_1} = -\frac{(n-k)}{\lambda''} + \left(\frac{\alpha''+1}{\lambda''}\right) \sum_{i=k+1}^{n} \left(\frac{x_i^{\beta''}}{\lambda'' + x_i^{\beta''}}\right) + \sum_{i=k+1}^{n} x_i^{\beta''} \left(\frac{\lambda''}{\lambda'' + x_i^{\beta''}}\right)^{\alpha''+1} \left[\frac{a''\alpha''(b''-1)}{\lambda''^2 \Upsilon_i''(\Upsilon_i''^{-a''} - 1)} - \frac{\alpha''(a''-1)}{\lambda''^2 \Upsilon_i''}\right] = 0$$

Step-2: The MIC under the null hypothesis (H_0) is defined as

$$H_0 : MIC_n = -2 \log L_{H_0} + 5 \log n \tag{3}$$

where ‘5’ is the number of parameters to be estimated in the KPL distribution. The MIC under the alternative hypothesis (H_1) is defined as

$$H_1 : MIC_k = -2 \log L_{H_1} + \left(10 + \left(\frac{10}{n} - 1\right)^2\right) \log n \tag{4}$$

for a fixed change at location k.

Once, the MIC_n and MIC_k are obtained, we check for the condition given in step-3 of binary segmentation procedure. *In CPA, the change point is the index of each data point from 1 to n samples. Even after partition, the detection of change point will be indicated by its actual index of ‘n’ samples.*

4. Applications of KPL distribution in Change Point Detection

The practical applications of Change point detection was considered to the following real data sets as Floyd river data (Mudholkar and Hutson, 1996) and Eruption data (da Silva, de Andrade, Maciel, Campos and Cordeiro, 2013).

Floyd River Data: The dataset is about the recordings of annual flood discharge rates (in ft^3/s) from the Floyd River at James, Iowa. There are total of 39 samples measured between the period of 1935-1973 with 10 years of split.

Table 1: The Annual Flood Discharge Rates of Floyd River

Years	**Flood Discharge in** (ft^3/s)
1935-1944	1460, 4050, 3570, 2060, 1300, 1390, 1720, 6280, 1360, 7440,
1945-1954	**5320**, 1400, 3240, 2710, 4520, 4840, 8320, **13900**, 71500, 6250,
1955-1964	2260,318, 1330, 970, **1920**, 15100, 2870, 20600, 3810, 726,
1965-1973	7500, **7170**, 2000, 829, 17300, 4740, 13400, 2940, 5660.

Now applying the binary segmentation procedure, the following points are observed.

The $MIC_n > \underset{1\leq k<n}{min} MIC_k$ i.e., 771.167 > 743.3059. So, this indicates that there is a shift/change in the floods of the years.

1. The first change point location is at $k = 18\,(MIC_k = 742.3649)$. So, the total sample $(n = 39)$ gets divided into two segments. First segment is from 1 to 18 data points and rest of them fall under second segment.

2. Continuing the procedure with 18 samples and 21 samples, two more change points are observed. One in the first segment $(n = 18)$ and other in the second segment $(n = 21)$.

3. The second change point is detected at 11^{th} location $(MIC_k = 297.6745)$ of the first 18 samples. Similarly, the 3^{rd} change point is observed at 32^{nd} location of the second segment $(n = 21)$ with $MIC_k = 421.9489$.

4. One more change point is noticed between the 19^{th} and 32^{rd} location, i.e., at $k = 25$ with $MIC_k = 248.5449$.

5. In the above step, the condition given in step-3 of binary segmentation procedure is satisfied, hence the procedure will get terminated.

6. In total, in this dataset, four change points are detected at $k = 18$ (Flood discharge rate=13900ft^3/s); $k = 11$ (Flood discharge rate=5320ft^3/s); $k = 32$ (Flood discharge rate=7170ft^3/s) and $k = 25$ (Flood discharge rate=248.5449ft^3/s).

The entire description is depicted in Figures (2-4). The distribution fit at every iteration is computed and the same is presented in Figure (4).

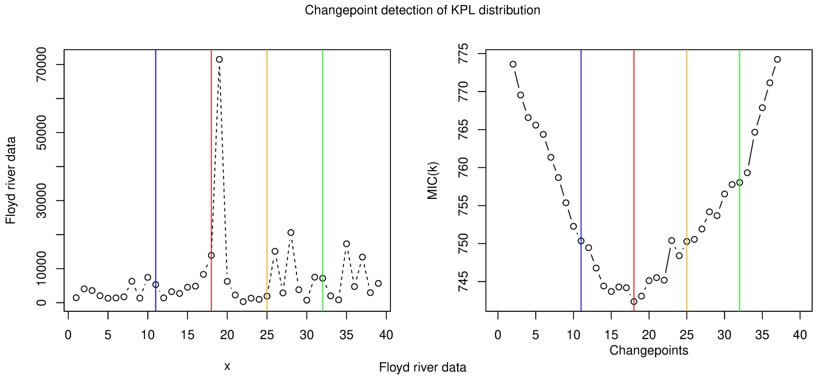


Figure 2: Change point detection using KPL for Floyd river data

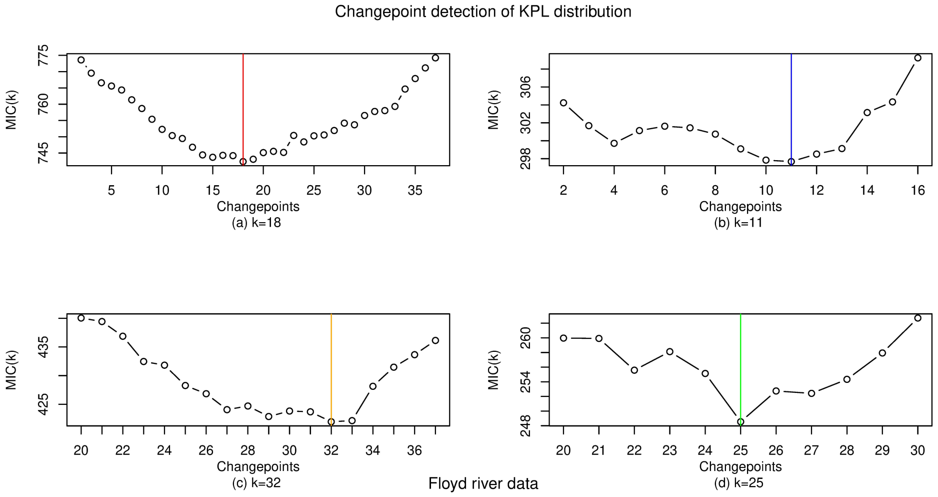
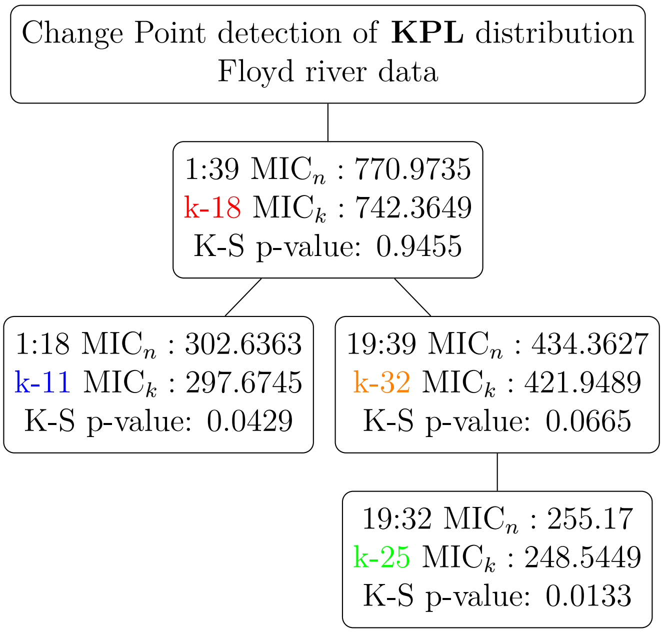


Figure 3: Change point detection values using KPL distribution



Change Point detection of **KPL** distribution
Floyd river data

1:39 MIC_n : 770.9735
k-18 MIC_k : 742.3649
K-S p-value: 0.9455

1:18 MIC_n : 302.6363
k-11 MIC_k : 297.6745
K-S p-value: 0.0429

19:39 MIC_n : 434.3627
k-32 MIC_k : 421.9489
K-S p-value: 0.0665

19:32 MIC_n : 255.17
k-25 MIC_k : 248.5449
K-S p-value: 0.0133

Figure 4: Floyd river data - Decision tree

Eruption Data:

This dataset is about the waiting times (in seconds), between 65 successive eruptions of the Kiama Blowhole. These values were recorded with the aid of digital watch on Jim Irish and the data values are: 83, 51, 87, 60, 28,**95**, 8, 27, 15, 10, 18, 16, 29, 54,**91**, 8, 17, 55, 10, 35, 47, 77, 36, 17, 21,**36**, 18, 40, 10, 7, 34, 27, 28, 56, **8**, 25, 68, 146, 89, 18, 73, **69**, 9,37, 10, 82, 29, 8, 60, 61, **61**, 18, 169, 25, 8, 26, 11, 83, 11, 42, 17, 14, 9, 12.

Now applying the binary segmentation procedure, the following points are observed.

The $MIC_n > \underset{1 \leq k < n}{min} MIC_k$ i.e.,609.2285 $>$ 541.1450. So, this indicates that there is a shift/change in the eruption waiting times.

1. The first change point location is at $k = 26\,(MIC_k = 541.1450)$. So, the total sample $(n = 64)$ gets divided into two segments. First segment is from 1 to 26 data points and rest of them fall under second segment.

2. Continuing the procedure with 26 samples and 38 samples, two more change points are observed. One in the first segment $(n = 26)$ and other in the second segment $(n = 38)$.

3. The second change point is detected at 15^{th} location $(MIC_k = 128.2241)$ of the first 25 samples. Similarly, the 3^{rd} change point is observed at 42^{nd} location of the second segment $(n = 38)$ with $MIC_k = 334.2032$.

4. Here, again the second segment gets divided into two segments. We observed that, two more change points are detected i.e., at $k = 35$ (between 27^{th} and 42^{th} data points) $(MIC_k = 94.8768)$ and $k = 60$ (between 43^{rd} and 64^{th} data points) $(MIC_k = 201.8235)$.

5. One more change point is noticed between the 1^{st} and 15^{th} location, i.e., at $k = 6\,(MIC_k = 55.5510)$.

6. In the above step, the condition given in step-3 of binary segmentation procedure is satisfied, hence the procedure will get terminated.

7. In total, in this dataset, six change points are detected at $k = 26$ (Eruption waiting time=36s); $k = 15$ (Eruption waiting time=91s); $k = 42$ (Eruption waiting time=9s); $k = 35$ (Eruption waiting time=8s); $k = 51$ (Eruption waiting time=61s); and $k = 6$ (Eruption waiting time=95s).

The entire description is depicted in Figures (5-7). The distribution fit at every iteration is computed and the same is presented in Figure (7).

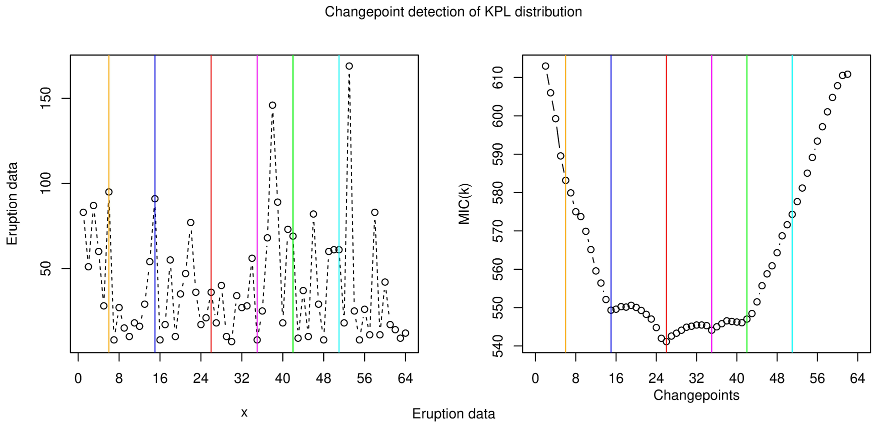


Figure 5: Change point detection using KPL CPD for Eruption data

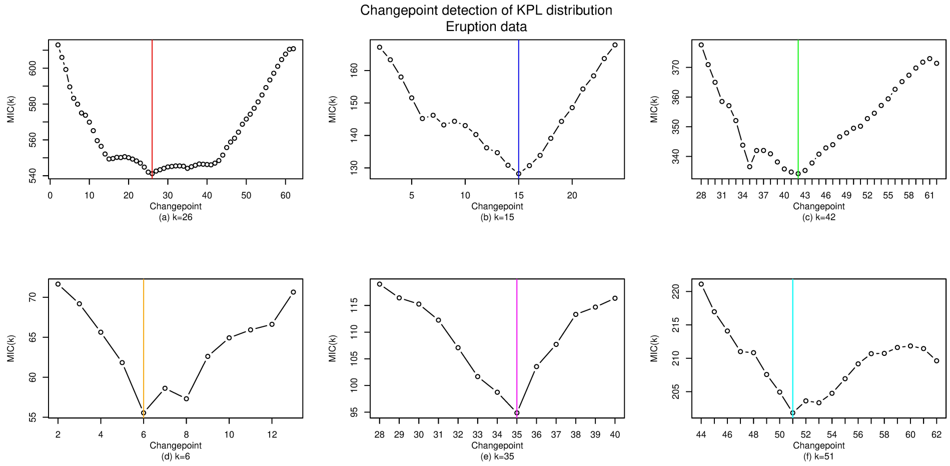


Figure 6: Change point detection values using KPL distribution

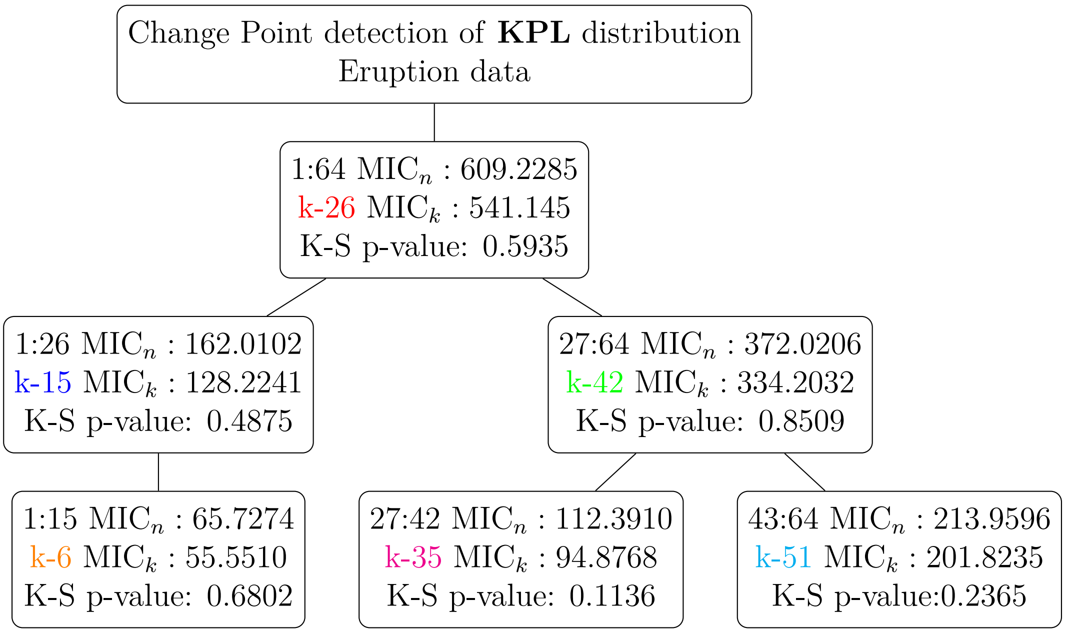

Change Point detection of **KPL** distribution
Eruption data
1:64 MIC_n : 609.2285
k-26 MIC_k : 541.145
K-S p-value: 0.5935
1:26 MIC_n : 162.0102
k-15 MIC_k : 128.2241
K-S p-value: 0.4875
27:64 MIC_n : 372.0206
k-42 MIC_k : 334.2032
K-S p-value: 0.8509
1:15 MIC_n : 65.7274
k-6 MIC_k : 55.5510
K-S p-value: 0.6802
27:42 MIC_n : 112.3910
k-35 MIC_k : 94.8768
K-S p-value: 0.1136
43:64 MIC_n : 213.9596
k-51 MIC_k : 201.8235
K-S p-value:0.2365

Figure 7: Eruption data - Decision tree

5. Discussions

In this paper, we discussed that, the KPL distribution plays an important role to detect multiple change points. For the datasets considered, at each and every iteration test for goodness of fit and computed MIC values are computed. The tree diagram representation is based on the algorithm of binary segmentation.

References

Al-Marzouki, S., Jamal, F., Chesneau, C., and Elgarhy, M. (2020). Type II Top Leone Power Lomax distribution with applications. *Mathematics*, **8(1)**, 1-26.

Chen, J., Gupta, A., and Pan, J. (2006). Information criterion and change point problem for regular models. *Sankhya: The Indian Journal of Statistics*, **68(2)**, 252-282.

Chen, J., and Gupta, A. K. (1997). Testing and locating variance change points with application to stock prices. *Journal of the American Statistical association*, **92(438)**, 739–747.

Chen, J., and Gupta, A. K. (2011). Parametric statistical change point analysis: with applications to genetics, medicine, and finance. *Springer Science & Business Media.*

Cordeiro, G. M., Ortega, E. M., and Nadarajah, S. (2010). The kumaraswamy weibull distribution with application to failure data. *Journal of the Franklin Institute*, **347(8)**, 1399-1429.

da Silva, R. V., de Andrade, T. A., Maciel, D. B., Campos, R. P., and Cordeiro, G. M. (2013). A new lifetime model: The gamma extended frechet distribution. *Journal of Statistical Theory and Applications*, **12(1)**, 39-54.

Lomax, K. (1954). Business failures: Another example of the analysis of failure data. *Journal* of the American Statistical Association, **49(268)**, 847-852.

Mudholkar, G. S., and Hutson, A. D. (1996). The Exponentiated Weibull family: some properties and a flood data application. *Communications in Statistics–Theory and Methods*, **25(12)**, 3059–3083.

Nagarjuna, B.V. , and Vishnu Vardhan, R. (2020). Marshall-olkin Exponential Lomax distribution: Properties and its application. *Stochastic Modelling and Applications*, **24**, 161-177.

Nagarjuna, B. V., Vardhan, R. V., and Chesneau, C. (2021a). Kumaraswamy generalized Power Lomax distributionand its applications. *Stats*, **4(1)**, 28–45.

Nagarjuna, B. V., Vardhan, R. V., and Chesneau, C. (2021b). On the accuracy of the sine Power Lomax model for data fitting. *Modelling*, **2(1)**, 78–104.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42(3/4)**, 523–527.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41(1/2)**, 100–115.

Parana, P. F., Ortega, E. M., Cordeiro, G. M., and Pascoa, M. A. d. (2013). The Kumaraswamy Burr XII distribution: theory and practice. *Journal of Statistical Computation and Simulation*, **83(11)**, 2117–2143.

Rady, E. H. A., Hassanein, W. A., and Elhaddad, T. A. (2016). The Power Lomax distribution with an application to bladder cancer data. *SpringerPlus*, **5(1)**, 1–22.

Shams, T. M. (2013). The Kumaraswamy-Generalized Lomax Distribution. **17(5)**, 641-646.

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. *In Doklady akademii nauk*, **259**, 270–274.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 171- 187

Multi-Criterion Approach for State and District Level Agricultural Infrastructural Adequacy

Rajni Jain, Prem Chand and Priyanka Agarwal

ICAR-National Institute of Agricultural Economics and Policy Research

Received: 04 August 2022; Revised: 10 August 2022; Accepted: 13 August 2022

Abstract

Agricultural infrastructure is essential to transform subsistence agriculture to commercial and dynamic farming system by lowering farming costs and increasing the farm income. This study developed a methodology for quantifying the status of both physical and institutional infrastructures for agriculture in India. The study identified the relative state-level agricultural infrastructural adequacy status in the country based on secondary datasets and also validated the methodology at district level using a case study of Bundelkhand region. Finally, the composite infrastructural category is identified for both state level in the country and district level in Bundelkhand Region.

Key words: Multi-criterion approach; Infrastructural adequacy; Pair wise comparison.

1. Introduction

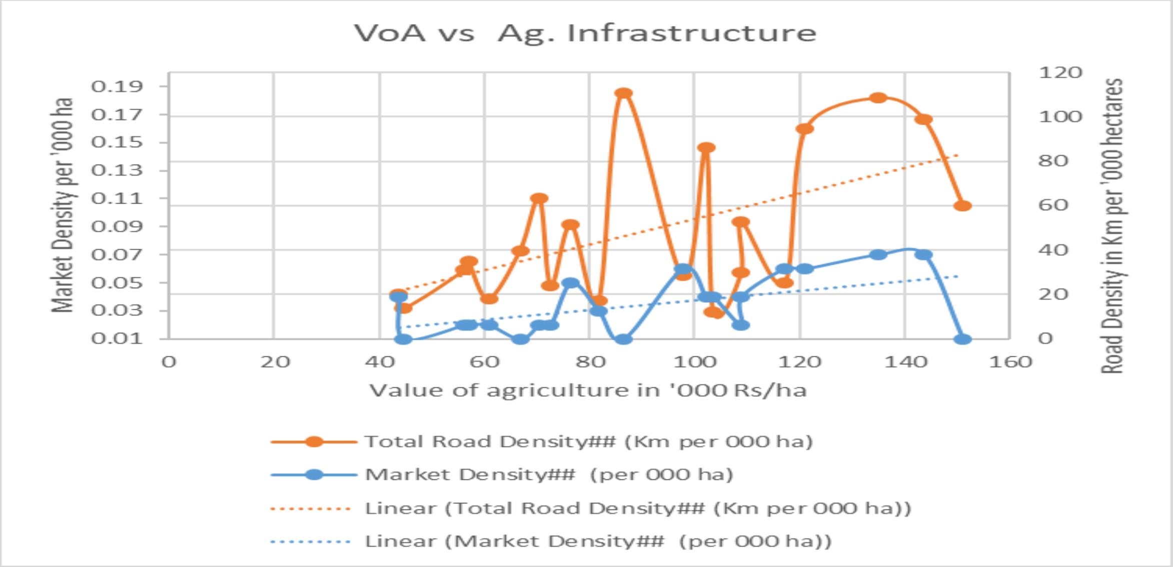
Inadequate rural infrastructure has been considered as a major reason for low agricultural productivity. Agricultural productivity depends on rural infrastructure, well-functioned domestic markets, appropriate institutions and access to appropriate technologies (Andersen and Shimokowa, 2006). Improved roads lead to rise in rural small non-farm business (Fan and Zhang, 2004). One per cent increase in the stock of infrastructure is associated with one per cent increase in GDP across all countries (Patel, 2010). An efficient marketing system leads to enhanced farm income (Kamara, 2004). Connectivity to rural roads may result in change in cropping pattern. One per cent increase in irrigated area has bought about 0.32 per cent increase in productivity of all inputs (Casella and Schilling, 2017). Scientific and holistic development of storage structures promotes horticulture commodities and also controls the food inflation. The relation between agricultural productivity and two important infrastructures on the basis of state level estimates from 21 major states is shown in Figure 1. Pearson correlation coefficient for both market density and road density was observed significantly positive with values 0.52 and 0.56 at 1% and 5% respectively.

Agricultural infrastructure has potential to transform subsistence agriculture to commercial and dynamic farming system as adequate markets, roads, irrigation, extension services, credit facilities, storage *etc*., facilitates lowering farming costs and increase in farm income. The extension personnel and communication technologies help the farmers in better understanding of various supportive policies and schemes. Credit has positive and significant effect on agricultural production. Thus, financial institutions help in increasing production investments, which in turns enhances farmers' return. It is easier to classify a district or state into adequate or inadequate category on the basis of a single parameter, but classifying a region into these categories is difficult with multiple parameters. Considering the importance of agricultural infrastructure, allocation to the Agriculture Infrastructure Fund (AIF) increased to

Corresponding Author: Rajni Jain
E-mail address: Rajni.jain@icar.gov.in

Rs. 500 crores in 2022-23 from Rs. 200 crores in RE for 2021-22. However, identifying the suitable regions having inadequate infrastructure for further infrastructural development is a challenge.

In this study, we propose a methodology for multi-criteria based agricultural infrastructural classes. The methodology is useful for any region, state level as well as at district level. Land suitability classes proposed by FAO (2017) motivated authors to identify the agricultural infrastructure adequacy classes. However, there are differences between land evaluation methodology and the one required for adequacy level estimation of infrastructure as (i) Crop wise adequacy level not needed and (ii) Standard requirement for different classes not available.



Source: Prepared by authors based on data regarding SGDP-agriculture from MOSPI, number of markets from AGMARKETNET, road density from Ministry of road, transport and highways, GoI

Figure 1: Relation between agricultural productivity and infrastructure

2. Data and Methodology

This section presents the data sources used for the development of infrastructural classes and the methodology used for classifying the infrastructural status of each dimension as well as for composite infrastructure.

2.1. Data

The data availability regarding various dimensions of agricultural infrastructure from the authentic government websites is presented in Table 1.

2.2. Weight determination using the Analytical Hierarchical Process

The Analytical Hierarchical Process (AHP) method is considered among the best available approaches to deal with relative importance of one criterion over another for determining the parameter weights, as per the AHP preference scale (Table 2). A scale of 9 indicates that one factor is more important than the other, while 1 means equal importance. The reciprocals of 1 to 9 (1/1 and 1/9) show that one is less important than the other (Saaty and

Vargas, 2001). In the pairwise comparison matrix (PWCM), the importance of parameters is decided by the experts as given in Table 2.

Table 1: Dimensions and data sources

Dimension	Website for data source
Markets	https://agmarknet.gov.in
Irrigation	https://eands.dacnet.nic.in/
Road density	https://censusindia.gov.in
KVK	https://kvk.icar.gov.in/
Credit	https://censusindia.gov.in
Communication	https://censusindia.gov.in
Storage	http://www.nccd.gov.in

Table 2: Preference scale between two parameters in AHP

Relative importance	**Definition**	**Description**
1	Equally important	Two factors contributing uniformly to the predefined goal.
3	Moderately important	Experience and judgment are negligibly in favor of one as compared to the another.
5	Strongly important	Experience and judgement strongly in favor of one in comparison to the other.
7	Very strong important	Experience and judgments very strongly favour one over the another. Its necessity is revealed in practice.
9	Extremely important	The sign favoring one as compared to the other parameter is of the maximum possible validity.
2,4,6,8	Intermediate	When compromise is needed
Reciprocals	Less importance	
← 1/9 1/7 1/5 1/3 1 3 5 7 → 9		
Less Importance more		

Source: (Saaty and Vargas, 2001).

After getting the importance from the experts, the weights for each parameter can be determined using the Satty method (Satty and Vargas, 2001). In the AHP method, while executing the pairwise comparisons of criteria, a certain level of variation may follow. To tackle this problem, Consistency Ratio is used for preventing bias through criteria weighting. As a solution, eigenvectors and the largest eigenvalue of the respective matrix were computed, and the consistency index (*CI*) was examined using the following equation:

$$CI = (\lambda_{max} - n)/(n-1) \tag{1}$$

Here, λ_{max} represents the maximum eigenvalue of the pairwise comparison matrix and *n* is the number of criteria in each Pair Wise Comparison Method (PWCM). At last, the uniformity of the PWCM is examined using the random consistency index (*RI*) value as shown in Table 3. Consistency Ratio (*CR*) was computed by using the method given below:

$$CR = CI/RI \tag{2}$$

To be valid, its consistency ratio should be ≤ 0.10. If the acquired value is larger than 0.10, it is essential to develop the pairwise comparison matrix again. Random Index value for varying "*n*" is shown in Table 3 (Chang, 2007; Shaloo *et al*. 2022).

Table 3: Random index (*RI*) value for varying "*n*" in the AHP

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

** The RI value for 8 criteria is 1.32*
Source: Shaloo et al. 2022

2.3. Infrastructure adequacy classes

In this study, the FAO land suitability evaluation methodology (FAO, 2017; Elsheikh *et al*., 2013) which classifies the crop suitability classes into five major (*S*1, *S*2, *S*3, *N*1 & *N*2) classes was adapted for measuring the socio-economic adequacy (Table 4). However, we have combined the two not suitable classes i.e., *N*1 and *N*2 into *N*.

Table 4: Land suitability classes

Class ID	**Class**	**Class description**
*S*1	Highly Suitable	Land having no significant limitations to sustained application of a given use, or only minor limitations that will not significantly reduce productivity or benefits and will not raise inputs above an acceptable level.
*S*2	Moderately Suitable	Land having limitations which in aggregate are moderately severe for sustained application of a given use; the limitations will reduce productivity or benefits and increase required inputs to the extent that the overall advantage to be gained from the use, although still attractive, will be appreciably inferior to that expected on Class *S*1 land.
*S*3	Marginally Suitable	Land having limitations which in aggregate are severe for sustained application of a given use and will so reduce productivity or benefits, or increase required inputs, that this expenditure will be only marginally justified.
*N*1	Currently Not Suitable	Land having limitations which may be surmountable in time but which cannot be corrected with existing knowledge at currently acceptable cost; the limitations are so severe as to preclude successful sustained use of the land in the given manner.
*N*2	Permanently Not Suitable	Land having limitations, which appear so severe as to preclude any possibilities of successful sustained use of the land in the given manner.

Source: FAO, 2017

There are two basic differences between land evaluation methodology and the one required for adequacy level estimation of infrastructure. Firstly, infrastructure is mostly common for all crops; hence, there is no need for estimation at crop level. Secondly, for land evaluation framework, standard requirement of a crop is known based on agronomic practices and field level research done for the crops. On the other hand, to the best of our knowledge, there is no standard infrastructural requirement available in the known literature which can be used to match the present availability status of infrastructure with the required level. In the presented methodology, the authors attempt to quantify the existing status of the agricultural infrastructure in the study states. After quantification in all the states, four infrastructural classes have been identified along with the respective range for each class. The identified ranges are applicable to any region, state or district in India, as it is based on village level data from 21 major states in the country.

Rest of this section presents the quantification models and the criteria identified for adequacy classes of each infrastructural parameter. For estimating the market suitability, data related to number of markets was transformed into a variable called radial distance. To estimate road, communication, extension and credit suitability a corresponding score was estimated as explained in following sub-headings.

2.4. Market concentration

The market concentration is expressed by radial distance catered by a market in kilometres. Market concentration (number of markets per 1000 hectares of NSA) was modified to radial distance (R) using equation (3). R is inversely related to market availability as lower radial distance means ease of market availability.

Radial distance catered by one market in kilometres is as follows:

$$R = \sqrt{\frac{Net\ sown\ area\ (in'000ha)}{100 * \Pi *\ no.of\ markets}} \tag{3}$$

Further, the radial distance of each state level was obtained to ascertain the range of values for each market based suitability class and the obtained range for each suitability class is specified in Table 5.

2.5. Irrigation infrastructure score

Area under irrigation per ha of net sown area was used as a proxy for availability of irrigation infrastructure in the state. Criteria for irrigation infrastructure suitability across states was determined as presented in the Table 7. The identified criteria depict that the states having irrigation availability to more than 82 per cent of Net sown area are under $S1$ category while the ones, which are less than 17 per cent, are under N category.

2.6. Road density score

Criteria for road suitability class was developed using the village level data in the country and presented in Table 5. Data on seven types of roads namely national highways (NH), state highways (SH), district roads (DR), other district roads (ODR), pucca road (PR), kuchcha road (KR) and water bound macadam (WBM) from village amenities dataset was used for estimating the road density score. The qualitative road data was converted to quantitative values using scores in the range 0-10 (Table 5).

Table 5: Scoring criteria based on availability or the distance of a facility (road, communication or credit)

Distance	**Score**
Available	10
Available within 5 km range	5
Available within 5 to 10 km range	3
Available at more than 10 km	0

Source: Authors based on expert opinion

Further, relative importance of each type of roads is estimated using pairwise comparison method developed by Satty and Vargas, 2001 as mentioned above in Section 2.1. The weights as estimated for each category of roads are given in Table 6.

Table 6: Estimation of weights of different category of roads using pair-wise comparison matrix

Road types	**NH**	**SH**	**DR**	**ODR**	**PR**	**KR**	**WBM**	**Weights in fraction (w)**
NH	1	1	3	4	5	8	9	0.3
SH	1	1	2	3	4	8	9	0.3
DR	1/3	½	1	2	3	6	7	0.2
ODR	1/4	1/3	1/2	1	2	5	7	0.1
PR	1/5	¼	1/3	1/2	1	5	7	0.1
KR	1/8	1/8	1/6	1/5	1/5	1	3	0.0
WBM	1/9	1/9	1/7	1/7	1/7	1/3	1	0.0

Note: National Highways (NH), State Highways (SH), District Roads (DR), other district roads (ODR), Pucca Road (PR), Kuchcha Road (KR) and Water Bound Macadam (WBM)

Source: Authors based on expert opinion

Now, using obtained quantitative individual scores as well as weights associated to each road variable, road score of villages (S_{v_i}) were estimated and aggregated using area based weightage of each village in the state to obtain road density score of a higher region *e.g.* district or a state (Equation 5).

$$S_V = \sum_{j=1}^{7} W_j R_j \tag{4}$$

where,

S_v = Road suitability score of a village

R_j = score for j^{th} type of road in the village (Table 3)

W_j is the weight assigned to the j^{th} type of road (Table 4)

$$Road\ Density\ Score\ (S_s) = \sum_{i=1}^{n} \frac{a_i}{A} * S_{v_i} \tag{5}$$

where,

a_i = area of i^{th} village
$A = \sum a_i$ "*i.e.,* sum of areas of all villages in the higher region *i.e.,* state or district."

2.7. Extension suitability score

Extension suitability score is an aggregate score of KVK score and communication score.

2.7.1. KVK score

The sufficiency of extension personnel in a state is estimated using number of subject matter specialist including heads and other staff working in the KVK of the state. Thus, vacant posts denote the lack of extension personnel in the state. KVK score (Ks) is estimated using the ratio of filled posts to total number of posts in a KVK (Equation 6)

$$KVK\ Score\ (Ks) = \frac{Number\ of\ \ Posts\ Filled}{Total\ Approved\ Posts} \tag{6}$$

2.7.2. Communication score

The score is estimated using data on village amenities as extracted from census of India, 2011. Availability status of Landline, PCO, Mobile and Internet were taken as the main communication infrastructure. The qualitative data on communication score was first converted to quantitative data as presented in Table 5 for the road data.

Then, an aggregate communication score was estimated by giving weights to each mode based on their importance (based on expert's opinion). The mathematical equation (7) is shown under.

$$C_V = \sum_{j=1}^{4} W_j M_j \tag{7}$$

where,

C_v is communication score of the village 'v' out of 10,

W_j is the weight assigned to the jth mode of communication (0.35 for landline and mobile each, 0.1 for PCO and 0.2 for internet),

$M_j = j^{th}$ mode of communication *i.e.,* Landline (j=1), PCO (j=2), Mobile (j=3) and Internet (j = 4)

An aggregate communication score of a state (C_S) is then obtained by combining weighted village communication score (C_V) of all the villages in the state (Equation 8)

$$C_s = \sum_{i=1}^{n} \frac{a_i}{A} * C_{v_i} \tag{8}$$

where,

C_S: State communication score
C_{v_i}: Communication score of i^{th} village
n: Number of village in the state

(Notations 'a_i' and 'A' are similar to equation 5)

Extension suitability score (*Es*) of a state is finally estimated using the Equation 9, allotting 0.6 weight to communication and 0.4 weight to KVK (based on expert opinion)

$$Es = 0.6^* C_s + 0.4^* K_s \tag{9}$$

2.8. Credit suitability score

For assigning score to credit facilities, equal weightage were given to four institutional setups commercial bank, cooperative bank, agricultural credit societies and self-help groups. In first stage, credit suitability of a village was estimated using Equation 10

$$L_v = \sum_{j=1}^{4} 0.25 L_j \tag{10}$$

where,

L_v = credit suitability score of the village

L_j = score of j^th^ institutional setup for availing credit

The credit suitability of the state (L_s) was estimated by using area weightage of each village (Equation 11)

$$L_s = \sum_{i=1}^{n} \frac{a_i}{\mathrm{A}} * L_{v_i} \tag{11}$$

where,

L_s = state/ region credit suitability score

L_{v_i}= credit suitability score of i^{th} village

(Notations 'a_i' and '*A*' are similar to equation 5)

Table 7: Criteria used for assigning suitability classes to road, extension and credit adequacy

Category	mean plus standard deviation (S1)	Mean to mean plus standard deviation (S2)	Mean minus standard deviation to mean (S3)	< Mean minus standard deviation (N)
Irrigation score (I)	> 8.2	8.2 - 4.9	4.9 - 1.7	1.7
Road score (S)	> 5.48	4.53 - 5.48	3.58 - 4.53	3.58
Extension score (E)	> 8.3	8.30 - 6.51	6.51 - 4.72	4.72
Credit score (L)	> 5.73	5.73 - 4.00	4.00 - 2.28	2.28
Radial distance (R)	< 6.45	6.45 - 10.88	10.88 - 15.10	15.10

Source: Estimated by the authors

2.9. Estimation of storage suitability score

Data on state-wise requirement and availability of cold-storage structures was collected from All India Cold-chain Infrastructure Capacity (*Assessment of Status & Gap*) by NCCD (2015) to estimate the storage gap in '000 MT. The gap in the study was assessed solely on current consumption patterns of the urban population in the country.

Surplus and deficit states based on the per cent gap between requirement and availability of the storage capacity with respect to availability were identified. Surplus indicates higher availability and deficit represents higher requirement. Further, distribution of deficit states with respect to severity of gap was obtained. Per cent gap with respect to availability up to 25 per cent were categorized as marginally deficit, 25 to 50 as moderately deficit, 50 to 75 as deficit and more than 75 percent gap were considered as highly deficit states.

2.10. Estimation of composite infrastructural suitability

Composite infrastructural suitability of a state (O_i) was estimated using the worst criteria principle (Rezaei, 2015) as presented using equation 12.

$$O_i = min\ (R_i, I_i, S_i, E_i, L_i, W_i) \tag{12}$$

Where, R_i, I_i, S_i, E_i, L_i and W_i refer to the estimated suitability classes for market, irrigation, road, extension, credit and storage for i^{th}state.

3. Results and Discussion

The methodology as expressed in Section 2 was used to calculate the infrastructural adequacy at state level and district level for Bundelkhand region. The level of adequacy is presented in section 3.1 and 3.2 respectively.

3.1. State level

Spatial variation in selected agricultural infrastructure based suitability classes amongst various states are illustrated through Table 8 - 9. Based on the criteria score of different classes as shown in Table 5, the states are categorised into *S*1 (highly suitable), *S*2 (moderately suitable), *S*3 (marginally suitable), and *N* (not suitable) in Figures 2-7.

Table 8: Market Radial distance (in Kms), road density, communication, KVK, extension and credit (scores out of 10) for major states, India

S. No.	States	Radial Distance	Irrigation	Road Density	Comm-unication	KVK	Extension	Credit Suitability
1.	Andhra Pradesh	9.89	4.50	4.56	7.60	8.60	8.00	4.98
2.	Arunachal Pradesh	6.77	1.00	2.81	2.30	9.10	5.02	1.29
3.	Assam	18.92	0.60	5.53	5.10	9.80	6.98	3.58
4.	Bihar	17.22	6.40	4.99	4.60	7.60	5.80	3.87
5.	Chhattis-garh	8.87	3.00	3.79	4.20	8.30	5.84	3.75
6.	Gujarat	10.37	4.20	4.94	9.10	7.40	8.42	5.32
7.	Haryana	9.10	8.80	6.17	8.10	7.30	7.78	6.23
8.	Himachal Pradesh	6.50	2.00	2.84	3.40	8.30	5.36	1.77
9.	J&K	8.23	4.00	4.18	5.10	7.70	6.14	2.19

10.	Jharkhand	12.65	3.50	3.80	3.50	6.40	4.66	2.44
11.	Karnataka	12.52	2.70	4.13	7.60	8.60	8.00	5.14
12.	Kerala	7.60	2.90	6.13	9.40	9.60	9.48	8.93
13.	Madhya Pradesh	12.07	4.50	4.04	4.80	6.50	5.48	3.51
14.	Maharash-tra	12.92	8.00	4.76	7.70	8.60	8.06	5.51
15.	Odisha	11.48	2.20	3.93	5.70	7.00	6.22	3.19
16.	Punjab	7.30	9.80	5.41	8.00	8.40	8.16	4.84
17.	Rajasthan	18.89	3.10	4.49	8.10	7.40	7.82	3.72
18.	Tamil Nadu	8.49	5.70	5.70	7.00	8.60	7.64	5.04
19.	Uttar Pradesh	2.96	8.40	4.38	5.20	8.30	6.44	4.16
20.	Uttara-khand	45.94	4.00	5.06	5.70	6.20	5.90	2.96
21.	West Bengal	14.77	5.20	3.51	6.30	8.40	7.14	4.17

Source: Estimated by authors

Table 8 depicts that four states are agriculturally not suitable as per marketing status including Uttarakhand, Assam, Rajasthan and Bihar. Uttarakhand has the highest radial distance of about 46 Km, indicating lack of agricultural markets in the state. Uttar Pradesh having least radial distance of about 3 Km and Himachal Pradesh with radial distance of 6.5 Km are better off than other states and are the only two states which are under highly suitable category. Though, the radial distance catered by markets in agriculturally developed states like Punjab and Haryana is relatively higher (7.3-9.1 Km respectively), yet presence of adequate road infrastructure compensates this to some extent.

Road adequacy infrastructure status helps the policy makers in prioritizing the needs of the states. The value of score is lowest *i.e.,* about 2.8 for Arunachal Pradesh and Himachal Pradesh, followed by West Bengal, Chhattisgarh, Jharkhand and Odisha, indicating lack of road infrastructure in these states and demand immediate focus from policy makers in order to facilitate agricultural development in the region.

Irrigation score of the states (Table 9) shows that even though water is a major constraint, the states like Punjab, Haryana, Uttar Pradesh, and Maharashtra have more than 80 per cent of the net sown area under irrigation, with Punjab having highest, 98 per cent of the net sown area under irrigation, indicating adequate infrastructure availability in these states. On the other hand, most of the north-eastern states like Arunachal Pradesh and Assam have less than 10 per cent area under irrigation, which can be due to either lack of irrigation facility or no requirement of irrigation in these states.

Table 9: Percent of villages having access to irrigation across irrigation categories

Category→ State↓	Irrigation categories				
	< 20%	20-40%	40-60%	60-80%	> 80%
Andhra Pradesh	31	15	12	10	27
Arunachal Pradesh	11	0	0	0	0
Assam	74	2	2	1	2
Bihar	9	11	16	24	39
Chhattisgarh	60	10	6	5	13
Gujarat	31	15	17	16	18
Haryana	4	3	4	6	83
Himachal Pradesh	64	7	5	3	8
Jammu & Kashmir	40	9	10	11	28
Jharkhand	68	11	6	3	7
Karnataka	43	23	12	7	8
Kerala	44	13	12	10	20
Madhya Pradesh	29	21	19	15	14
Maharashtra	6	5	7	13	66
Odisha	60	4	4	5	10
Punjab	1	1	1	2	94
Rajasthan	35	17	16	13	17
Tamil Nadu	12	12	13	15	43
Uttar Pradesh	3	2	5	11	77
Uttarakhand	69	8	4	3	11
West Bengal	23	15	12	15	34

The extension score indicates that Kerala and Gujarat with the highest extension score of 9.48 and 8.42 are the highly suitable states. On the other hand, Jharkhand, Mizoram and Meghalaya are agriculturally not suitable as per extension suitability score. The major producing states like Punjab, Maharashtra, Haryana and West Bengal with an extension score varying between 8.30 - 6.51 are moderately suitable states. While with score lying between 4.00 - 2.28, Uttar Pradesh and Madhya Pradesh are found marginally suitable. Thus, there is need of strengthening of extension services in the states.

Based on credit suitability score, Kerala is the most suitable state in the country with the score of 8.93. It has the highest credit suitability as cent per cent of the villages in Kerala have SHG, 78 per cent have commercial bank, 92 per cent cooperative bank and 63 per cent of the villages have agricultural credit societies (Census, 2011). Himachal Pradesh and Arunachal Pradesh are found not suitable, indicating dearth of banking infrastructures in these states. Bihar, Chhattisgarh, Rajasthan, Madhya Pradesh, Odisha, Uttarakhand and Jharkhand are marginally suitable states with credit suitability score in the range 1.29 -3.87.

Cold storage: It has been found that there is an overall requirement-gap of 10101 (‘000 MT) in the country for fruits, vegetables, dairy and meat products. The requirement-availability gap is highest in Bihar (3876 ‘000MT) and West Bengal (3586 ‘000MT) followed by states like Maharashtra (2527 ‘000MT), Madhya Pradesh (1905 ‘000MT), Jammu Kashmir (843 ‘000MT), Gujarat (520 ‘000MT) and Karnataka (500 ‘000MT). This indicates the huge scope of agricultural development through cold storage infrastructural development. States like Uttar Pradesh show surplus availability of cold storage structures up to the range of 2874 ‘000 MT. This reveals the scope for increasing production of high value crops like fruits and vegetables besides development of dairy and livestock thereby enhancing the farmer's income. States like Punjab and Andhra Pradesh also display sufficient cold storage facilities for the perishables. State level data on cold storage indicates Uttar Pradesh with surplus while Madhya Pradesh having huge gaps in the cold storage capacity.

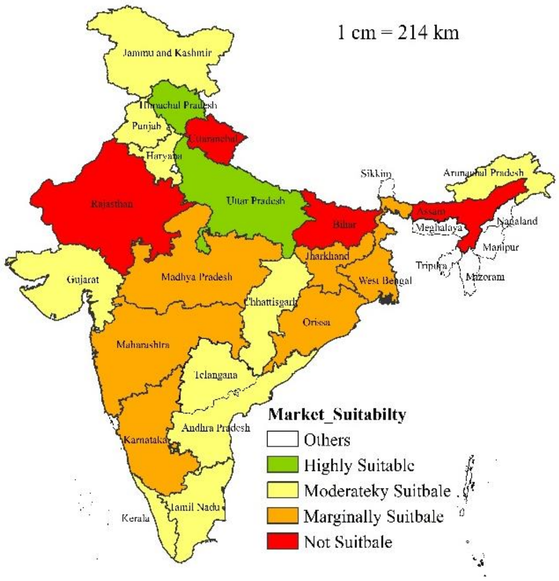


Figure 2: Market infrastructure

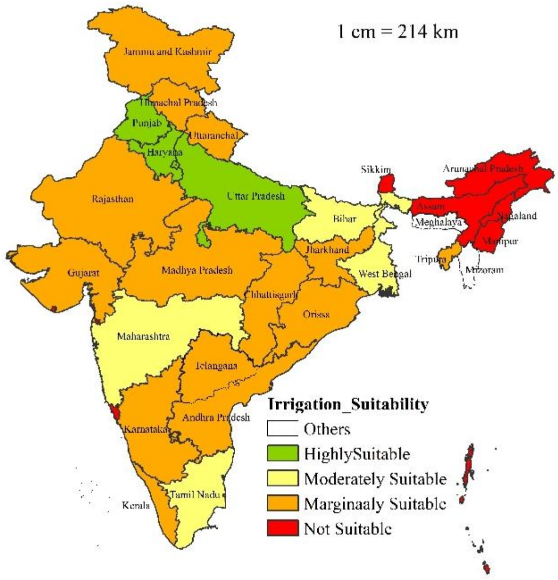


Figure 3: Irrigation infrastructure

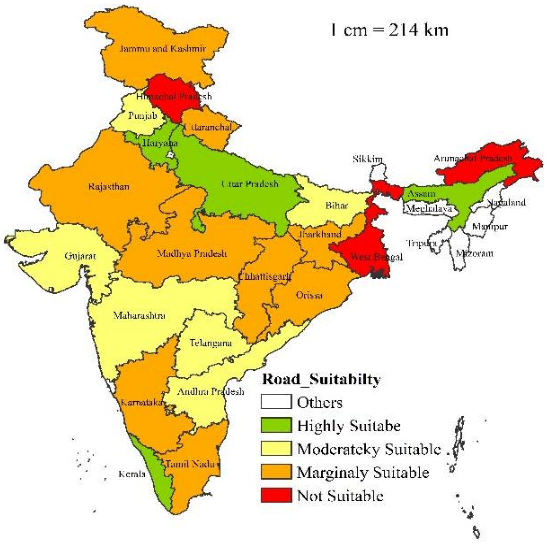


Figure 4: Road infrastructure

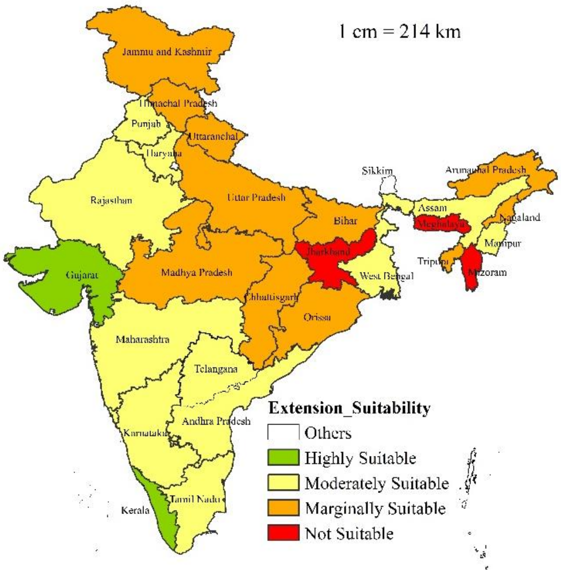
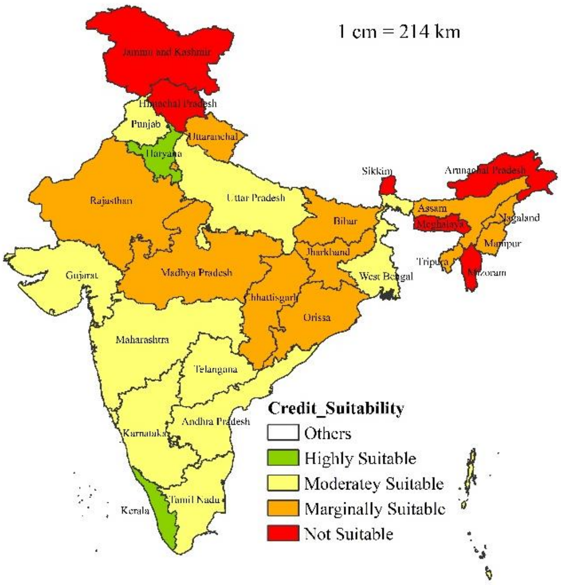


Figure 5: Agricultural Extension



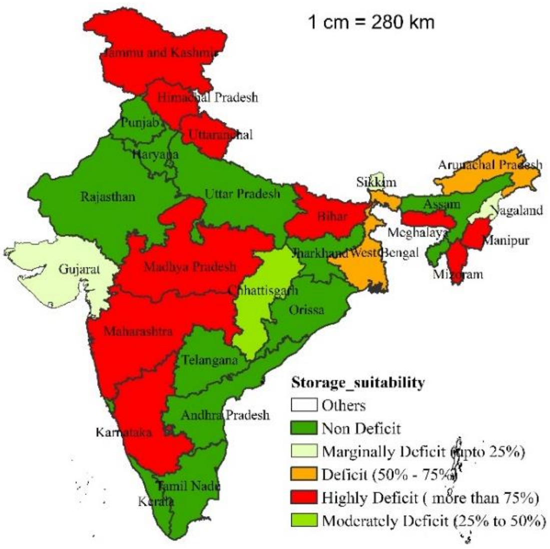


Figure 6: Agricultural credit **Figure 7: Gaps in cold storage**

Source: Authors

3.2. Composite infrastructural suitability of India

Based on suitability of each individual infrastructure facility, the state level composite infrastructural suitability status of the states was identified using the worst criteria principle (Table 10, Figure 8). The results show that none of the states is observed as having adequate infrastructure. Haryana, Punjab, Gujarat, Andhra Pradesh and Kerala are moderately suitable in terms of infrastructural adequacy (%). Other states are having marginal or not suitable infrastructure adequacy (%) indicating the dire need to improve the status of the one or more aspects of agricultural infrastructure (Figure 8).

Table 10: Distribution of the states as per suitability classes as per agricultural infrastructure adequacy

Suitability Class	**Name of the states falling in this category**	**Percent to selected states (*n*=21)**
*S*1	None	0.00
*S*2	Haryana and Punjab	9.52
*S*3	Andhra Pradesh, Chhattisgarh, Gujarat, Kerala, Orissa, Tamil Nadu and Uttar Pradesh	33.34
N	Arunachal Pradesh, Assam, Bihar, Himachal Pradesh, Jammu and Kashmir, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Rajasthan, Uttaranchal and West Bengal	57.14

Source: Authors

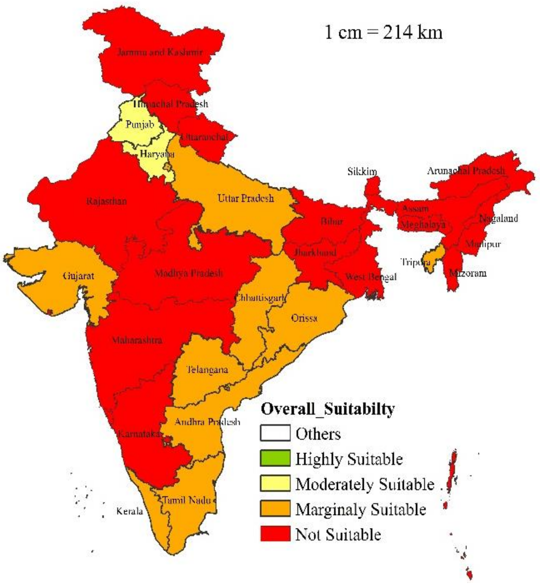


Figure 8: State level infrastructural adequacy for agriculture in India

Source: Authors

3.3. District level case study

We validated the proposed methodology at district level for Bundelkhand region. The Bundelkhand Region of central India is a semi-arid plateau that comprises seven districts of Uttar Pradesh (U.P.) *viz.,* Jhansi, Jalaun, Lalitpur, Mahoba, Hamirpur, Banda and Chitrakoot and six districts of Madhya Pradesh (M.P.) *viz.*, Datia, Tikamgarh, Chhatarpur, Panna, Damoh and Sagar. Agriculture in Bundelkhand is rainfed, diverse, complex, under-invested, risky and vulnerable. The yields obtained by the Bundelkhand farmers are usually lower than the state average for majority of the crops. District level assessment of infrastructural adequacy can help to determine level and kind of development needed in each district.

The district wise suitability score for market, roads, irrigation, extension and credit is estimated as per the criteria (Table 5) and infrastructural suitability class is shown in Figures 9-13.

Bundelkhand region is drastically deprived of the infrastructural facilities for agricultural commodities' market access in comparison to agriculturally developed states of India. The UP-Bundelkhand as well as MP-Bundelkhand are under marginally suitable category in terms of market infrastructure (Figure 9).

With reference to irrigation, out of 13 district, one districts namely Lalitpur of UP Bundlekhand region is suitable (S1), six districts *viz.,* Sagar, Tikamgarh, Jalun, Jhansi, Datia and Banda are moderately suitable and six districts of which three districts *viz.,* Hamirpur, Chirakoot, Mahoba are of UP Bundlekhand and remaining three districts *viz.,* Chhatarpur, Panna and Datia of MP Bundlekhand are marginally suitable (Figure 10). The results confirm the report of NITI Aayog indicating lowest irrigated area with respect to gross cropped area in Damoh and Chitrakkot district of Bundlekhand region (NITI, 2016).

Road density score among districts in Bundelkhand varies from 3.03 to 5.32 with an average score of 4.15. Thus, districts of Bundelkhand were categorised as moderately to not suitable classes in terms of road density score implying lack of road infrastructure facilities in the region (Figure 11). Thus, there is ample scope to improve agricultural income by road infrastructure development in the region.

Regarding Extension infrastructure in Bundelkhand MP region, 3 districts out of 6, namely Chhatarpur, Datia and Panna with extension score of 4.46, 2.46 and 4.54 are under ‘not suitable’ category. While the other three are marginally suitable. In Bundlekhand UP, 2 out of 7 district namely Chitrakoot and Lalitpur are moderately suitable and the remaining 5 district are marginally suitable (Figure 12). Inadequate technology delivery system coupled with acute shortage of staff were the major backdrops for the region. Therefore, focus on improving the extension services in the region is essentially required.

The credit suitability score across the districts of Bundelkhand region shows that overall two district of Bundelkhand UP *viz.*, Banda and Jalaun are moderately suitable. While the remaining 11 districts are marginally suitable (Figure 13). There is a strong need for development of rural financial infrastructure in Bundelkhand region.

Closer inspection of cold storage availability and their capability in districts of Bundelkhand indicates lack of cold storage facility (Indiastat.com). The scenario demands for the inclusion of storage facilities in the development plan for the region besides other infrastructural facilities.

Composite infrastructural suitability of Bundlekhand region show that Five out of 13 districts of the are in ‘not suitable’ category while remaining districts are ‘marginally suitable’, indicating lack of agricultural infrastructural adequacy (Figure 14). These results call for the attention of the policy makers towards the need to intensify the development of agricultural infrastructure in the region.

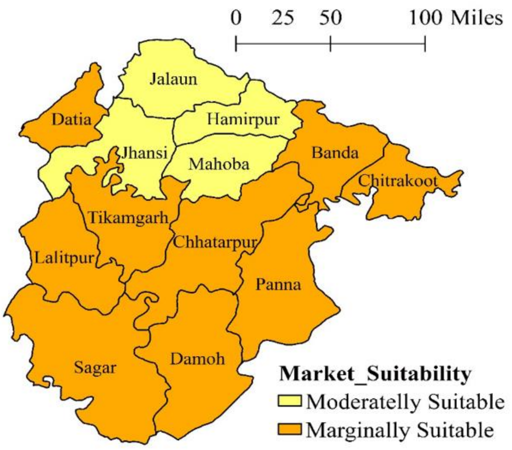


Figure 9: Market infrastructure

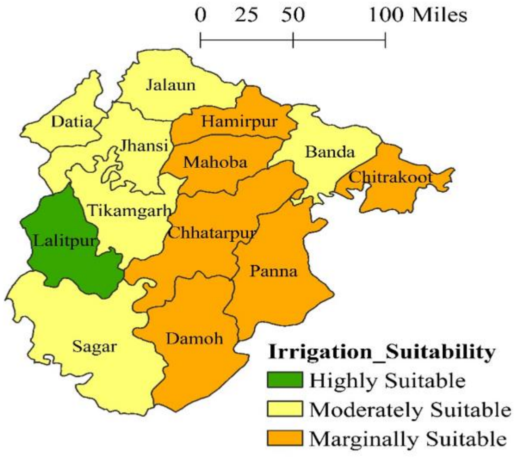


Figure 10: Irrigation infrastructure

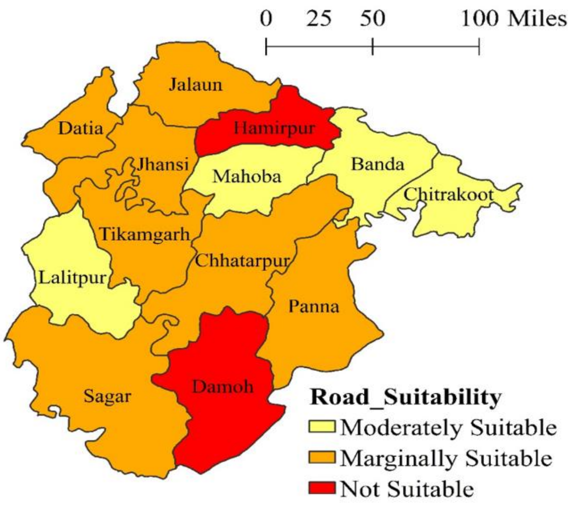


Figure 11: Roads

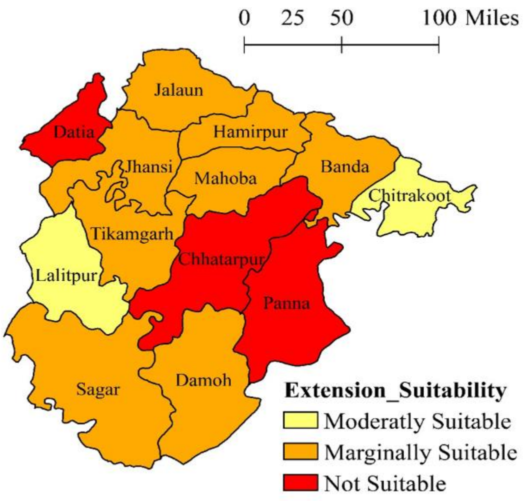


Figure 12: Extension

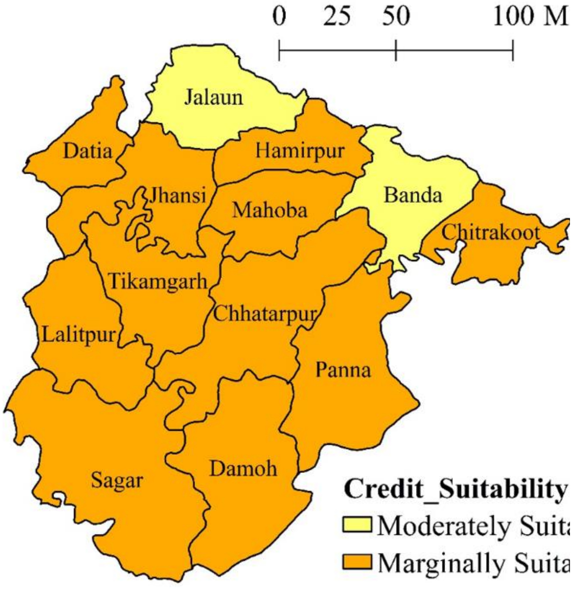


Figure 13: Credit

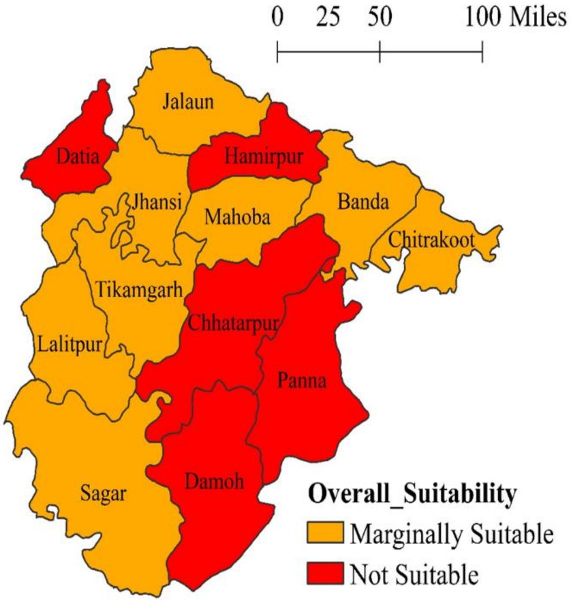


Figure 14: Composite suitability

Source: Authors

4. Conclusions and Policy Implications

The present study developed the methodology for determining agricultural infrastructural suitability status and validated the same for the country at both state levels as well as district level using a case study of districts from Bundelkhand region. The study contributes mainly by (i) developing methodological framework for quantification of agricultural infrastructure and (ii) estimation of adequacy level of agricultural infrastructure at state level and district level

The strong point of the proposed methodology is its simplicity and availability of data in public domain. The identified criteria for four infrastructural suitability classes are same for state level as well as district level. The methodology was implemented using omnipresent Excel spreadsheet. However, there is a scope for improvement in the methodology by developing the separate criteria for different agro-ecoregions.

Acknowledgement

This research was a part of ICAR Funded network project “Resource use planning for sustainable Agriculture” of ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi

References

Andersen, P. and Shimokowa, S. (2006). Rural infrastructure and agricultural development. *Annual Bank Conference on Development Economics*, Tokyo, Japan, 29-30 May, 2006; retrieved from: http://siteresources.worldbank.org/I NTDECABCTOK2006/ Resources/ Per_Pinstrup_Andersen_Rural_Infrastructure.pdf

Casella, M. I., and Schilling, B. (2017). Necessary role of extension in development of agricultural regulations. *Journal of Extension Education*, **55**(6), 6COM1; retrieved from: https://www.joe.org/joe/2017december/comm1.php.

Chang, C. W., Wu, C. R., Lin, G. T. and Lin, H. L. (2007). Evaluating digital video recorder using analytic hierarchy and analytic network processes. *Information Sciences,* **177**, 3383–3396.

Elsheikh, R., Shariff, A. R. B. M., Amiri, F., Ahmad, N. B., *et al*. (2013) Agriculture land suitability evaluator (ALSE): A decision and planning support tool for tropical and subtropical crops. *Computer and Electronics in Agriculture*, **93**, 98-110.

Fan, S. and Zhang, X. (2004). Infrastructure and regional economic development in rural China. *China Economics Review*, **15**, 203-214.

FAO (2017). The future of food and agriculture; trends and challenges. Food and Agriculture Organization, United Nations, retrieved from: http://www.fao.org /3/t8654e/t8654e04.htm.

Kamara, A. B. (2004). The impact of market access on input use and agricultural productivity: Evidence from Machakos district, Kenya, Agrekon. *Agricultural Economics Association of South Africa*, **43** (2**)**, 1-15, doi:10.22004/ag.econ.9485

NCCD (2015) All India cold-chain infrastructure capacity (Assessment of Status and Gap), Autonomous Body of Ministry of Agriculture and Farmers Welfare, Delhi, retrieved from: https://nccd.gov.in/PDF/CCSG_Final%20Report_Web.pdf

NITI. 2016. Human Development Report: Bundelkhand 2012. NITI Aayog-UNDP Project on Human Development: towards Bridging Inequalities.

Patel, A. (2010). Infrastructure for agriculture and rural development in India: Need for a comprehensive program and adequate investment. retrieved from: https://www.microfinancegateway.org.

Saaty, T. L. and Vargas, L. G. (2001) How to make a decision. *In Models, Methods, Concepts and Applications of the Analytic Hierarchy Process*, Springer, Boston, MA.

Shaloo Singh, R., Bisht, H., Jain, R., *et al*. (2022). Crop-Suitability Analysis Using the Analytic Hierarchy Process and Geospatial Techniques for Cereal Production in North India, *Sustainability*, **14**, 5246. https://doi.org/10.3390/su14095246

Rezaei, J. (2015) Best-worst multi-criteria decision-making method. *Omega*, **53**, 49–57. doi:10.1016/j.omega.2014.11.009.

Census (2011) available at https://censusindia.gov.in

Special Proceedings: ISBN #: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 189-204

Merchant Transactions Through Digital Payments

Bhavna Sharma[1] and Ashish Das[2]
[1]S. P. Jain School of Global Management, Sydney 2141, Australia
[2]Department of Mathematics, Indian Institute of Technology Bombay, Mumbai 400076, India

Received: 10 August 2022; Revised: 15 August 2022; Accepted: 17 August 2022

Abstract

To promote small ticket debit card and Unified Payments Interface (UPI) merchant transactions up to Rs. 2000, the government during the calendar years 2018 and 2019 made merchant discount rate (MDR) zero for the merchants. In contrast, effective January 1, 2020, the government made MDR zero for UPI and RuPay debit card transactions. Neither merchants nor the government paid the banks any MDR for such UPI and RuPay based merchant transactions. However, banks were allowed to impose MDR onto the merchants for every transaction using mastercard/VISA debit cards.

This report provides a follow-up of the conclusions drawn in an earlier IIT Bombay Technical Report (October 2021) “Merchant transactions through debit cards” http://dspace.library.iitb.ac.in/jspui/handle/100/36651. We assess the trends and progress of debit card usage since November 2019. Though the Covid-19 pandemic has distorted the trends in two spells (waves 1 and 2), our objective here is to get a general feel of the possible impact of MDR, on debit card usage. Based on data sources of RBI, DFS and, NPCI, we mine some interesting statistics.

With RBI’s prohibition for auto-debit payments in recurring transactions and the upcoming restrictions for storage of debit card details by merchants, it is bound to create enough friction for people to move towards alternative payment means like the UPI. It is expected that such risk-mitigating measures taken by RBI would act as a catalyst to see people migrating from debit cards to UPI.

1. Introduction

Debit cards and mobile Apps are provided by banks to facilitate account holders to withdraw cash at ATMs and carry out merchant transactions. The debit cards are primarily issued under one of the three card schemes – mastercard, VISA, or RuPay. Historically, card payments for merchant transactions had a well-defined revenue-generating structure for banks, where the revenue came from Merchant Discount Rate (MDR). Among the mobile Apps based payments system, Unified Payments Interface (UPI) is the most successful and accepted mode of digital payments.

1.1. The history of MDR regulation in India

In September 2012, the Reserve Bank of India (RBI) mandated a debit card MDR cap at 0.75%, for transactions valued up to Rs. 2000, and 1%, for transactions valued above Rs. 2000. This continued till November 8, 2016.

Corresponding Author: Bhavna Sharma
Email: bhavna.bs20dmu016@spjain.org

Immediately after the November 8, 2016 demonetization of the specified bank notes, the government instructed the banks to temporarily waive MDR imposed on merchants.

As an interim measure RBI, effective January 1, 2017, rationalized the MDR on debit cards. RBI set an MDR cap at (i) 0.25%, for transactions valued up to Rs. 1000, (ii) 0.5%, for transactions valued in excess of Rs. 1000 but not exceeding Rs. 2000, and (iii) 1%, for transactions valued in excess of Rs. 2000. RBI's new caps on debit card MDR were a substantial reduction to the RBI's pre-demonetization cap of 0.75% for transactions valued up to Rs. 2000.

Subsequently, effective January 1, 2018, RBI tweaked MDR rules claiming that such tweaks would encourage some small businesses to accept debit card payments. For businesses with annual turnover below Rs. 20 lakh, RBI capped the debit card MDR at 0.4% of transaction value or Rs. 200, whichever is lower. For others, *i.e.*, businesses with an annual turnover of Rs. 20 lakh or more, the debit card MDR was capped at 0.9% of the transaction value or Rs. 1000, whichever is lower. For QR-code-based debit card acceptance, the MDR caps were set 10 basis points lower than the physical point-of-sale (POS) and online debit card acceptance infrastructure.

In parallel, effective January 1, 2018, the government made MDR zero for the merchants and decided to bear the MDR cost for two years on all debit card and UPI transactions valued up to Rs. 2000. However, for the banks, the government fixed the POS- and online-based MDR at 0.4% for debit card and UPI transactions up to Rs. 2000. In effect, due to the government's intervention, RBI's decision to allow banks to charge up to 0.9% as debit card MDR for businesses with an annual turnover of Rs. 20 lakh or more (even for transaction amounts less than Rs. 2000), got overruled and the banks got only 0.4% as MDR for such sub Rs. 2000 ticket debit card transactions.

Corresponding to this government-provided-MDR of 0.4%, the interchange fixed by card payment networks had been 0.15%. Thus, RBI's MDR mandates could never get implemented since the government felt otherwise on small ticket transactions up to Rs. 2000, reducing the MDR to zero for all merchant categories.

In fact, the National Payments Corporation of India (NPCI) was the only card network to adopt an MDR that was lower than the MDR-cap set by RBI. The MDR pricing structure arrived at (effective October 2019) for RuPay debit cards had been 0.4% (0.3% when the transaction is QR-code based) for transactions up to Rs. 2000 and 0.6% (0.5% when the transaction is QR-code based) for transactions exceeding Rs. 2000, with a ceiling on MDR of Rs. 150 for any transaction. For transactions exceeding Rs. 2000, RuPay's 0.6% MDR applied only to businesses with an annual turnover of Rs. 20 lakh or more (vis-à-vis RBI's MDR cap of 0.9%).

1.2. The present avatar of MDR

Effective January 1, 2020, the government decided not to bear MDR any further on all debit card transactions valued up to Rs. 2000. In effect, due to this decision, RBI's mandate got re-invoked and banks got the leverage to charge MDR @ 0.9% or less from businesses with an annual turnover of Rs. 20 lakh or more for transactions of any value. Furthermore, for businesses with an annual turnover of less than Rs. 20 lakh, banks got the freedom to impose an MDR of 0.4% or less.

Nevertheless, the government simultaneously brought in a new law where RuPay debit card and UPI had been identified as a prescribed payment mode for which banks and system providers could no longer charge any fee to the merchants. Consequently, any charge, including the MDR, was no longer applicable on payments made through RuPay debit cards and UPI.

However, the government effective April 1, 2021, started to fund towards paying MDR to acquiring banks. All RuPay debit card transactions received 0.4% MDR capped at Rs. 100 for one year starting since April 2021. Also, low-value UPI (or BHIM-UPI) transactions (upto Rs. 2,000) received 0.25% MDR for one year starting since April 2021. These rates were further reduced for Industry Programmes.

Banks were reimbursed 95% of their claimed amount in each quarter of the scheme. To get the remaining 5% of each quarter, every bank was required to show at least a 50% y-o-y growth rate in the number of BHIM- UPI transactions and a 10% y-o-y growth rate in the number of RuPay debit card transactions at the end of last quarter of the scheme. In general, the banks were to achieve this target to pocket the 5% of the monetary support totaling around Rs. 60 Cr. The February 2022 budget announcement mentions about continuation of the financial support towards MDR in 2022-23.

2. Trends and Progress of Indian Payment Modes

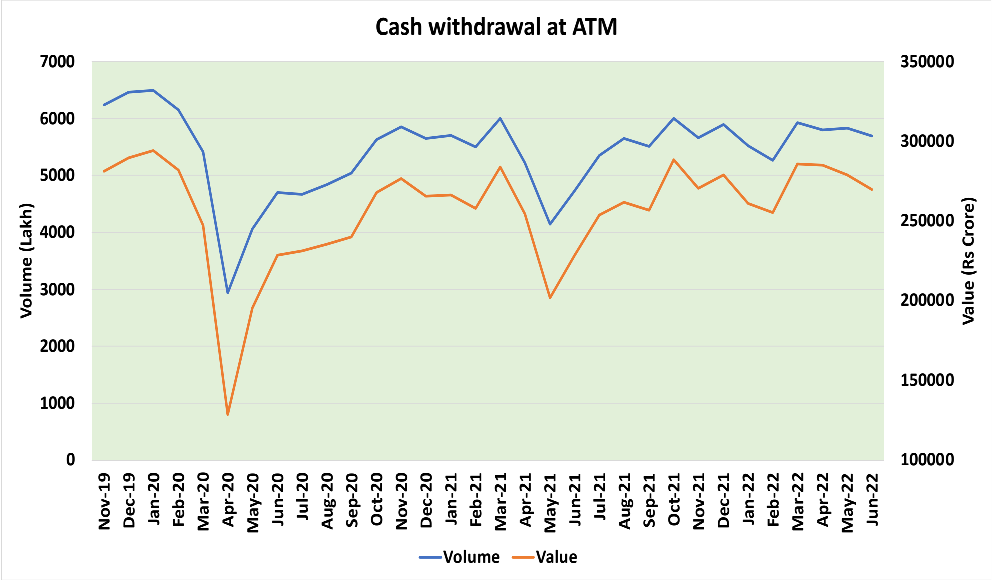
Debit cards are extensively used by bank account holders towards cash withdrawals at ATMs. Currently, RBI and banks are absorbing significant costs while providing cash as a prominent mode of payment. There is a need to arrest RBI's promotion of free excessive cash by mandating significant amounts of free ATM cash withdrawals. Such arrests would not only reduce cash handling costs for the banks but also save enough to support digital payments.

2.1. Cash from ATM

Cash is predominantly promoted in India with 8 to 10 free ATM withdrawals per month. This potentially amounts to bank's disbursement of at least Rs. one lakh of free cash per month to an individual holding a bank account. Starting November 2019, RBI, in its monthly ATM data dissemination, has started reporting cash withdrawals at ATMs using debit cards, instead of debit card usage at ATMs.

Figure 1 shows how the cash withdrawal at ATMs using debit cards behaved during the period November 2019 – June 2022. The periods March-May 2020 and April-June 2021 show the effects of Covid-19 waves.

But for RBI's mandate allowing a significant amount of free ATM cash withdrawals for many bank customers, technically speaking, banks would not have incurred such avoidable and non-remunerating expenses. There is nothing that RBI appears to have done as a deterrent, which strongly prompts a reduction of large amounts of free ATM cash withdrawals in a month. Digital payment modes are now amply available (especially, in tier I and II cities) where large and frequent cash is still in use. RBI advocating banks to charge a reasonable fee, in a tiered fashion, for total cash withdrawals in excess of a reasonable amount per month, could create enough deterrents. Such a move would allow generating desirable revenue for the banks to meet their cash handling costs and to provide support towards costs in maintaining digital payments infrastructure.



Source: RBI

Figure 1: Cash withdrawal at ATM using debit cards

2.2. Merchant transactions using debit cards

Starting November 2019, RBI, in their monthly bulletin, is disseminating bifurcated card transaction data comprising POS and ‘others’. For debit cards, ‘others’ primarily include E-com transactions, card-to-card transfers, and digital bill payments through ATMs. The same bulletin also publishes the Cash Withdrawal data. Furthermore, a combined POS cum E-com cum Cash Withdrawal data is provided in the monthly “Bank-wise ATM/POS/Card Statistics” that RBI releases. Based on these data sources, we derive the extent of E-com transactions. Table 1 and Table 2 provide the debit card transaction volume and value, respectively, for POS, E-Com, and Cash Withdrawals.

Table 1: Debit card transactions (Volume) and the extent of E-com transactions

Volume (Lakh)	POS	E-Com	Cash Withdrawal	POS + E-Com + Cash Withdrawal
Nov-2019	2483	1690	63	4236
Dec-2019	2634	1806	71	4512
Jan-2020	2587	1905	92	4584
Feb-2020	2456	1781	59	4297
Mar-2020	1925	1676	31	3632
Apr-2020	676	1371	36	2083
May-2020	1121	1528	36	2686
Jun-2020	1475	1518	28	3021
Jul-2020	1462	1691	31	3184
Aug-2020	1647	1733	28	3409
Sep-2020	1759	1741	29	3529
Oct-2020	1984	1940	29	3953
Nov-2020	2113	1645	32	3790
Dec-2020	2165	1622	35	3822
Jan-2021	2132	1560	31	3723
Feb-2021	2009	1430	20	3458
Mar-2021	2229	1526	19	3774
Apr-2021	1794	1412	19	3225
May-2021	1128	1422	18	2568
Jun-2021	1505	1434	7	2946
Jul-2021	1902	1496	6	3404
Aug-2021	2120	1461	6	3586
Sep-2021	2053	1430	5	3488
Oct-2021	2303	1543	4	3850
Nov-2021	2112	1288	4	3404
Dec-2021	2203	1289	4	3496
Jan-2022	1925	1261	2	3188
Feb-2022	1845	1104	2	2952
Mar-2022	2078	1190	2	3270
Apr-2022	2132	1174	2	3308
May-2022	2150	1163	2	3316
Jun-2022	2013	1079	2	3095

*POS + eCom + CashWithdrawal = Total number of financial transactions done by the debit card at POS terminals

*E-Com may include some failed transactions

Source: RBI/NPCI and authors' computation

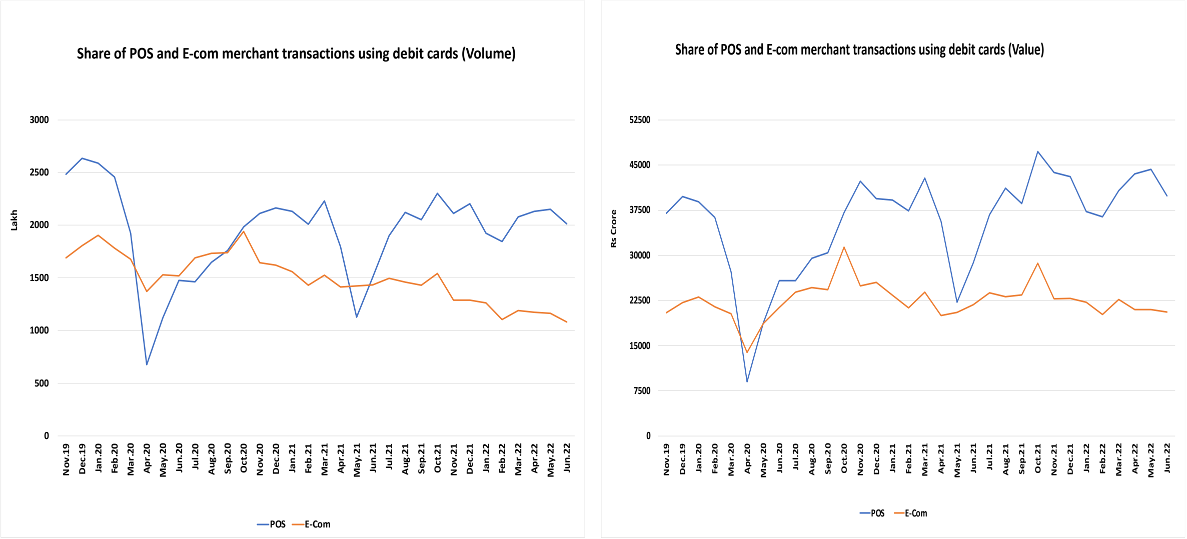
Table 2: Debit card transactions (Value) and the extent of E-com transactions

Value (Rs Crore)	POS	E-Com	Cash Withdrawal	POS + E-Com + Cash Withdrawal
Nov-2019	37007	20453	129	57590
Dec-2019	39740	22124	134	61998
Jan-2020	38907	23083	163	62154
Feb-2020	36258	21450	132	57841
Mar-2020	27238	20303	105	47646
Apr-2020	9005	13887	105	22998
May-2020	18814	18692	116	37622
Jun-2020	25788	21354	114	47256
Jul-2020	25821	23886	132	49840
Aug-2020	29525	24624	129	54277
Sep-2020	30422	24307	118	54847
Oct-2020	37110	31352	129	68591
Nov-2020	42289	24931	137	67357
Dec-2020	39437	25513	142	65093
Jan-2021	39189	23369	135	62693
Feb-2021	37414	21255	114	58783
Mar-2021	42816	23889	114	66819
Apr-2021	35621	20006	111	55739
May-2021	22195	20523	102	42820
Jun-2021	28743	21789	44	50576
Jul-2021	36764	23792	43	60599
Aug-2021	41177	23132	42	64352
Sep-2021	38591	23443	40	62074
Oct-2021	47226	28672	39	75937
Nov-2021	43750	22814	39	66603
Dec-2021	43062	22852	35	65949
Jan-2022	37274	22201	20	59495
Feb-2022	36376	20180	20	56576
Mar-2022	40770	22687	21	63479
Apr-2022	43530	21007	22	64558
May-2022	44273	21006	22	65300
Jun-2022	39877	20567	22	60466

*POS + eCom + CashWithdrawal = Total number of financial transactions done by the debit card at POS terminals
*E-Com may include some failed transactions

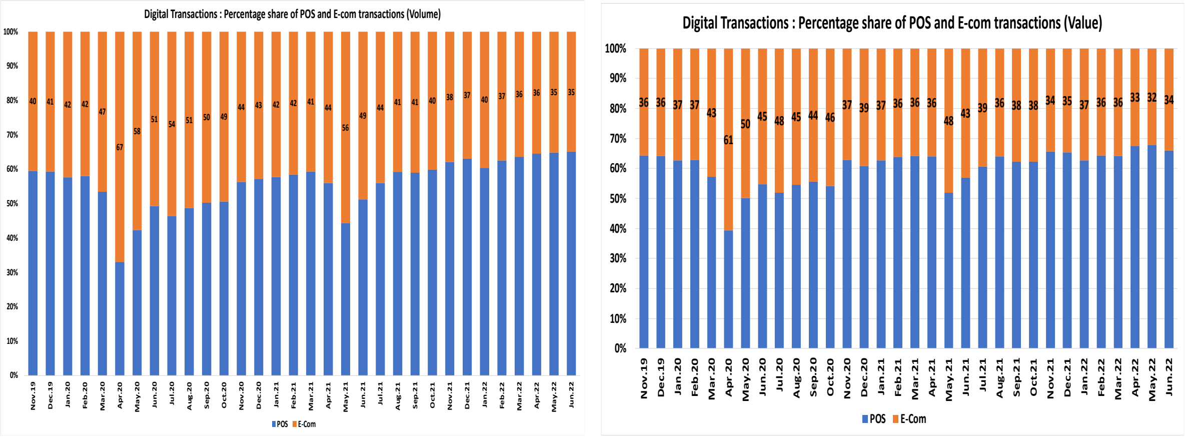
Source: RBI/NPCI and authors' computation

The extent of POS usage vis-à-vis E-Com indicates that POS still contributes more than E-Com. Figures 2 and 3 show the extent of POS vis-à-vis E-Com transactions.



Source: RBI

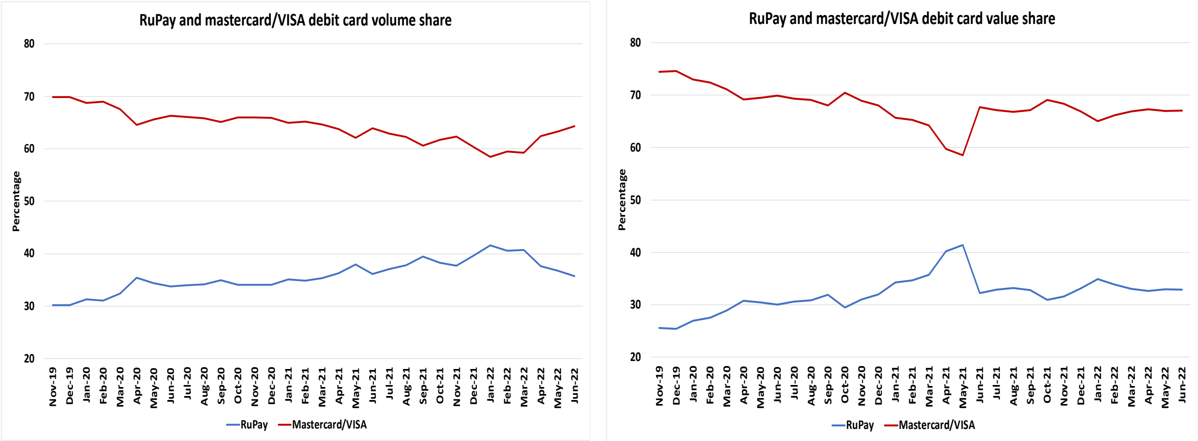
Figure 2: Share of POS and E-com merchant transactions using debit cards



Source: RBI

Figure 3: Percentage share of POS and E-Com transactions using debit cards

For the period November 2019 – June 2022, Figure 4 shows the extent of debit card transactions using RuPay and mastercard/VISA. Except for the recent months, both in volume and value terms, the share of RuPay transactions vis-à-vis mastercard/VISA had been consistently growing. With RuPay cards being primarily issued to accounts opened under the Pradhan Mantri Jan-Dhan Yojana (PMJDY), this relative growth in RuPay transactions could be attributed to the decline in discretionary expenditure of the urban population due to Covid-19 lockdowns. There was a fall in transactions by the non-PMJDY account holders (urban affluent/middle class) holding mastercard/VISA debit cards rather than RuPay debit cards, linked to PMJDY accounts.



Source: RBI and NPCI

Figure 4: Percentage share of RuPay and mastercard/VISA transactions

3. Performance of RuPay Debit Cards

Though the zero MDR for RuPay debit cards will lead to some savings for merchants, an important question remains as to whether it would serve the purpose of promoting card payments in the presence of merchants still being overburdened by the fee for accepting other debit cards. Note that for mastercard/VISA, effective January 1, 2020, the merchants no longer enjoy zero MDR on transactions up to Rs. 2000.

As it stands now, mastercard/VISA have been provided open grounds to see the promotion of their cards and demotion of RuPay cards. The lack of a level playing field would only give more earnings for mastercard/VISA at the cost of equitable promotion of RuPay.

If there is a revenue differential for banks between RuPay and mastercard/VISA, banks would always, in their commercial interest, tend to promote that card scheme which generates more revenue for them. DFS disseminates data to reflect the progress report of PMJDY. The time-series data is released every Wednesday updating information on the number of accounts opened under PMJDY, and RuPay debit cards issued. We have used this data to show monthly status. We take data points for every Wednesday of a month falling between the 1st through 9th of each month. Such monthly data points are used to reflect the status at the end of the previous month and are shown in Table 3.

During the one-year period early-July 2019 through end-June 2020, there had been a net issuance of about 376 lakh RuPay debit cards, and about 107 lakh accounts were added under the PMJDY. In contrast, for the subsequent two corresponding tenures, we see a subdued issuance of RuPay debit cards despite a fair number of PMJDY accounts being added.

Table 3: PMJDY accounts added vs RuPay debit cards issued

1-Year Periods	PMJDY accounts added (Lakh)	RuPay debit cards issued (Lakh)
Jul'19-Jun'20	376	107
Jul'20-Jun'21	282	161
Jul'21-Jun'22	331	72

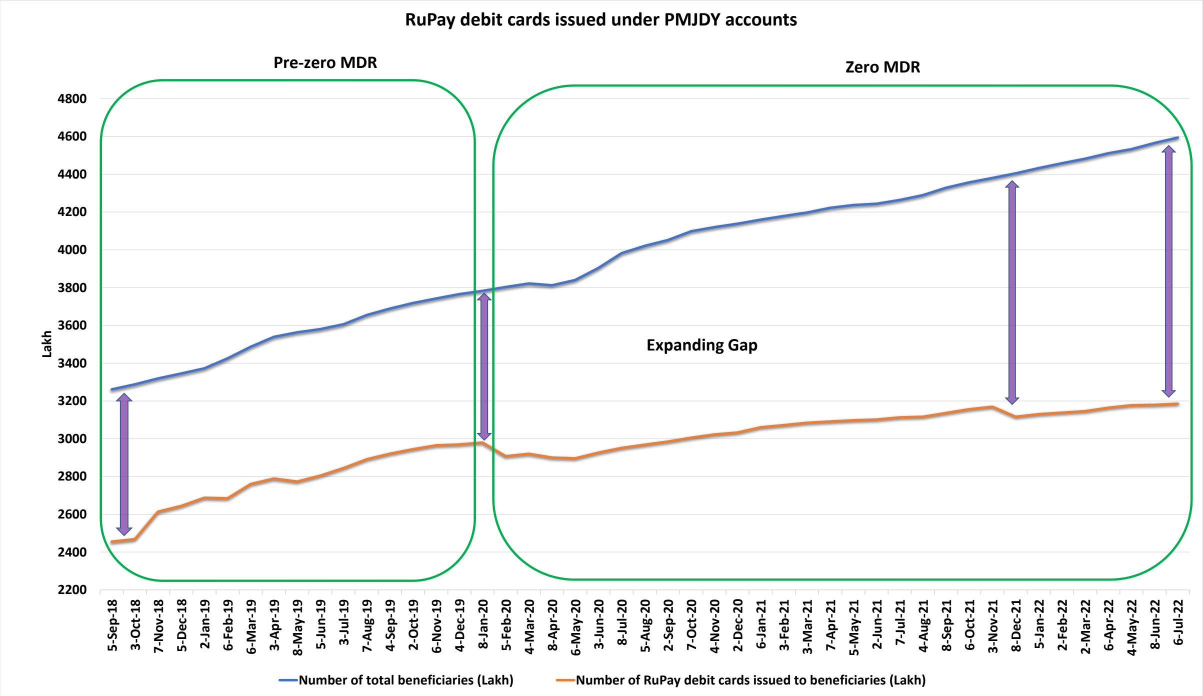
The expanding gap between the new PMJDY accounts added and the RuPay debit cards issued gets clearly reflected in Figure 5. Unless there are other extraneous causes (*e.g.* Covid-

19, *etc.*), a possible cause for such a trend could be that banks have deliberately moved away from RuPay to promote a card scheme that generates more revenue for them.

Table 4: Progress-report of the PMJDY

Month-end	Date	Number of total beneficiaries (Lakh)	Deposits in Accounts (Rs Crore)	Number of RuPay debit cards issued to beneficiaries (Lakh)
Aug-18	05-Sep-18	3261.32	82490.98	2455.76
Sep-18	03-Oct-18	3288.15	85378.60	2468.74
Oct-18	07-Nov-18	3319.38	84689.14	2614.14
Nov-18	05-Dec-18	3345.73	84814.54	2644.28
Dec-18	02-Jan-19	3372.82	87033.42	2687.97
Jan-19	06-Feb-19	3425.59	90217.40	2684.60
Feb-19	06-Mar-19	3487.23	93567.18	2760.37
Mar-19	03-Apr-19	3539.37	97665.66	2789.45
Apr-19	08-May-19	3563.56	98437.41	2773.39
May-19	05-Jun-19	3580.65	98473.68	2804.63
Jun-19	03-Jul-19	3606.19	100495.95	2844.92
Jul-19	07-Aug-19	3654.99	101879.34	2891.29
Aug-19	04-Sep-19	3688.91	102645.70	2920.83
Sep-19	02-Oct-19	3718.79	104698.00	2944.79
Oct-19	06-Nov-19	3742.26	106846.62	2964.93
Nov-19	04-Dec-19	3765.89	107904.11	2969.97
Dec-19	08-Jan-20	3782.80	111714.82	2979.83
Jan-20	05-Feb-20	3803.55	114569.13	2907.57
Feb-20	04-Mar-20	3822.12	117015.50	2920.43
Mar-20	08-Apr-20	3812.33	127748.43	2900.63
Apr-20	06-May-20	3840.51	131825.49	2896.55
May-20	03-Jun-20	3904.20	131339.59	2926.80
Jun-20	08-Jul-20	3982.42	131576.08	2951.62
Jul-20	05-Aug-20	4021.08	129719.63	2968.33
Aug-20	02-Sep-20	4052.07	129929.28	2985.34
Sep-20	07-Oct-20	4098.43	130360.53	3005.84
Oct-20	04-Nov-20	4120.33	131639.95	3022.42
Nov-20	02-Dec-20	4138.33	130932.33	3032.60
Dec-20	06-Jan-21	4159.88	135739.52	3060.49
Jan-21	03-Feb-21	4179.27	138425.52	3072.07
Feb-21	03-Mar-21	4197.34	139668.24	3084.69
Mar-21	07-Apr-21	4222.29	146084.99	3091.64
Apr-21	05-May-21	4237.00	143814.60	3097.89
May-21	02-Jun-21	4244.18	144525.39	3101.61
Jun-21	07-Jul-21	4263.94	144277.46	3112.95
Jul-21	04-Aug-21	4289.49	143834.57	3116.60
Aug-21	08-Sep-21	4328.89	145050.76	3136.36
Sep-21	06-Oct-21	4357.21	145811.5	3156.18
Oct-21	03-Nov-21	4380.78	148935.05	3169.07
Nov-21	08-Dec-21	4405.04	147812.21	3116.61
Dec-21	05-Jan-22	4432.71	154916.47	3130.10
Jan-22	02-Feb-22	4458.28	157649.71	3138.07
Feb-22	02-Mar-22	4482.11	160839.87	3145.96
Mar-22	06-Apr-22	4511.16	167812.71	3163.98
Apr-22	04-May-22	4532.98	166803.08	3177.36
May-22	08-Jun-22	4566.08	170010.91	3179.67
Jun-22	06-Jul-22	4594.57	171802.13	3185.02

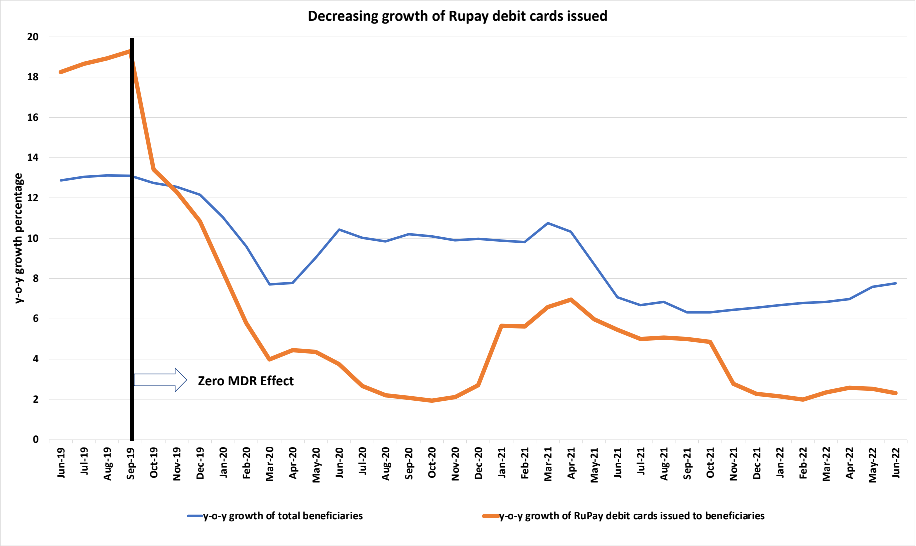
Source: Data submitted to DFS, Ministry of Finance, by Public Sector Banks, Regional Rural Banks and Major Private Sector Banks



Source: Table 4

Figure 5: RuPay debit card issued against the PMJDY accounts opened

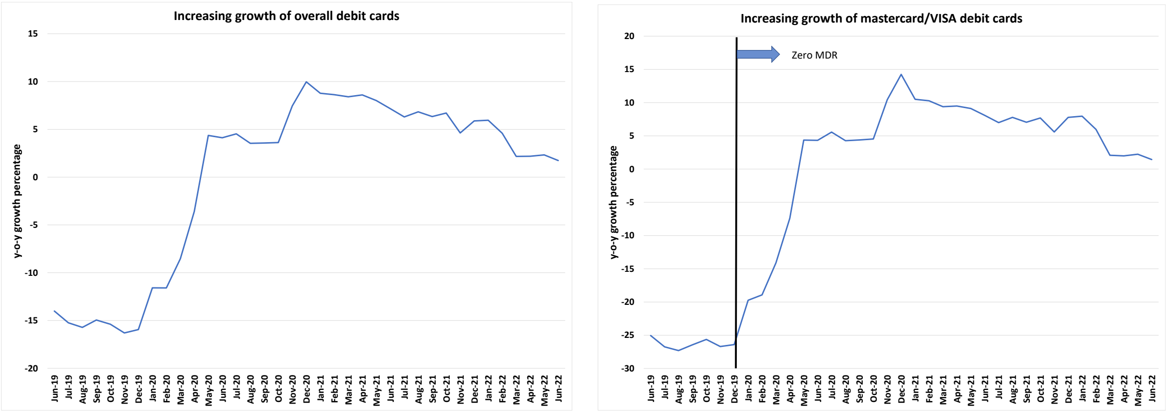
We now look at the y-o-y growth of RuPay cards issued and new accounts opened under the PMJDY. Starting early-September 2019, when the y-o-y growth of RuPay debit cards issued under the PMJDY was 18.9%, there has been a significant reduction in the y-o-y growth of Rupay debit cards. The early-September 2020 y-o-y growth has drastically reduced to 2.2%. This recovered a bit with the early-September 2021 y-o-y growth being 5.0% but has now returned to the older trend of descent with early-July 2022 y-o-y growth being 2.3%. The consistent decrease in the y-o-y growth since September 2019 is clearly depicted in Figure 8. In contrast, if one looks at the y-o-y growth of new PMJDY accounts opened, the figures for early-September 2019, early-September 2020, early-September 2021, and early-July 2022 are 13.1%, 9.8%, 6.8%, and 6.8%, respectively. Thus, though y-o-y growth of new PMJDY accounts opened had been relatively consistent, in comparison, the drastic downward y-o-y growth of RuPay debit cards issued is possibly a zero MDR effect.



Source: Table 4 and authors' computation

Figure 6: y-o-y growth of RuPay debit cards issued against PMJDY accounts opened

To assess the correct status of RuPay debit cards, we compare the extent of RuPay debit cards issued (under the PMJDY) against the overall debit cards outstanding. We consider the period since June 2019. Unlike a general decline in y-o-y growth of Rupay debit cards issued under the PMJDY, Figure 7(a) shows that there is a sharp uptick in y-o-y growth of overall debit cards outstanding (includes mastercard, VISA, and RuPay). Though implicit, inherently this showcases that there is a consistently increasing y-o-y growth of mastercard/VISA (Figure 7(b)) as against a consistently decelerating y-o-y growth of RuPay debit cards. During the recent months, the y-o-y growth of mastercard/VISA debit cards is above 2% with RuPay showing similar growth but this could still be due to their dwindling past numbers. We say this since, between the share of RuPay and mastercard/VISA, mastercard/VISA still has a significant share as we see below.

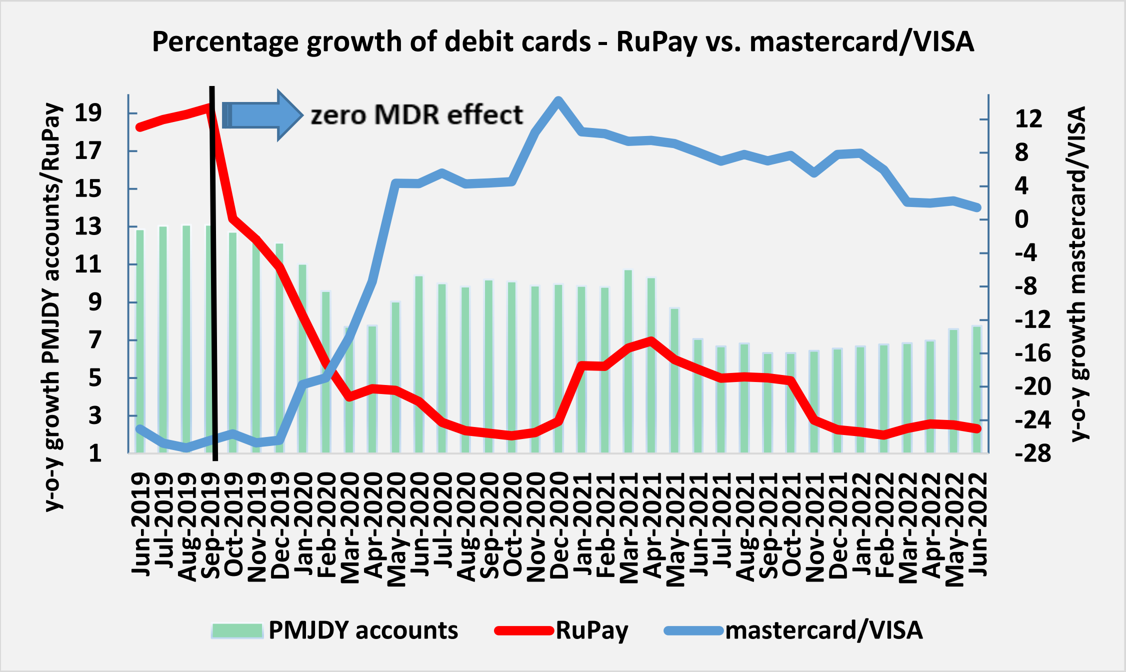


(a) (b)

Source: RBI/DFS and authors' computation

Figure 7: y-o-y growth of (a) overall debit cards, and (b) mastercard/VISA debit cards

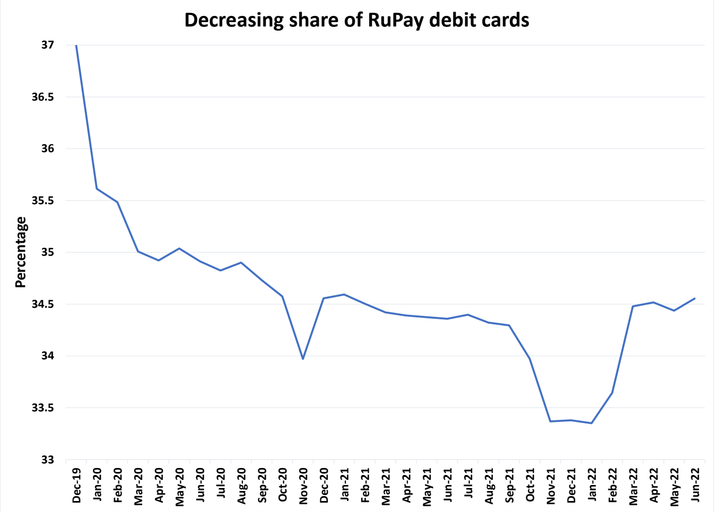
A comparative picture of growth in PMJDY accounts and RuPay/mastercard/VISA is presented in Figure 8.



Data Source: Table 4 and Figure 7

Figure 8: A comparative picture of growth in PMJDY accounts and RuPay/mastercard/VISA

Finally, in Figure 9, we show that the percentage share of RuPay debit cards issued under the PMJDY among the total debit cards outstanding has been consistently declining since December 2019. These data and charts showcase the negative impact of zero MDR on RuPay debit cards issued (at least for those issued under the PMJDY) and the unintended thrust that it provided to mastercard/VISA.



Source: RBI/DFS and authors' computation

Figure 9: Share of RuPay debit cards issued (PMJDY) among the total debit cards outstanding

With merchants' preference for debit card acceptance not being quite guided by debit card MDR alone, for the payments industry, the supply of RuPay debit cards by the producers (banks) would be impacted due to differentiated administered pricing on MDR. The administered MDR pricing for RuPay is zero while for mastercard/VISA it is 0.4-0.9%. In presence of such an imbalanced administered pricing, and with merchants' preference for card acceptance also being guided by considerations other than debit card MDR, the supply of debit cards by banks gets appositely restricted to products (mastercard/VISA) that generate higher MDR (interchange) or revenue for them, and as a result, the same is being promoted by the banks. Because of the twin administered pricing, competition is not resulting in an identical spread of the debit card products, leading to a welfare loss to the society in form of (i) merchants' and consumers' depleting choice to harness the benefits of zero MDR on RuPay (due to lack of adequate supply of RuPay debit cards), and (ii) banks no longer having the capacity to produce RuPay as a means for merchant payments (due to commercial considerations).

The debit card MDR alone has a limited impact in determining a merchant's choice or preference for acceptance of a particular card (over cash) as it is just one of the few other costs associated with card acceptance. Where cash is an alternative and where the merchant attaches significance to 'cost to merchant', even if debit card MDR is zero, the merchant should think twice to agree to accept cards. This is so since he has to pay for (i) the high credit card MDR @ 2-4% and (ii) the high monthly rentals in the range of Rs. 200-600 for point-of-sale (POS) terminals, etc. Thus, the preference for debit card acceptance by the merchant is not quite guided by debit card MDR alone since there exist other deterrents, for many small and medium merchants.

For the payments industry, the supply of RuPay debit cards by the producers (banks) would be impacted due to differentiated administered pricing on MDR. The administered MDR pricing for RuPay is zero while for mastercard/VISA it is 0.4-0.9%. In presence of such an imbalanced administered pricing, and with merchants' preference for card acceptance also being guided by considerations other than debit card MDR, the supply of debit cards by banks gets appositely restricted to products (mastercard/VISA), which generate higher MDR (interchange) or revenue for them, and as a result, the same is promoted by the banks. Because of the twin administered pricing, competition is not resulting in an identical spread of the debit card products, leading to a welfare loss to the society in form of (i) merchants' and consumers' depleting choice to harness the benefits of zero MDR on RuPay (due to lack of adequate supply of RuPay debit cards), and (ii) banks no longer having the capacity to produce RuPay as a means for merchant payments (due to commercial considerations).

4. Current Policy Impacts

Starting April 1, 2021, the government funded around Rs. 1,300 crore towards paying MDR to acquiring banks. All RuPay debit card transactions received 0.4% MDR capped at Rs. 100 for one year starting since April 2021. Also, low-value BHIM-UPI transactions (up to Rs. 2,000), received 0.25% MDR for one year starting since April 2021. For Industry Programmes, these rates were further reduced.

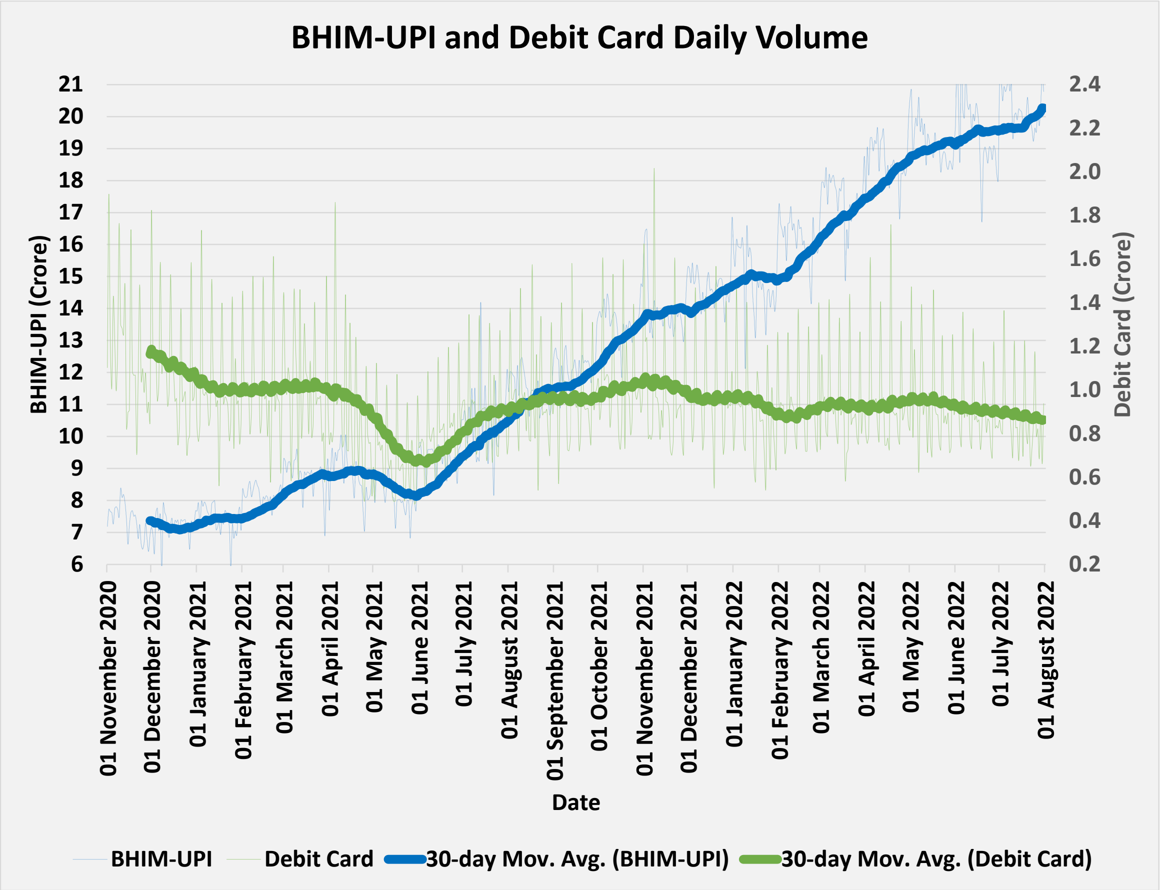
Each bank was reimbursed 95% of its claimed amount in each quarter of the scheme. To get the remaining 5% of each quarter, every bank was required to show at least a 50% y-o-y growth rate in the number of BHIM- UPI transactions and a 10% y-o-y growth rate in the number of RuPay debit card transactions at the end of last quarter of the scheme. In general, the banks were to achieve this target to pocket the 5% of the monetary support totalling around Rs. 60 Cr. The February 2022 budget announcement mentions about continuation of the support in 2022-23.

4.1. Debit cards unparalleled to BHIM-UPI

Mastercard/VISA debit cards generating revenue through MDR may be unfair for RuPay, but this is just the beginning. It is likely that the present approach is only a temporary measure to test how the card payments market responds. We see that the asset-lite mobile phone-based BHIM-UPI is cannibalising the payments arena.

Acceptance of cards is usually via a combined credit/debit/pre-paid card acceptance product. The MDR for credit cards is not regulated (reigns as high at 2% to 4%) and thus is an expensive proposition for merchants. Merchants may choose to enable themselves for card acceptance, however, it may not be advisable for all small and medium merchants to go for the relatively expensive card-based acceptance when, as an alternative to cash, there exists BHIM-UPI as a payment mode. BHIM-UPI does not cost the merchants anything under extant laws. Moving the country away from cash would hinge on having an equivalent, simple, convenient, and secure digital alternative, like the BHIM-UPI (in this era of mobile phones). In July 2022, there were 629 crore BHIM-UPI transactions amounting to a total of Rs. 10.63 lakh crore.

RBI has provided the daily data from payment networks for BHIM-UPI and debit cards. Comparing the trends since November 2020, we see that BHIM-UPI volumes have consistently increased, while the debit card transaction volumes are showing signs of decline. Currently, BHIM-UPI is hitting over 20 crore daily transactions, as against less than one crore daily off-us debit card transactions (Figure 10).



Data Source: RBI

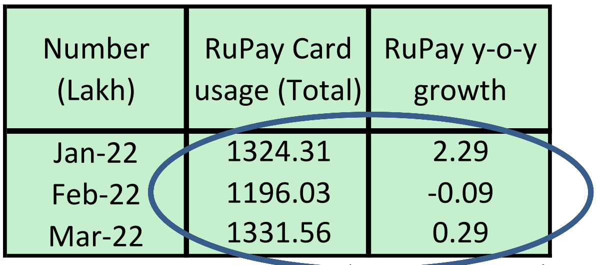
Figure 10: Standing on the shoulders of a giant – The rising BHIM-UPI against debit cards

As discussed, UPI is currently cannibalising the payment arena, which can be seen in Figure 10. BHIM-UPI is having an almost constant y-o-y growth of over 100%, and thus, reaching the 50% y-o-y growth rate by the end of the last quarter requirement was an easy task to achieve. As expected, banks have been successful in meeting this requirement reaching 98% y-o-y growth at the end of the last quarter of the scheme (Jan-Mar 2022).

4.2. RuPay debit card transactions

Though the banks were able to successfully achieve the BHIM-UPI target set by the government, however, to receive the 5% remaining remuneration, they had to also achieve the target of 10% y-o-y growth in the number of RuPay transactions. On average, the banks were unsuccessful in achieving this target and closed the final quarter with the y-o-y growth of approximately 0.3% in terms of RuPay (Table 5). This was much below what was set out to be achieved and thus, the banks were unable to receive the remaining 5% remuneration.

Table 5: y-o-y RuPay growth



Number (Lakh)	RuPay Card usage (Total)	RuPay y-o-y growth
Jan-22	1324.31	2.29
Feb-22	1196.03	-0.09
Mar-22	1331.56	0.29

Source: RBI/NPCI and authors' computation

An estimate of the funding provided by the government towards BHIM-UPI and RuPay debit card merchant transactions for FY22 has been worked out in Das and Das (2022).

5. Ending Remarks

The Indian government's intervention to provide monetary support to the banks to run the BHIM-UPI platform has led to banks recovering their cost to run the BHIM-UPI system. With BHIM-UPI showing a consistent y-o-y growth of 100%, this has its effect on other digital payment modes, including the RuPay Debit card, where we see people migrating to BHIM-UPI as a more convenient means of payment. With the increasing trend of smartphones, one can expect more and more people to be enabled on BHIM-UPI. In this scenario, as seen in the past year, it is unlikely to see in the future a 10% y-o-y growth for RuPay debit card transactions. It is thus recommended that the Indian government revisits the incentive reimbursement clause of a 10% y-o-y growth for RuPay Debit card transactions.

References

Concept Paper on Card Acceptance Infrastructure. RBI, March 8, 2016.

Das, Amogh and Das, Ashish (2022). *Unified Payments Interface – A Giant in the Digital Payments Space*. IIT Bombay Technical Report. August 1, 2022.

Das, Ashish (2016). *Promotion of Payments through Cards and Digital Means*. IIT Bombay Technical Report. April 24, 2016.

Das, Ashish and Das, Praggya (2016). *Sanitising Distortions in Digital Payments*. IIT Bombay Technical Report. November 28, 2016.

Das, Ashish (2020a). *Deviating from the BHIM-UPI Law*. IIT Bombay Technical Report. August 24, 2020.

Das, Ashish (2020b). *Discriminatory Approach for RuPay Debit Cards: Some Suggestions for Corrective Measures*. IIT Bombay Technical Report. January 7, 2020.

Promotion of Payments through Cards and Digital means. Government of India (Ministry of Finance), February 29, 2016.

Report of the Committee to review the ATM Interchange Fee Structure. Report of the RBI constituted committee. October 2019.

Report of the Committee on the analysis of QR (Quick Response) Code. RBI. July 22, 2020.

Sharma, Bhavna and Das, Ashish (2021). *Merchant Transactions Through Debit Cards*. IIT Bombay Technical Report. October 1, 2021.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 205-213

Are Some People More Likely Than Others to Develop Disease When Exposed to an Infection?

Shyamal Peddada
Biostatistics and Bioinformatics Branch
Eunice Shriver National Institute of Child Health and Human Development (NICHD), NIH
Bethesda, MD

Received: 14 August 2022; Revised: 20 August 2022; Accepted: 24 August 2022

Abstract

When a pandemic strikes a population, not everyone is similarly affected. Some people are more susceptible to a disease than others, and the question is: why? A person with a “good” innate immune system potentially has better defense mechanisms to confront or respond to an infection than those who do not have good immune system. Understanding the underlying heterogeneity in immunity is a question of scientific interest. There are potentially numerous reasons for heterogeneity within a population, such as, genetics, environment, socioeconomics, and so on. It is well-established that microbiome plays an important role in inflammation and immune response. In this article we summarize the findings of Chen *et al*. (2021), on the role of microbiome in the HIV infection and AIDS, who demonstrated that changes in gut microbiome takes place months before the onset of HIV infection and years before the HIV patients progressed to develop AIDS. Given the findings of that study, one may speculate a similar phenomenon for other infectious diseases, such as the novel coronavirus (COVID-19).

Key Words: Alpha Diversity; Beta Diversity: Differential Abundance; HIV infection; Infectious diseases; Microbiome.

1. Background

Human health is dependent on complex interactions between our genome, the external factors, and the internal environment over time. The external environment broadly includes chemicals to which we are exposed, the air we breathe, our water supply, diet, physical activity, and other factors. The internal environment includes hormones, the microbiome and microbial biproducts such as the short chain fatty acids, various metabolites, and so on. The interaction of gene by external factors over time as been extensively studied in the literature. During the past decade, researchers began to recognize the important role played by the microbiome on human health. Humans are estimated to have 45.6 million bacterial genes in oral and gut microbiome alone (*cf*., Tierney *et al*., 2019), which is about 2000-fold more than human genes. Therefore, the microbiome is sometimes referred to as the "second genome", or another "organ" of human body

Corresponding Author: Shyamal Peddada
Email: shyamal.peddada@nih.gov

(*cf*., O'Hara and Shanahan, 2006, Relman and Falkow, 2017, Hurst, 2017). It is now well-established that the microbiome is involved in metabolism, immune response, and inflammation. Gut microbiome also regulate mood (Sampson and Mazmanian, 2015, Steenbergen *et al*., 2015), anxiety, cognition, pain, aging, and a host of other factors of human health and behavior. It is not surprising that numerous diseases such as allergies, asthma, obesity (Turnbaugh *et al*., 2009), Crohn's disease (Gevers *et al*. 2014), inflammatory bowel diseases, and HIV are associated with alterations in the microbiome (Lozupone *et al*., 2013).

The role of the microbiome on human health and disease has been demonstrated for many other diseases. For example, in the general population, there is greater microbial diversity among people with no asthma than those with asthma, with a greater abundance of taxa such as *Faecalibacterium prausnitzii, Sutterella wadsworthensis* and *Bacteroides stercoris*, and microbial byproducts such as short chain fatty acids, acetate and butyrate (Wang *et al*., 2018). Conversely, there is a greater abundance of pro-inflammatory bacteria such as *Clostridium bolteae*, *Clostridium ramosum*, *Clostridium spiroforme* and *Eggerthella lenta* among people with asthma (Wang *et al*., 2018). Gut microbiome is also involved in the production of immunoglobulin E (IgE) (Cahenzli *et al*., 2013) and are associated with lung functions such as forced expiratory volume (FEV) (Begley et *al*., 2018). Furthermore, as seen in recent literature (Vila *et al*., 2020, Weersma and Zhernakova, 2020), medications potentially affect the gut microbiome, which in turn impacts the efficacy of treatments (Weersma and Zhernakova, 2020).

We use the terms "taxa", "bacteria", and "microbe" interchangeably. Often the terms "microbiome" and "microbiota" are used in the literature interchangeably although these two are distinct terms. Microbiota refers to the taxa describing various organisms whereas microbiome is a broader term that includes microbiota and their genes.

Given the recent, and ongoing COVID-19 pandemic, some natural questions to ask are: What is the effect of the microbiome on infectious diseases such as COVID-19, HIV, etc.? What caused some people to be more susceptible to acquiring a disease than others? Are there generally significant differences between the gut microbiome of people who acquire a disease and those who do not?

The role of the microbiome in infectious diseases is well-documented in the literature, with several review articles and Perspectives written on this subject in recent years, e.g., Harris *et al*., (2017), Libertucci and Young (2019), Cai *et a*l., (2021), Giovanni *et al*. (2021), Harper *et al*. (2021), and Hussain *et al*. (2021). Interactions between the human microbiome and pathogens, and the role of microbiome in stimulating the host immune system to defend against pathogens and hence protect against infections is well-characterized in these review articles. Interest in this area has increased with the recent pandemic to understand the mechanistic role of the microbiome for developing prebiotics, probiotics, fecal microbiome transplantation (FMT), and other treatments for infectious diseases. For example, according to Yeoh *et al*. (2021) the gut microbiota plays an important role in modulating markers of host immune system such as the cytokines and inflammatory markers, and thus plays a crucial part in lessening the severity of COVID-19.

Although the existing literature suggests a change in the gut microbial composition after the onset of a disease among infected people, the question remains whether people with gut dysbiosis are prone to develop a disease if exposed to an infection. This question is difficult to answer because it is not a common practice to collect stool samples in the general population to investigate who in the future gets a disease and who does not. The prospective Multicenter AIDS Cohort Study (MACS) provided an opportunity for Chen *et al.* (2021) to answer the above question in the context of HIV infection and AIDS. During the HIV pandemic of the 1980's, the MACS was established at four centers in the United States, namely, Pittsburgh, Baltimore, Chicago, and Los Angeles. The study recruited men who had sex with men (MSM) before any of the recruited individuals seroconverted, *i.e*., before becoming HIV positive (HIV+). This cohort allowed Chen *et al.* (2021) to investigate differences in gut microbial compositions between a group of men who became HIV+, or developed AIDS after becoming HIV+, with those who remained HIV negative (HIV-). In Section 2 we briefly describe the microbiome data and methodologies for analyzing those data, and we summarize in Section 3 the findings of Chen *et al*. (2021). Concluding remarks are provided in Section 4.

2. A Brief Overview of Statistical Methodology

Observed microbiome data are count data derived from sequencing a specimen obtained from an ecosystem, in the present case the human gut. Two popular technologies used to generate microbiome data are the technology based on 16s ribosomal RNA (16s rRNA) and the shotgun metagenomics. Since 16s rRNA is highly conserved in almost all bacteria and its function has not evolutionarily changed (*cf*. Janda and Abbott, 2007), it is commonly used by researchers conducting microbiome surveys. Microbial count data obtained by using 16s rRNA are commonly referred to as "16s data." Although 16s rRNA technology is specific for bacterial profiling, the shotgun metagenomics surveys not only bacteria but also sequences all genomic DNA. Thus, in addition to high taxonomic resolution, at the level of species and strain, the shotgun metagenomics allows host DNA inference, functional profiling, and metabolic pathway analysis. Since the shotgun metagenomics method currently is far more expensive than 16s rRNA, some researchers apply informatic tools, such as PICRUSt (Langille *et al*., 2013), to 16s data for functional profiling.

In the following description, our focus is on the analysis of 16s data. The observed microbiome data are a matrix of counts, with rows representing various taxa and columns representing the samples. Two important characteristics of these microbiome data are that
(1) typically, a very large proportion of the entries of this matrix are zero. The zero entries may arise for several reasons as detailed in Kaul *et al*. (2017); and (2) for reasons explained below, the observed counts are compositional, *i.e*., reside in a simplex.

A variety of statistical parameters are considered when comparing two or more experimental groups. Common parameters of interest are alpha diversity, beta diversity, taxon abundance, and taxon relative abundance. The alpha diversity parameter measures the diversity within samples. Numerous measures of alpha diversity appear in the microbiome literature; some examples include Shannon's entropy, the Gini-Simpson index, and the Chao1 index. The beta diversity parameter measures diversity in taxa between samples or between groups and, similar to alpha diversity, the microbiome literature provides a variety of measures of beta diversity due to differing concepts of distances between samples. For a review of these measures of diversities, we refer to Weiss *et al*. (2017) and references therein.

In addition to measures of diversity, researchers are interested in identifying taxa that are differentially abundant between experimental groups. This area of research, known as *differential abundance analysis*, is centered on testing the null hypothesis of equality of (relative) abundance of a taxon between two or more groups against various alternative hypothesis.

Suppose there are G experimental groups and n_i subjects in the i^{th} experimental group $i = 1,2,\dots,G$, In a unit volume of an ecosystem of the k^{th} subject, $k = 1,2,..,n_i$ in the i^{th} experimental group $i = 1,2,\dots,G$, let A_{ijk} denote the unobservable true abundance of the j^{th} taxon, $j = 1,2,..,m$. Using the 16s or shotgun metagenomics technologies we obtain O_{ijk}, the observed counts of the j^{th} taxon, on the k^{th} subject in the i^{th} experimental group.

Define $T_{ik} = \sum_{j=1}^{m} O_{ijk}$, sometimes called the *library size* of the k^{th}subject in the i^{th} experimental group. Due to sample collection methods and technology, T_{ik}, in practice, is highly variable among subjects. Also, within each subject k, as the sample collection and preparations change, the observed counts O_{ijk} are assumed to change proportionally. Thus, the observed counts O_{ijk} within each subject are compositional and hence are in a simplex. Statisticians often convert these observed counts to relative abundances, $R_{ijk} = O_{ijk}/T_{ik}$, so that $\sum_{j=1}^{m} R_{ijk} = 1$. One may view the observed counts O_{ijk} as an unknown fraction (or multiple) c_{ik} of the true abundance A_{ijk}. The population abundance parameter of interest is $\mu_{ij} = E\left(A_{ijk}\right), i = 1, 2, \dots, m, j = 1,2,\dots,G$. Unfortunately, this parameter μ_{ij}cannot be estimated unbiasedly unless the bias due to the nuisance parameter c_{ik} is eliminated. Suppose that $\lambda_{ijk} = A_{ijk}/\sum_{j=1}^{m} A_{ijk}$ denotes the relative abundance of the j^{th} taxon, $j = 1,2,\dots,m$, on the k^{th} subject, $k = 1,2,\dots,n_i$, in the i^{th}experimental group, $i = 1,2,\dots,G$, then λ_{ijk} can be estimated by the observed relative abundance of R_{ijk}. For this reason, as an alternative to μ_{ij}, researchers sometimes are interested in making inferences about the mean relative abundance $\theta_{ij} = E(\lambda_{ijk})$. Although it is natural to study the relative abundance because it does not involve the nuisance parameter c_{ik}, but from a clinical or scientific point of view, the relative abundance parameter may be difficult to interpret. Consider the two ecosystems in the toy example provided in Tables 1a and 1b. Table 1a consists of the abundances of five taxa in the two ecosystems. The counts of the first four taxa are identical across the two ecosystems, with only Taxon 5 differentially abundant between the two ecosystems. Often researchers are interested in identifying Taxon 5, the differentially abundant taxon. However, if one were to consider relative abundances (Table 1b), all five taxa have differential relative abundances between the ecosystems. Although it is mathematically correct that the relative abundances differ between the ecosystems, clinically or scientifically it may not be a useful piece of information. Thus, there are reasons to prefer to test for equality of abundances rather than the equality of relative abundances of taxa between ecosystems. However, it is a challenging problem to test for equality of abundances between two or more ecosystems because of the nuisance parameter mentioned above. Several methods have been proposed in the literature to eliminate the bias due to the unknown nuisance parameter. Some methods commonly used in the literature include

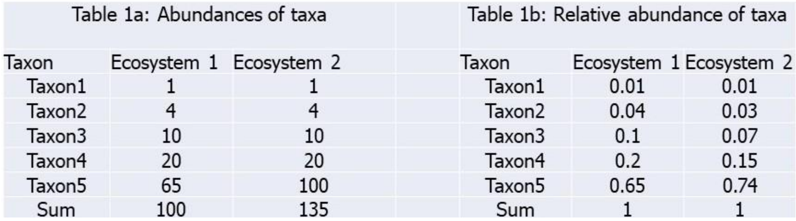


Table 1a: Abundances of taxa

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	20
Taxon5	65	100
Sum	100	135

Table 1b: Relative abundance of taxa

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	0.01	0.01
Taxon2	0.04	0.03
Taxon3	0.1	0.07
Taxon4	0.2	0.15
Taxon5	0.65	0.74
Sum	1	1

ALDEx2 (Fernandes *et al*., 2014), ANCOM (Mandal *et al*., 2015), ANCOM-BC (Lin and Peddada, 2020), and RNA-seq-based methods such as edgeR (Robinson *et al*., 2010), DESeq2 (Love *et al*., 2014).

Although the above methods are among the popular methods for conducting differential abundance analysis, they are rapidly getting outdated with several new methods introduced in the literature on a regular basis (Zhou *et al*., 2021, Hu *et al*., 2022). Recently, Nearing *et al*. (2022) conducted an exhaustive numerical study involving 38 different 16s data sets to evaluate different available methods for differential abundance analysis and they concluded that ALDEx2 and ANCOM-II produce the most consistent results. However, these authors did not include LiNDA (Zhou *et al*., 2021) or LOCOM (Hu *et al*., 2022) which were perhaps not available at the time Nearing *et al*. (2022) was published.

3. The Findings of Chen *et al*. (2021)

Using the stool and blood samples collected from men during their first clinical visit in the 1980's by MACS, Chen *et al*. (2021) investigated the differences in microbial compositions of men who developed HIV infection at a future time and those who did not. Furthermore, they also investigated differences in the microbial compositions among those who developed AIDS at different time points in the future.

The study consisted of 265 participants who were HIV negative (*i.e*., did not seroconvert, denoted as negative controls (NC)) at the beginning of the study, and among these 156 remained HIV- but 109 seroconverted, *i.e*., became HIV+, within about six months after the first samples were collected (denoted as seroconverters (SC)). Of the 109 who seroconverted, 32 of them developed AIDS within 5 years, 31 developed AIDS between 5 to 10 years and 46 took more than 10 years to develop AIDS. The data are summarized in the schematic provided in Figure 1. Chen *et al*. (2021) compared the SC and the NC groups using their microbiome data collected at the first visit when all of them were HIV-, the SC group did not yet seroconvert. Similarly, they compared the different AIDS groups (G1, G2 and G3) using their microbiome data collected at the first visit when all of them were HIV-.

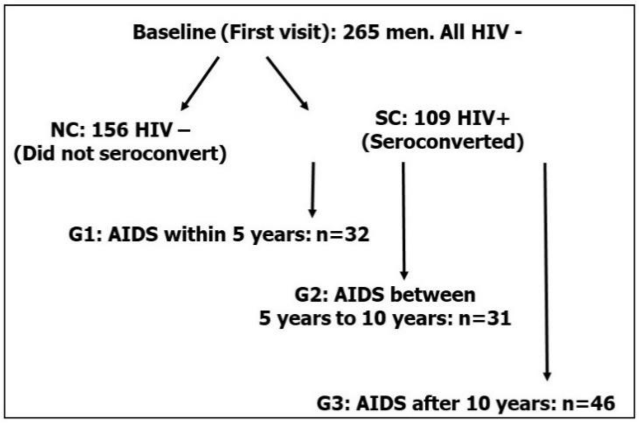

Baseline (First visit): 265 men. All HIV -
NC: 156 HIV –
(Did not seroconvert)
SC: 109 HIV+
(Seroconverted)
G1: AIDS within 5 years: n=32
G2: AIDS between
5 years to 10 years: n=31
G3: AIDS after 10 years: n=46

Figure 1: Study participants in the Multicenter AIDS Cohort Study (MACS)

Using ANCOM-BC for differential abundance analysis of microbiome and some standard regression-based methods for other data collected in the study, such as alpha diversity, cytokines data, ratio of CD4/CD8 counts, and short chain fatty acids data, Chen *et al*. (2021) made several interesting and important discoveries. At the baseline, or the first visit, when all 265 men in the sample are HIV negative, not surprisingly Chen *et al*. (2021) did not find differences in the ratio of CD4/CD8 between those who became HIV+ in the future versus those who stayed HIV-.

However, as expected, by the next visit when some became HIV+, the CD4/CD8 ratio was significantly different between the two groups of men at the second visit. Although there was no significant difference in the alpha diversity at the baseline between men who later became HIV+ and those who remained HIV-., very interestingly, Chen *et al*. (2021) did find differential abundance of various pro and anti-inflammatory taxa at the baseline between the two groups of men. Thus, they discovered intestinal dysbiosis months before men developed HIV infection, characterized by increase in pro-inflammatory taxa such as *Prevotella Stercorea* and a reduction in commensal bacteria such as *Bacteroides spp*, *Akkermansia Muciniphila, Alistipes spp, and Ruminococcus Spp*. Not only did they see such differences at the baseline, but Chen *et al*. (2021) also discovered that ratio of *Prevotellaceae to Bacteroidaceae* was highly correlated with HIV infection at a later time point. In view of these findings, it is not surprising that they also discovered elevated levels of circulating cytokines IL-6, LBP, sCD14, sCD163, before developing HIV infection. Increased levels of these cytokines suggest immune response to changes in the bacterial composition. There is growing evidence in the literature demonstrating the important roles played by short chain fatty acids produced by the gut microbiota. Propionate is one such short chain fatty acid which Chen *et al*. (2021) found to be positively correlated with the levels of CD4/CD8 counts. Taken all these findings together, it appears that differences in gut microbial composition could be in the pathway of a person developing HIV infection. Chen *et al.* (2021) also found significant increase in pro-inflammatory bacteria and a significant decrease in the anti-inflammatory commensal bacteria at baseline among those who developed AIDS soon after becoming HIV+ compared to those who were either slow to develop or never developed AIDS.

4. Conclusions

Although the focus of Chen *et al.* (2021) was on understanding the role of microbiome in HIV infection and the development of AIDS, their work together with emerging literature cited in this paper, suggests that gut microbiome may potentially play an important role in other infectious diseases, including COVID-19. There appear to be differences in the composition of gut microbiome in people who later became HIV+ than those remained HIV-, and those who rapidly developed AIDS versus those who did not. Before one can assert about other infectious diseases, more carefully planned studies similar to MACS' HIV/AIDS study are needed. The MACS study provided a unique opportunity because, at the beginning of the study none of the men were HIV+, but over time some became HIV+, and the comparisons of microbiome data were performed before anyone became HIV+. Such studies are not easy to conduct, unless it becomes a common practice to collect microbiome samples routinely, for example during the annual physical exams. Another possibility is to collect stool samples from subjects when they have an infectious disease and again obtain stool samples after they are fully recovered from their disease. Of course, such a design assumes that the subjects gut microbiome did not change permanently once a person is infected.

From a statistical perspective, these data and this line of research provide opportunities to develop statistical methods for analyzing these complex data.

Acknowledgement

The author's research was funded by the intramural research program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD, NIH).

References

Begley L., Madapoosi S., Opron K., Ndum O., Baptist A., Rysso K., Erb-Downward J. R. and Huang Y. J. (2018). Gut microbiota relationships to lung function and adult asthma phenotype: a pilot study. *BMJ Open Respiratory Research*, **5**(**1**), e000324.

Cahenzli J., Köller Y., Wyss M., Geuking M. B. and McCoy K. D. (2013). Intestinal microbial diversity during early-life colonization shapes long-term IgE levels. *Cell Host Microbe*, **14**(**5**), 559e570.

Cai Y., Chen L., Zhang S., Zeng L. and Zeng G. (2021). The role of gut microbiota in infectious diseases. *WIREs Mechanisms of Disease*. DOI: 10.1002/wsbm.1551.

Chen Y., Lin H., Cole M., Morris A., Martinson J., Mckay H., Mimiaga M., Margolick J., Fitch A., Methe B., Srinivas V. R., Peddada S. D. and Rinaldo C. R. (2021). Signature changes in gut microbiome are associated with increased susceptibility to HIV-1 infection in MSM. *Microbiome*. doi: 10.1186/s40168-021-01168-w.

Fernandes, A. D. *et al*. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.

Gevers D., Kugathasan S., Denson L. A., Vazquez-Baeza Y., Treuren W. V., Ren B., Schwager E., Knights D., Song S. J., Yassour M., *et al*. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host and Microbe*, **15**(**3**), 382-392.

Giovanni M. Y., Schnieder J. S., Calder T. and Fauci A. S. (2021). Refocusing human microbiota research in infectious and immune-mediated diseases: Advancing to the next stage. *Journal of Infectious Diseases*. DOI: 10.1093/infdis/jiaa706.

Harper A., Vijaykumar V., Ouwehand A. C., ter Haar J., Obis D., Espadaler J., Binda S., Desiraju S. and Day R. (2021). Viral infections, the microbiome, and probiotics. *Frontiers in Cellular Infections and Microbiology*. https://doi.org/10.3389/fcimb.2020.596166.

Harris V. C., Haak B. W., van Hensbroek M. B. and Wiersinga W. J. (2017). The intestinal microbiome in infectious diseases: The clinical relevance of a rapidly emerging field. *Open Forum Infectious Diseases*, doi: 10.1093/ofid/ofx144.

Hu Y., Satten G., Hu Y. -J. (2022). LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. https://doi.org/10.1073/pnas.2122788119.

Hurst G. D. D. (2017). Extended genomes: symbiosis and evolution. *Interface Focus*, **7**(**5**), 20170001.

Hussain I., Cher G. L. Y., Abid M. A. and Abid M. B. (2021). Role of gut microbiome in COVID-19: An insight into pathogenesis and therapeutic potential. *Frontiers in Immunology*. doi: 10.3389/fimmu.2021.765965.

Janda J. M. and Abbott S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology*, **45**, 2761-2764.

Kaul A., Mandal S., Davidov O. and Peddada S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology*. doi: 10.3389/fmicb.2017.02114

Langille M. G. I., Zaneveld J., Caporaso J. G., McDonald D., Knights D., Reyes J. A., Clemente J. C., Burkepile D. E., Vega Thurber R. L., Knight R., Beiko R. G. and Huttenhower C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, **31**, 814–821.

Libertucci J. and Young V. B. (2019). The role of the microbiota in infectious diseases. *Nature Microbiology*, https://doi.org/10.1038/s41564-018-0278-4

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Lozupone C. A., Li M., Campbell T. B., Flores S. C., Linderman D., Gebert M. J., Knight R., Fontenot A. P. and Palmer B. E. (2013). Alterations in the gut microbiota associated with HIV-1 infection. *Cell Host and Microbe*, **14(3)**, 29 - 339.

Mandal S., Van Treuren W., White R. A., Eggesbø M., Knight R. and Peddada S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, **26**, 1–7.

Nearing, J. T., Douglas, G. M., Hayes, M. G. *et al.* (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, **13,** 342. https://doi.org/10.1038/s41467-022-28034-z

O'Hara A. M. and Shanahan F. (2006). The gut organ as a forgotten organ. *EMBO Reports*, **7**(**7**), 688-693.

Relman D. A. and Falkow S. (2017). The meaning and impact of the human genome sequence for microbiology. *Trends in Microbiology*, **9**(**5**), 206-208.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Sampson T. R. and Mazmanian S. K. (2015) Control of brain development, function, and behavior by the microbiome. *Cell Host Microbe*, **17**(**5**), 565–576. https://doi.org/10.1016/j.chom.2015.04.011

Steenbergen L., Sellaro R., van Hemert S., Bosch J. A. and Colzato L. S. (2015). A randomized controlled trial to test the effect of multispecies probiotics on cognitive reactivity to sad mood. *Brain, Behaviour and Immunity*, **48**, 258–264.

Tierney B. T., Yang Z., Luber J. M., Beaudin M., Wibowo M. C., Baek C., Mehlenbacher E., Patel C. J. and Kostic A. D. (2019). The landscape of genetic content in the gut and oral human microbiome. *Cell Host and Microbe*, **26**(**2**), 283-295.

Turnbaugh P. J., Hamady M., Yatsunenko T., Cantarel B. L., Duncan A., Ley R. E., Sogin M. L., Jones W. J., Roe B. A., Affourtit J. P., Henrissat B., Heath A. C., Knight R. and Gordon J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, **457**(**7228**), 480.

Vila A. V., Collij V., Sanna S., Sinha T., Imhann F., Bourgonje A. R., Mujagic Z., Jonkers D. M. A. E., Masclee A. A. M., Fu J., Kurilshikov A., Wijmenga C., Zhernakova A.and Weersma R. K. (2020). Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. *Nature Communications*, **11**, 362. https://doi.org/10.1038/s41467-019-14177-z.

Weersma R. K., Zhernakova A. and Fu J. (2020). Interaction between drugs and the gut microbiome. *Gut*, **0**, 1–10.

Wang Q. I., Li F., Liang B., Liang Y., Chen S., Mo X., Ju Y., Zhao H., Jia H., Spector T. D., Xie H. and Guo R. (2018). A metagenome-wide association study of gut microbiota in asthma in UK adults. *BMC Microbiology*, **18**(**1**), 114.

Weiss S., Xu Z. Z., Peddada S., Amir A., Bittinger K., Gonzalez A. *et al*. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27. doi: 10.1186/s40168-017-0237-y.

Wu H. J. and Wu E. (2012) The role of gut microbiota in immune homeostasis and autoimmunity. *Gut Microbes*, **3**(**1**), 4–14.

Yeoh Y. K., Zuo T., Lui G. C.-Y. *et al*. (2021). Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19. *Gut*, doi:10.1136/gutjnl-2020-323020.

Zhou, H., He, K., Chen, J. *et al.* (2022). LinDA: Linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, **23,** 95. https://doi.org/10.1186/s13059-022-02655-5.

Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 215-230

Spatial Hierarchical Bayes Small Area Inference

Priyanka Anjoy
Assistant Director, National Accounts Division (NAD), National Statistical Office (NSO)
Ministry of Statistics and Program Implementation (MOSPI), Government of India

Received: 16 August 2022; Revised: 25 August 2022; Accepted: 28 August 2022

Abstract

The Hierarchical Bayes predictor of small area proportions (HBP) under an area level version of generalized linear mixed model with logit link function is widely used in small area estimation for binary variable. However, this predictor does not account for the presence of spatial effect between contiguous or neighbouring regions. Conditional Autoregressive and Simultaneous Autoregressive specifications do incorporate spatial associationship while considering the spatial correlation effects among areas. But none of these approaches implement the idea of spatially varying covariates through spatially dependent fixed effect parameters. Such approach in statistics is known as spatial nonstationarity. This article introduces a spatially nonstationary extension to the Hierarchical Bayes predictor of small area proportions that accounts for the presence of spatial nonstationarity. The proposed predictor is referred as the spatial nonstationary Hierarchical Bayes predictor (HBNSP). The impact of survey design information is also explored in the proposed predictor. The empirical results from simulation studies using spatially nonstationary data indicate that the HBNSP method performs better, in terms of relative bias and relative mean squared error, than the alternative HBP method that ignore this spatial nonstationarity. The results further show that use of survey-weight to incorporate the sampling design appears to be imperative when sample data is informative.

Key Words: Hierarchical Bayes, Small area estimation, Spatial nonstationarity, Survey-weight.

1. Introduction

In recent years, small area estimation (SAE) technique has emerged as one of the most important topics in survey estimation because of an increasing demand for reliable small area statistics by various government and international agencies, see for example, Rao and Molina (2015). United Nations Sustainable Development Agenda has also marked the developmental strategy through availing and utilizing disaggregate level statistics in the programmes and planning aimed at uprooting social and regional inequalities. Sample surveys are generally designed so that direct estimators (*i.e.* estimators that use only the sample data from the domain of interest) for larger domains provide reliable estimates for parameters of interest. On many occasions, however, the interest is in estimating parameters for domains that contain only a small number of sample observations or sometimes no sample observations. The term ‘small areas’ is used to describe domains whose sample sizes are not large enough to allow sufficiently precise direct estimation. Hereafter, refer to these smaller domains as 'small areas' or simply 'areas'. When direct estimation is not possible, one must rely on alternative, model-based

Corresponding Author: Priyanka Anjoy
EMail: anjoypriyanka90@gmail.com

methods for producing small area estimates. Further, large scale surveys produce reliable estimates at higher geographical level and such estimates often mask variations which is available at local levels. This restricts targeting of heterogeneity at higher levels of spatial disaggregation and limits the scope for monitoring and evaluation of parameters locally within and across administrative units. Model-based SAE techniques are now widely used in practice to meet the indispensable need of reliable disaggregate level statistics from the existing survey data. Such SAE methods depend on the availability of population level auxiliary information related to the variable of interest and are commonly referred to as indirect methods. The industry standard for SAE is to use unit or area level models (Fay and Herriot, 1979; Battese, Harter and Fuller, 1988). In the former case these models are for the individual survey measurements and include area effects, while in the latter case these models are used to smooth out the variability in the unstable area-level direct estimates. Area-level small area modeling is usually employed when unit-level data are unavailable, or, as is often the case, where model covariates (*e.g.* census variables) are only available in aggregate form. In this article solely focus is on area (or aggregated) level small area modeling.

Fay-Herriot (FH) model is one of the popular examples of aggregated level small area model. For continuous survey variable, this model is widely used in practice and has led the phenomenal development of small area literatures based on this model. However, binary or count data is often of interest in many practical applications. In epidemiological, environmental, poverty related studies such data is much common, where interest generally lies in estimation of proportions. A generalized linear mixed model (GLMM) with logit link function (also referred to as logistic linear mixed model) is commonly used for estimation of small area proportions. The basic structure of area level small area models includes sampling model for direct survey estimates and associated sampling error; linking model to link the parameter of interest with area-specific auxiliary variables and random effects. The area random effect in small area models explains unstructured heterogeneity between areas. Two basic approaches for drawing inferences about the small area parameters of interest are known to be popular: The empirical best prediction method is based on frequentist idea to estimate unknown model parameters and the hierarchical Bayes (HB) approach assumes particular prior distributions for the hyperparameters to obtain posterior quantities of the parameter of interest. The HB approach has the flexibility to deal with complex SAE model as it overcomes the difficulties of analytical mean squared error (MSE) estimation in frequentist set up and provides quick and easier posterior variance computation based on Markov Chain Monte Carlo (MCMC) simulation. Refer Jiang and Lahiri (2001), You and Zhou (2011), Liu *et al.* (2014), Rao and Molina (2015) and Chandra *et al.* (2018) for frequentist and Bayesian related studies and various real life applications. This article in particular focuses on estimation of small area proportions in hierarchical Bayes framework. Among the previous literatures, Liu *et al.* (2014) and Anjoy *et al.* (2019) have applied hierarchical Bayes version of GLMM (HBGLMM) to estimate survey-weighted small area proportions considering different cases of known and unknown sampling variance structure (denoted by HBP). The linking model of HBP incorporates random effect which is assumed to be independent and identically distributed. As a result, spatial associationship between geographical areas cannot be described through this structure of the model. However, in many small area problems like disease prevalence and poverty estimation, spatial contiguity between neighbouring areas is very common. Therefore, induction of spatial variability in GLMM can be a way of reducing the variances or Coefficient of Variation (CV) in final estimates. One approach to incorporating such spatial dependency among the areas is to extend the GLMM to allow for spatially correlated area effects using, for example, a Simultaneous Autoregressive (SAR) model (Cressie, 1993). This model allows for spatial correlation in the area effects, while keeping the fixed effects parameters spatially invariant (Chandra and Salvati, 2018). There are data situations, where this assumption is inappropriate and parameters associated with the model covariates (*i.e.* the fixed effects parameters) vary spatially. This phenomenon is often referred to as spatial nonstationarity (Brunsdon *et al.*, 1996). An alternative approach to incorporating spatial information in SAE is therefore to assume that the parameters associated with the model covariates vary spatially. In frequentist framework, Chandra *et al.* (2017) has devised the concept of spatial nonstationarity in area level version of GLMM (NSGLMM) for estimating small area proportions. A key feature of this approach is that it tries to capture spatial variability through incorporating spatially varying covariates in the linking model. It is worth noting that Chandra *et al.* (2017) approach does not use the sampling weights or clustering information in estimation of small area proportions under NSGLMM. However, use of this sampling information is essential for valid inference from survey data collected by complex survey designs. Baldermann *et al.* (2018) has also forwarded spatial nonstationarity concept for explaining spatial variability between areas, but their model is for unit-level data. Contrary to the previous studies, this article describes a spatial nonstationary version of hierarchical Bayes approach for SAE that incorporates the sampling information when estimating small area proportions under an area level small area models (denoted by HBNSP). Unlike frequentist approach, the HBNSP method offers the flexibility of MSE estimation through posterior variance computation based on MCMC simulation.

Standard model-based approaches to the analysis often ignore the sampling mechanism. The GLMM technique implicitly considers equal probability sampling (simple random sampling with replacement) within each small area and thus ignores the survey-weight (Chandra *et al.*, 2019). But, this may result in potentially large biases in the final estimates. In FH model for estimation of small area population mean, direct design-based estimators are modeled directly and the survey variance of the associated direct estimator is introduced into the model via the design-based errors. The Horvitz-Thompson estimator, weighted Hájek estimator are the structures here to incorporate survey design information (Hidiroglou and You, 2016). However, this method for continuous data requires extension for binary or count data for estimating more representative small area proportions. Consequently, the strategic idea is to modeling survey-weighted proportions (Liu *et al*., 2014). Hence, the next attempt in this article is to check the impact of complex survey design information in HBNSP. In next section, two versions of HBNSP predictor denoted as HBNSP1 and HBNSP2 are described, which respectively account for HB modeling of unweighted and survey-weighted small area proportions. In section 3, empirical evaluation studies first include a model-based simulation set up to evaluate the performance of proposed HBNSP as compared to HBP. Secondly, a design-based simulation study is carried out for comparing the performance of HBNSP1 and HBNSP2 which respectively, ignores and considers the modeling of survey-weighted proportions. The article ends with relevant concluding remarks.

2. Methodology

Let us consider a finite population U of size N which is partitioned into D distinct small areas or simply areas. The set of population units in area i is denoted as U_i with known size N_i, such that $U = \cup_{i=1}^{D} U_i$ and $N = \sum_{i=1}^{D} N_i$. A sample s of size n is drawn from population U using a probabilistic mechanism. This resulted in sample s_i in area i with size n_i, so that $s = \cup_{i=1}^{D} s_i$ and $n = \sum_{i=1}^{D} n_i$. Assume that y_{ij} be the value of target variable y for unit j (j=1,…,n_i) in small area *i*. The target variable with values y_{ij} has binary response, taking value either 1 or 0. Our aim is to estimate the small domain proportions $P_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$. When the sample *s* is drawn following a complex survey design, with each unit y_{ij} in small area *i* design weight w_{ij} is attached, which is alternatively known as survey-weights or sampling weights.

2.1. Estimation of small area proportions

The area level version GLMM is widely used for estimation of small area proportions to improve the precision of direct survey estimates. Consider, p_i be the direct survey estimator for the parameter of interest P_i . In aggregated level model, it is customary to assume that,

$$p_i = P_i + e_i;\ i = 1,...,D\ ,$$

where e_i 's are independent sampling error associated with direct estimator p_i . Sampling error e_i is assumed to have zero mean and known sampling variance σ_{ei}^2. The linking model of P_i attempt to relate area-specific auxiliary variables and random effect component,

$$g(P_i) = \mathbf{x}_i'\boldsymbol{\beta} + v_i; i = 1,...,D,$$

where the linking function $g(.)$ is logit for binary data and log for count data, $\mathbf{x}_i$ represent matrix of area-specific auxiliary variables, $\boldsymbol{\beta}$ is the regression coefficient or fixed effect parameter vector and v_i being the area-specific random effect, independent and identically distributed as $\mathrm{E}\left(v_i\right) = 0$ and $\mathrm{var}\left(v_i\right) = \sigma_v^2$. Random area-specific effects are included in the linking model to account for between areas dissimilarities. Working under HB set up, certain prior distributions are assumed for the hyperparameters. For estimating small area proportions P_i , the sampling and linking models of HBP are represented as,

$$p_i \mid P_i \sim N(P_i, \sigma_{ei}^2),\ i = 1,...,D \text{ and } \mathrm{logit}(P_i) \mid \boldsymbol{\beta}, \sigma_v^2 \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_v^2),\ i = 1,...,D\ .$$

Following standard literature, prior choice for $\boldsymbol{\beta}$ is usually taken to be $N(0,\sigma_0^2)$ and for σ_v^2 prior choice is $IG(a_0, b_0)$, (*IG* stands for Inverse Gamma) where σ_0^2 is set to be very large (say, 10^6) and very small value for a_0 and b_0 (usually $a_0 = b_0 \to 0$) to reflect lack of prior knowledge about variance parameters (Rao, 2015; You and Zhou, 2011). Then, inferences about the small area parameter of interest are drawn from posterior distribution. Posterior mean is taken as the point estimate of the parameter and posterior variance as a measure of the uncertainty associated with the estimate. However, an inbuilt postulation in HBP is that fixed effect parameter or regression coefficient vector $\boldsymbol{\beta}$ is spatially invariant, this is what customarily known as spatial stationarity. In contrary, spatial nonstationarity approach tends to describe/define spatially varying regression parameters, *i.e.*, values of the regression coefficients are necessarily different at different spatial locations. Small area estimation of proportions in presence of such spatial nonstationarity is described in next subsection.

2.2. Hierarchical Bayes version of spatial nonstationary GLMM

For spatial nonstationary version of HBGLMM or HBNSP, regression coefficients in the small area model may be expressed as explicit functions of the spatial points of the sample observations. Unlike HBP, where one restrict to a single global model with fixed parameter, HBNSP technique defines local relationships to exist between study and auxiliary variables. This approach is quite like the geographically weighted regression (GWR) in a multiple regression framework which takes nonstationary auxiliary variables into consideration (Brunsdon *et al.*, 1996). Let, l_i be the coordinates of an arbitrarily defined spatial location (longitude and latitude) for i^{th} small area; generally, this will be its centroid. Consider, $\mathbf{l}=(l_1,...,l_D)'$ denoting the D component vector of such spatial locations *i.e.*, having available longitude and latitude for all the D spatial locations or areas of interest. Assume that nonstationarity is characterized by an area specific vector of fixed effects,

$$\mathbf{x}_i'\boldsymbol{\beta}(l_i)=\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{x}_i'\boldsymbol{\gamma}(l_i),$$

where $\boldsymbol{\beta}(l_i)=\boldsymbol{\beta}+\boldsymbol{\gamma}(l_i)$ and $\boldsymbol{\gamma}(l_i)=(\gamma_1(l_i),...,\gamma_p(l_i))'$. The linking model of P_i in HBNSP attempt to relate nonstationary auxiliary variables and random effect component,

$$\text{logit}(P_i)=\mathbf{x}_i'\boldsymbol{\beta}(l_i)+v_i; i=1,...,D, \text{ with } v_i \sim N(0,\sigma_v^2).$$

Aggregating D area level models lead to the population level version of the HBNSP as

$$\mathbf{p}=\mathbf{X}\boldsymbol{\beta}(\mathbf{l})+\mathbf{v}+\mathbf{e}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\Psi}\boldsymbol{\Theta}+\mathbf{v}+\mathbf{e},$$

where $\mathbf{p}=(p_1,...,p_D)'$ is the vector of direct survey estimates, $\mathbf{X}=(\mathbf{x}_1',...,\mathbf{x}_D')'$ be $D\times p$ matrix of auxiliary variates, $\boldsymbol{\beta}$ is the fixed effect parameter vector, $\mathbf{v}=(v_1,...,v_m)'$ is a vector of domain random effects such that $\mathbf{v}\sim N(\mathbf{0},\sigma_v^2\mathbf{I}_D)$, $\mathbf{I}_D$ is the unit matrix of dimension D, $\mathbf{e}=(e_1,...,e_D)'$ is the vector of sampling errors with $\mathbf{e}\sim N(\mathbf{0},\boldsymbol{\Omega})$, where $\boldsymbol{\Omega}=\text{diag}\{\sigma_{ei}^2;1\le i\le D\}$ is the matrix of design variances. $\boldsymbol{\Psi}=\{diag(\mathbf{x}_1'),...,diag(\mathbf{x}_D')\}'$ is a $D\times pD$ matrix of known auxiliary data; $\boldsymbol{\Theta}=(\boldsymbol{\gamma}'(l_1),...,\boldsymbol{\gamma}'(l_D))'$ is a spatial Gaussian random vector of dimension $pD\text{x}1$ such that $\text{E}(\boldsymbol{\Theta}\,|\,\boldsymbol{\Psi},\mathbf{l})=\mathbf{0}$ and covariance matrix $\text{var}(\boldsymbol{\Theta}\,|\,\boldsymbol{\Psi},\mathbf{l})=\boldsymbol{\Sigma}_\eta=\mathbf{W}\otimes(\mathbf{cc}')$, where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{W}=1/(1+L(l_i,l_j))$ defines the spatial distances between sample spatial locations (l_i,l_j), specifically distances between centroids of two locations (i,j). In general, the only constraint on the vector $\mathbf{c}$ is that $\boldsymbol{\Sigma}_\eta=\mathbf{W}\otimes(\mathbf{cc}')$ is symmetric and non-negative definite. Following Chandra et al. (2017), consider $\mathbf{c}=\sqrt{\eta}\mathbf{1}_p$, where $\eta\ge 0$ and $\mathbf{1}_p$ denotes the unit vector of order p. So, $\boldsymbol{\Sigma}_\eta=\eta\mathbf{W}\otimes(\mathbf{1}_p\mathbf{1}_p{}')$ involves non zero covariance $\text{cov}(\gamma_k(l_i),\gamma_h(l_j))=\eta/(1+L(l_i,l_j))$ between $\gamma_k(l_i)$ and $\gamma_h(l_j)$ for sample spatial locations (i,j), with $k\ne h=1,...,p$ and diagonal elements as η. The parameter η denotes the strength of spatial heterogeneity being explained by nonstationary auxiliary variables. In particular $\eta = 0$ indicates the situation where the model is spatially homogeneous. In HB framework, the sampling and linking models for HBNSP are then expressed as

$$\mathbf{p}|\mathbf{P} \sim N(\mathbf{P},\boldsymbol{\Omega}) \text{ and } \text{logit}(\mathbf{P})|\boldsymbol{\beta},\eta,\sigma_v^2 \sim N(\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}' + \sigma_v^2\mathbf{I}_\text{D}).$$

The prior for hyper-parameter $\boldsymbol{\beta}$ is $N(0,\sigma_0^2)$ and for variance parameters η and σ_v^2 prior is $IG(a_0,b_0)$, where σ_0^2 is set to be very large (say, 10^6) and very small value for a_0 and b_0. Note that HBNSP reduces to HBP when $\eta = 0$. Gibbs sampling method is implemented to estimate posterior mean $\text{E}\left(P_i\,|\,\mathbf{p}\right)$ and posterior variance $\text{var}\left(P_i\,|\,\mathbf{p}\right)$. The required full conditional distribution of parameters under HBP and HBNSP models are given in Section 2.3.

2.3. Survey-weighted estimation

The HB modeling of respectively unweighted and survey-weighted small area proportions is a way to check the impact of complex survey design information in the resultant estimates. Survey-weighted direct estimates used for HB modeling purpose have the potentiality to reduce the bias or design error of the final estimates. Consider sample *s* of size *n* is drawn from population *U* using a complex design or at least unequal probability scheme. Let p_{ij} be the selection probability attached to j^{th} sampling unit y_{ij} in the area *i*. The basic design weight can be given by $w_{ij} = (n_i p_{ij})^{-1}$. These weights can be adjusted to account for non-response and/or auxiliary information (Hidiroglou and You, 2016). Normalized survey-weights d_{ij} may also be constructed, $d_{ij} = w_{ij}\left(\sum_j w_{ij}\right)^{-1}$. Liu *et al.* (2014) and Anjoy *et al.* (2019) have considered HB modeling of survey-weighted small area proportions, where GLMM structure was used for estimation of area proportions. But the effects of taking informative samples were not discussed. Here, two alternative models of HBNSP are definedto study the impact of design informativeness while aim is to estimate small area proportions in presence of spatial nonstationary auxiliary variables using the above furnished HBNSP technique. Let, $p_{i.uw}$ be the direct survey unweighted estimator for small area proportion P_i ,

$$p_{i.uw} = \left(n_i\right)^{-1}\sum_{j=1}^{n_i} y_{ij} \text{ and the variance of } p_{i.uw} \text{ is given as } \sigma_{ei.uw}^2 = n_i^{-1}P_i\,(1\text{-}P_i).$$

The survey-weighted estimator denoted as, $p_{i.sw}$ and its variance is expressed as,

$$p_{i.sw} = \left(\sum_{j=1}^{n_i} w_{ij}\right)^{-1}\sum_{j=1}^{n_i} w_{ij}y_{ij} \text{ and } \sigma_{ei.sw}^2 = \left(\sum_{j=1}^{N_i} w_{ij}\right)^{-2}\left\{\sum_{j=1}^{N_i} w_{ij}(w_{ij}-1)(y_{ij}-P_i)^2\right\}.$$

Two HBNSP methods are explored for the impact of complex survey deign, denoted as HBNSP1 and HBNSP2. These models are furnished below:

HBNSP1: Does not incorporate survey-weight

Sampling model: $\mathbf{p}_{uw}|\mathbf{P} \sim N(\mathbf{P},\boldsymbol{\Omega}_{uw})$

Linking model: $\mathbf{g(P)}|\boldsymbol{\beta},\eta,\sigma_v^2 \sim N(\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}})$

HBNSP2: Incorporate survey-weight

Sampling model: $\mathbf{p}_{sw}|\mathbf{P}\sim N(\mathbf{P},\boldsymbol{\Omega}_{sw})$

Linking model: $\mathbf{g(P)}|\boldsymbol{\beta},\eta,\sigma_v^2 \sim N(\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}})$

The required full conditional distributions of HBNSP1 and HBNSP2 under Gibbs sampler are given as below. Within the Gibbs sampler, particularly Metropolis-Hastings (M-H) algorithm is used for drawing random samples from full conditional distributions of posterior quantities. For HBP model, the full conditional distributions for the Gibbs sampler are given as,

$$P_i\,|\boldsymbol{\beta},\sigma_v^2,p_i \propto \frac{1}{P_i(1-P_i)\sqrt{\sigma_{ei}^2\sigma_v^2}}\,exp\left(-\frac{\left(p_i-P_i\;\right)^2}{2\sigma_{ei}^2}-\frac{\left(\log\mathrm{it}\left(P_i\right)-\mathbf{x}_i'\boldsymbol{\beta}\;\right)^2}{2\sigma_v^2}\right),$$

$$\boldsymbol{\beta}\,|P_i,\sigma_v^2\sim N\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log\mathrm{it}(\mathbf{P}),\sigma_v^2(\mathbf{X}'\mathbf{X})^{-1}\right),\text{ and}$$

$$\sigma_v^2\,|\boldsymbol{\beta},P_i,\sim IG\left(a+\frac{D}{2},\;b+\frac{\sum_{i=1}^{D}\left(\log\mathrm{it}(P_i)\text{-}\mathbf{x}_i'\boldsymbol{\beta}\right)^2}{2}\right).$$

For HBNSP model, the full conditional distributions for the Gibbs sampler are given as,

$$\mathbf{P}|\boldsymbol{\beta},\eta,\sigma_v^2,\mathbf{p}\propto|\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)|^{-\frac{1}{2}}|\boldsymbol{\Omega}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\{(\mathbf{p}-\mathbf{P})'\,\boldsymbol{\Omega}^{-1}(\mathbf{p}-\mathbf{P})$$

$$+\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)^{-1}\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta})\}\right]\left|\frac{\partial\log\mathrm{it}(\mathbf{P})}{\partial\mathbf{P}}\right|,$$

$$\boldsymbol{\beta}|\mathbf{P},\eta,\sigma_v^2\sim\mathrm{MVN}\left[\left(\mathbf{X}'\boldsymbol{\Pi}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\boldsymbol{\Pi}^{-1}\log\mathrm{it}(\mathbf{P})\right),\left(\sigma_v^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'\right)\left(\mathbf{X}'\boldsymbol{\Pi}^{-1}\mathbf{X}\right)^{-1}\right],$$

$$\eta|\boldsymbol{\beta},\sigma_v^2,\mathbf{P}\sim\mathrm{IG}\left[a_0+\frac{D}{2},b_0+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)}{2}\right],\text{ and}$$

$$\sigma_v^2|\boldsymbol{\beta},\eta,\mathbf{P}\sim\mathrm{IG}\left[a_1+\frac{D}{2},b_1+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)}{2}\right].$$

where, $\boldsymbol{\Pi}=\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_{\mathrm{v}}^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Omega}$. Recall that $\boldsymbol{\Sigma}_\eta=\eta\mathbf{W}\otimes\left(\mathbf{1}_p\mathbf{1}_p{}'\right)$ with distance matrix $\mathbf{W}=1\big/\left(1+L\left(l_i,l_j\right)\right)$.

For HBNSP1 model, the full conditional distributions for the Gibbs sampler are given as,

$$\mathbf{P}|\boldsymbol{\beta},\eta,\sigma_v^2,\mathbf{p}_{uw}\sim|\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)|^{-\frac{1}{2}}|\boldsymbol{\Omega}_{uw}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\{\left(\mathbf{p}_{uw}-\mathbf{P}\right)'\boldsymbol{\Omega}_{uw}{}^{-1}\left(\mathbf{p}_{uw}-\mathbf{P}\right)$$

$$+\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)^{-1}\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta})\}\right]\left|\frac{\partial\log\mathrm{it}(\mathbf{P})}{\partial\mathbf{P}}\right|,$$

$$\boldsymbol{\beta}|\mathbf{P},\eta,\sigma_v^2\sim\mathrm{MVN}\left[\left(\mathbf{X}'\boldsymbol{\Pi}_{uw}{}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\boldsymbol{\Pi}_{uw}{}^{-1}\log\mathrm{it}(\mathbf{P})\right),\left(\sigma_v^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'\right)\left(\mathbf{X}'\boldsymbol{\Pi}_{uw}{}^{-1}\mathbf{X}\right)^{-1}\right],$$

$$\eta|\boldsymbol{\beta},\sigma_v^2,\mathbf{P}\sim\mathrm{IG}\left[a_0+\frac{D}{2},b_0+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)}{2}\right],\text{ and}$$

$$\sigma_v^2|\boldsymbol{\beta},\eta,\mathbf{P}\sim\mathrm{IG}\left[a_1+\frac{D}{2},b_1+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)}{2}\right].$$

where, $\mathbf{p}_{uw}=\left(p_{1.uw},\ldots,p_{D.uw}\right)'$; $\boldsymbol{\Omega}_{uw}=\mathrm{diag}\left\{\sigma_{ei.uw}^2;1\le i\le D\right\}$ and $\boldsymbol{\Pi}_{uw}=\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_{\mathrm{v}}^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Omega}_{uw}$.

For HBNSP2 model, the full conditional distributions for the Gibbs sampler are given as,

$$\mathbf{P}|\boldsymbol{\beta},\eta,\sigma_v^2,\mathbf{p}_{sw}\sim|\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)|^{-\frac{1}{2}}|\boldsymbol{\Omega}_{sw}|^{-\frac{1}{2}}\exp[-\frac{1}{2}\{\left(\mathbf{p}_{sw}-\mathbf{P}\right)'\boldsymbol{\Omega}_{sw}^{-1}\left(\mathbf{p}_{sw}-\mathbf{P}\right)$$

$$+\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_v^2\mathbf{I}_{\mathrm{D}}\right)^{-1}\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta})\}\right]\left|\frac{\partial\log\mathrm{it}(\mathbf{P})}{\partial\mathbf{P}}\right|,$$

$$\boldsymbol{\beta}|\mathbf{P},\eta,\sigma_v^2\sim\mathrm{MVN}\left[\left(\mathbf{X}'\boldsymbol{\Pi}_{sw}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\boldsymbol{\Pi}_{sw}^{-1}\log\mathrm{it}(\mathbf{P})\right),\left(\sigma_v^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'\right)\left(\mathbf{X}'\boldsymbol{\Pi}_{sw}^{-1}\mathbf{X}\right)^{-1}\right],$$

$$\eta|\boldsymbol{\beta},\sigma_v^2,\mathbf{P}\sim\mathrm{IG}\left[a_0+\frac{D}{2},b_0+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\mathbf{v}\right)}{2}\right],\text{ and}$$

$$\sigma_v^2|\boldsymbol{\beta},\eta,\mathbf{P}\sim\mathrm{IG}\left[a_1+\frac{D}{2},b_1+\frac{\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)'\left(\log\mathrm{it}(\mathbf{P})-\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\Psi}\boldsymbol{\Theta}\right)}{2}\right].$$

where, $\mathbf{p}_{sw}=\left(p_{1.sw},\ldots,p_{D.sw}\right)'$; $\boldsymbol{\Omega}_{sw}=\mathrm{diag}\left\{\sigma_{ei.sw}^2;1\le i\le D\right\}$ and $\boldsymbol{\Pi}_{sw}=\boldsymbol{\Psi}\boldsymbol{\Sigma}_\eta\boldsymbol{\Psi}'+\sigma_{\mathrm{v}}^2\mathbf{I}_{\mathrm{D}}+\boldsymbol{\Omega}_{sw}$.

3. Empirical Evaluations

This section reports the empirical results on the comparative performances of different estimators of the small area proportions which have been described previously. In particular, empirical performance of the proposed small area estimator HBNSP as compared to HBP is evaluated. Further, empirical performance of HBNSP1 and HBNSP2 also has been evaluated. Two types of simulation studies are used here. Section 3.1 describes the model-based simulation set up to evaluate the performance of HBNSP and HBP. In model-based simulation, population data is generated using a specified model. In section 3.2, a design-based simulation study is presented for comparing the performance of nonstationary process HBNSP1 and HBNSP2 which respectively, ignores and considers the modeling of survey-weighted proportions. Here, the aim is to explore impact of the incorporation of complex survey information. Simulation studies have been implemented in R. Different performance indicators considered for comparison of small area estimators are as below. Let *t* is the subscript for *T* simulations.

- $\text{RB}_i = 100 \times \left(T^{-1} \sum_{t=1}^{T} P_i^{(t)} \right)^{-1} \left\{ T^{-1} \sum_{t=1}^{T} \left(\hat{P}_i^{(t)} - P_i^{(t)} \right) \right\}$ is the percentage relative bias (RB) for *i*th small area, where $\hat{P}_i^{(t)}$ is the estimate of true population mean $P_i^{(t)}$ for *i*th for small area at *t*th simulation.
- $\text{RRMSE}_i = 100 \times \left(T^{-1} \sum_{t=1}^{T} P_i^{(t)} \right)^{-1} \left\{ \sqrt{T^{-1} \sum_{t=1}^{T} \left(\hat{P}_i^{(t)} - P_i^{(t)} \right)^2} \right\}$ is the percentage relative root mean squared error (RRMSE) for *i*th for small area.
- $\text{CR}_i = 100 \times T^{-1} \sum_{t=1}^{T} I \left\{ LB\left(\hat{P}_i^{(t)} \right) \le P_i^{(t)} \le UB\left(\hat{P}_i^{(t)} \right) \right\}$ is the percentage coverage rate (CR) for *i*th small area, where $LB\left(\hat{P}_i^{(t)} \right)$and $UB\left(\hat{P}_i^{(t)} \right)$ are respectively Lower Bound (*LB*) and Upper Bound(*UB*) of the estimated population mean $\hat{P}_i^{(t)}$. *I*(.) indicates an indicator function which takes values 1 if true parameter value $P_i^{(t)}$ is within the computed interval, otherwise it takes value 0. This CR% particularly demonstrate the credible interval property of HB models.

For design-based simulation, $P_i^{(t)}$ is equal to P_i or true population mean. A better model should show smaller values for all the performance indicators expect CR. Higher the CR better is the model.

3.1. Model-based simulations

In model-based simulations the data were generated using both stationary and nonstationary processes. In stationary data generation process (SDGP), the regression coefficients are spatially invariant. The aim of this simulation set is to examine how HBNSP performs when the data follows spatial stationary process. Here, data is generated via the linking model:

$$\text{logit}(P_i) = 1 + x_i + v_i,\ i = 1, \ldots, D = 100.$$

In case of nonstationary data generation process (NSDGP), data is generated from the following model:

$$\text{logit}(P_i) = 1 + x_i + \sqrt{\eta} \left(\gamma_1 \left(l_i \right) + \gamma_2 \left(l_i \right) x_i \right) + v_i,\ i = 1, \ldots, D = 100$$

Here the values of x_i were independently drawn from the uniform distribution $x_i \sim Uniform[0,1]$ and area random effects independently drawn as $v_i \sim N(0, \sigma_v^2 = 0.0625)$. Again, the sampling model part is considered as $p_i = P_i + e_i;\ i = 1, \ldots, D$. The independent sampling errors e_i are generated from $N(0, \sigma_{ei}^2)$ with σ_{ei}^2 taking values 0.01,0.02,0.03 and 0.04 respectively for equal number of areas. To define $longitude_i$ and $latitude_i$ of spatial locations, it is assumed that observations have been drawn from a two-dimensional grid consist of a $\left(\sqrt{D}\,\text{x}\sqrt{D}\right)$ points uniformly spaced between -1 to 1 with a distance of $2\big/\left(\sqrt{D}-1\right)$ between any two neighbouring points along the vertical and horizontal axes. The *D* points or spatial locations are arranged in such a way that k_1 varies from –1 to 1 for each given k_2, which also then varies from −1 to 1. For example, when *D*=100, the set $\left(k_1, k_2\right)$ is, $\left\{k_1, k_2 = -1, -0.77, -0.55, -0.33, -0.11, 0.11, 0.33, 0.55, 0.77, 1\right\}$. Further, $(\gamma_1\left(l_i\right), \gamma_2\left(l_i\right))'$ has been defined as a random draw from $N(0, \mathbf{W}\otimes\mathbf{I}_2)$ with $\mathbf{W} = 1\big/\left(1+L\left(l_i, l_j\right)\right)$ being the distance matrix between spatial locations $\left(l_i, l_j\right)$. The values of η have been used as 0.5,1, 2, 4 in this study. This simulation set up is followed from Chandra *et al.* (2017).

The process of generating data and estimation of small area proportions by implementing HBP and HBNSP methods was independently replicated *T* = 500 times from both stationary and nonstationary data generation process. The empirical performance and relative efficiency of the proposed HBNSP is compared with the HBP which excludes spatial nonstationarity structure. Performance of the small area HB estimators under each model is compared with respect to different prior cases for variance parameter σ_v^2. Specifically, *IG*(0.01,0.01) and *IG*(0.1,0.1) prior cases were taken up for such sensitivity analysis with respect to prior for variance parameter σ_v^2 . However, the result from prior *IG*(0.1, 0.1) are only reported. As inferences based on different non-informative priors were found to be similar. The prior for hyperparameter $\boldsymbol{\beta}$ was taken as *N*(*0*,10^6). The prior for η was taken to be same as σ_v^2. To implement the Gibbs sampler, three independent chains are used each of length 10000. The first 5000 iterations are deleted as “burn-in” periods. Further, following Gelman and Rubin (1992), potential scale reduction factor ($\hat{R}$) is used to monitor the convergence of the M–H within Gibbs sampler. The $\hat{R}$ value close to 1 is expected and equal to 1 implies stationarity.

Table 1 shows the average values of relative biases (RB), relative root mean squared errors (RRMSE) and coverage rates (CR) for HBP and HBNSP methods investigated in model-based simulations. In Table 1 these values are presented as percentage and averages over the small areas of interest (*D* = 100). Summary statistics of RB, RRMSE and CR for HBP and HBNSP methods for different values of η under NSDGP for *D* = 64 and 100 areas are reported in Appendix (Table A1-A2). The differences between two small area predictors HBP and HBNSP in Table 1 are essentially as one would expect. When the underlying data is stationary (*i.e.*, data generated through SDGP), with identical value of RB, value of RRMSE of HBP is marginally smaller than the HBNSP. In contrast, in presence of nonstationarity in data (*i.e.*, data generated through NSDGP), the HBNSP method performs consistently better than the HBP method both in terms of RB and RRMSE for all values of nonstationarity parameter η. Additionally, HBNSP has shown better coverage properties. Noncoverage rate is marginally higher for HBP method.

Table 1: The average values of percentage relative biases (%RB), percentage relative root mean squared errors (%RRMSE) and percentage coverage rates (%CR) for HBP and HBNSP methods in model-based simulation. Averaged *D*=100 areas

Criterion	SDGP		NSDGP							
			η =0.5		η =1		η =2		η =4	
	HBP	HBNSP	HBP	HBNSP	HBP	HBNSP	HBP	HBNSP	HBP	HBNSP
RB	0.065	0.065	-0.974	-0.478	-0.891	-0.402	-0.747	-0.359	-0.602	-0.403
RRMSE	4.289	4.447	5.636	4.635	5.130	4.242	4.593	4.047	4.622	4.281
CR	71	86	81	92	89	95	93	95	93	94

3.2. Design-based simulations

As in real life small area applications, one cannot be confident that our data ideally follow an assumed model, rather a working model is fitted. The endeavor of design-based simulation is to evaluate the performance of different SAE methods in the context of a realistic population, where a model assumption is essentially an approximation. For this simulation study, debt-investment survey (AIDIS-2013) data of National Statistical Office for rural areas of the state of Karnataka in India is used. The sample size of AIDIS-2013 is 2340 units (rural households including both indebted and non-indebted) spread over 30 districts of Karnataka. The AIDIS sample data is considered as fixed population of size 2340 units (or households) and 30 districts as small areas. Population size of small areas ranges between a minimum of 55 to a maximum of 112 with an average of 78 households. The variable of interest y_{ij} is binary which takes value 1 if a household is indebted and 0 otherwise. The aim is to estimate the proportions of indebted farm households (*i.e.* or incidence of indebtedness in farm households) at the district level. Here, Probability Proportional to Size with Replacement (PPSWR) samples were drawn independently within each small area instead of Simple Random Sampling to take into account the effect of varying sampling weights. Motivated from the simulation set up in Hidiroglou and You (2016), PPSWR sampling was employed as follows: a size measure z_{ij} is defined for a given unit y_{ij}. Using these z_{ij} values, selection probabilities $p_{ij} = z_{ij}\left(\sum_j z_{ij}\right)^{-1}$ are computed and used to select PPSWR samples of equal size n_i from each small area. Then PPSWR samples of sizes n_i = 10, 15, 20 and 25 were drawn from each small area based on selection probabilities p_{ij}. This selection probability, computed from a size measure z_{ij} is a linear combination of two auxiliary variables, namely Household size and Area operated (in hectare). The basic design weight calculated as, $w_{ij} = (n_i p_{ij})^{-1}$. Further, two cases were considered for fitting HBNSP models. Case 1- No auxiliary variable is included in the HB models and linking model contains only intercept and random effect (*i.e.*, random mean form of model). Case 2-Available auxiliary variable (Area operated, in hectare) is used as covariate in the HB models and linking model contains intercept, one auxiliary variable and random effect (*i.e.*, random intercept form of model). The prior for hyperparameter $\boldsymbol{\beta}$ was N $(0,10^6)$. The prior for η and σ_v^2 was taken to be IG (0.1,0.1). Gibbs sampling method is implemented with three independent chains each of length 10000; the first 5000 iterations are deleted as “burn-in” periods. To monitor the convergence success potential scale reduction factor $\hat{R}$ is observed. The $\hat{R}$ value close to 1 determines that the MCMC sampler converged very well.

Table 2 presents the average values of RB, RRMSE and CR for the small area predictors defined by HBNSP1 and HBNSP2 methods investigated in design-based simulations under case 1. The average values of RB, RRMSE and CR for HBNSP1 and HBNSP2 under case 2 are reported in Table 3. Figure 1 plots the average values of bias for HBNSP1 and HBNSP2 methods in design-based simulations under case-1 (left side) and case-2 (right side). From these results, it is evident that design bias of survey-weighted predictor HBNSP2 is smaller than HBNSP1. Further, the values of RB for survey-weighted predictor reduces with sample size, which shows the property of design consistency of small area predictor HBNSP2. The RRMSE values are also smaller for HBNSP2 and having the same decreasing trend with increment of small area sample sizes. Investigation on coverage properties of both the models shows that noncoverage rate is higher for HBNSP1 model as compared to the other. As number of areas increases, HBNSP2 shows the better coverage percentage.

Table 2: The average values of percentage relative biases (%RB), percentage relative root mean squared errors (%RRMSE) and percentage coverage rates (%CR) for HBNSP1 and HBNSP2 methods in design-based simulation under case 1

Criterion	Method	$n_i = 10$	$n_i = 15$	$n_i = 20$	$n_i = 25$
RB	HBNSP1	2.52	3.28	4.15	4.10
	HBNSP2	1.97	1.74	1.45	1.35
RRMSE	HBNSP1	24.23	23.02	23.80	24.08
	HBNSP2	23.37	15.57	13.35	12.73
CR	HBNSP1	89	87	83	76
	HBNSP2	91	96	97	97

Table 3: The average values of percentage relative biases (%RB), percentage relative root mean squared errors (%RRMSE) and percentage coverage rates (%CR) for HBNSP1 and HBNSP2 methods in design-based simulation under case 2

Criterion	Method	$n_i = 10$	$n_i = 15$	$n_i = 20$	$n_i = 25$
RB	HBNSP1	3.45	3.77	4.90	4.88
	HBNSP2	2.41	1.67	1.02	1.07
RRMSE	HBNSP1	28.03	24.64	25.52	25.34
	HBNSP2	24.42	17.00	15.41	13.25
CR	HBNSP1	85	84	80	77
	HBNSP2	91	95	96	97

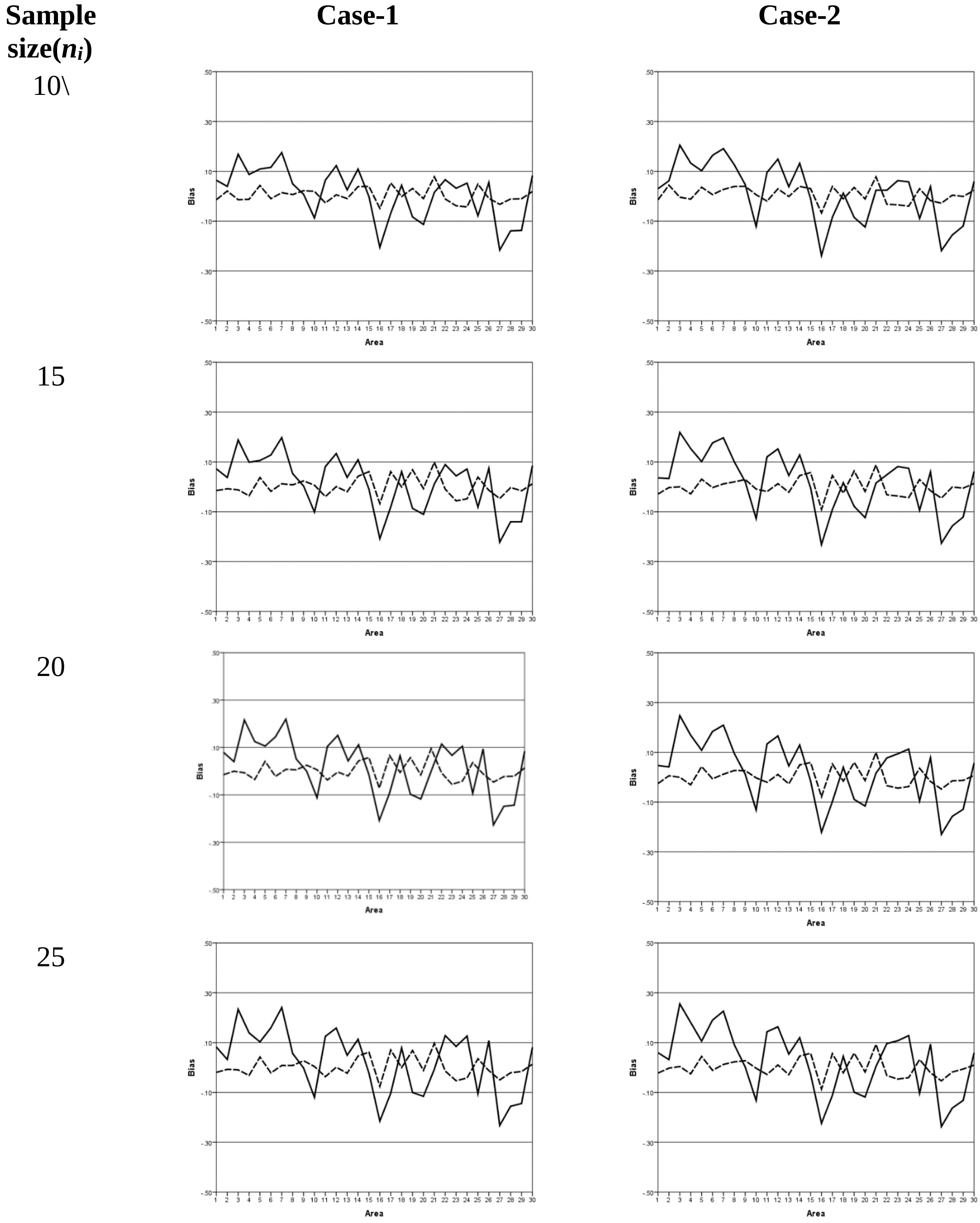


Figure 1: Comparison of bias of HBNSP1 and HBNSP2 (HBNSP1: Solid line, HBNSP2: Dash line) under case 1 (left side) and case 2 (right side)

4. Concluding Remarks

The article describes a spatial nonstationary extension of the area level version of the hierarchical Bayes generalized linear mixed model and considers SAE of proportions under this model. The corresponding predictor is referred to as the spatial nonstationary hierarchical Bayes predictor (HBNSP) for small area proportions. This predictor can account for the presence of spatial nonstationarity in the data where the parameters associated with the model covariates vary spatially.

Empirical results based on simulation studies provide evidence that the proposed HBNSP predictor is more efficient than the alternative hierarchical Bayes predictor under the area level generalized linear mixed model when there is a spatial nonstationarity in the data. The MSE estimation of the HBNSP predictor derived from associated posterior variance also performed reasonably well, with good coverage performance for nominal confidence intervals based on it. It is worth noting that in this article empirical studies were also carried out using survey weights to incorporate the sampling design in SAE. This seems more realistic to implement survey weighted estimation than assuming that the sampling design is customary non-informative.

The Census in India, like in other countries, usually has limited scope in collection of data. It focuses mainly on basic social and demographic information and that too at decennial interval. On the other hand, NSSO conducts regular surveys on several socio-economic indicators, but outcome is restricted to generate national and state level estimates, not administrative units below state because of small sample sizes for such units. Due to emphasis on disaggregate level Sustainable Development Goal indicators, Government of India as well as different State Governments are now struggling with generation of disaggregated level statistics. The SAE is only indispensable alternative to meet the growing demand for such disaggregated level statistics needed for decentralized policy planning. The SAE methodology discussed in this article can be used for calculating disaggregate level estimates of prevalence and proportions which is common in most of the socio-economic and health surveys.

References

Anjoy P., Chandra H. and Basak P. (2019). Estimation of disaggregate-level poverty incidence in Odisha under area-level Hierarchical Bayes small area model. *Social Indicators Research*, **144**, 251-273.

Baldermann C., Salvati N. and Schmid, T. (2018). Robust small area estimation under spatial non-stationarity. *International Statistical Review*. DOI: https://doi.org/10.1111/insr.12245.

Battese G. E., Harter R. M. and Fuller W. A. (1988). An error-component model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Brunsdon C., Fotheringham A. S. and Charlton M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity, Geographical Analysis, 28, 281–298.

Chandra H., Chambers R. and Salvati N. (2019). Small area estimation of survey weighted counts under aggregated level spatial model. *Survey Methodology*, **45**, 31-59.

Chandra H., Kumar S. and Aditya K. (2018). Small area estimation of proportions with different levels of auxiliary data. *Biometrical Journal*, **60**, 395–415.

Chandra H. and Salvati N. (2018). Small area estimation for count data under a spatial dependent aggregated level random effects model. *Communications in Statistics - Theory and Methods*, **47**(**5**), 1234 -1255.

Chandra H., Salvati N. and Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics*, **20**, 30-56.

Cressie N. (1993). *Statistics for Spatial Data*. Wiley, New York.

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.

Gelman A. and Rubin D. (1992). Inference from Iterative Simulation using Multiple Sequences. *Statistical Science*, **7**, 457-511.

Hidiroglou M. A. and You Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodology*, **42**, 41-61.

Jiang J. and Lahiri P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, **53**, 217-243.

Liu B., Lahiri P. and Kalton G. (2014). Hierarchical Bayes modeling of survey-weighted small area proportions. *Survey Methodology*, **40**, 1-13.

Rao J. N. K. and Molina I. (2015). *Small Area Estimation*. 2nd Edition. New York: John Wiley & Sons.

You Y. and Zhou M. Q. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, **37**, 25-37.

APPENDIX

Table A1: Summary statistics of percentage relative biases (%RB), percentage relative root mean squared errors (%RRMSE) and percentage coverage rates (%CR) for HBP and HBNSP methods in model-based simulations for different values of η under spatial nonstationary data generation process for *D*= 100 small areas.

Criterion	RB		RRMSE		CR	
	HBP	HBNSP	HBP	HBNSP	HBP	HBNSP
η=0.5						
Minimum	–9.50	–8.26	2.11	1.88	22	39
Q1	–3.70	–2.46	4.28	3.41	76	91
Mean	–0.97	–0.48	5.64	4.63	81	92
Median	–0.83	–0.50	5.39	4.38	86	95
Q3	0.76	1.09	6.46	5.47	93	98
Maximum	22.30	17.34	23.72	18.52	100	100
η=1						
Minimum	–9.30	–8.01	1.60	1.477	33	41
Q1	–3.48	–2.17	3.53	2.86	85	95
Mean	–0.89	–0.40	5.13	4.24	89	95
Median	–0.86	–0.32	4.72	4.03	94	98
Q3	0.79	1.00	6.06	5.16	98	99
Maximum	23.08	17.11	25.24	18.30	100	100
η=2						
Minimum	–8.73	–7.63	1.36	1.22	39	44
Q1	–3.25	–2.21	2.93	2.62	91	95
Mean	–0.75	–0.36	4.59	4.05	93	95
Median	–0.58	–0.06	4.25	3.73	98	99
Q3	0.79	0.89	5.63	5.05	99	100
Maximum	23.17	19.33	25.13	20.66	100	100
η=4						
Minimum	–8.17	–7.57	1.28	1.15	43	49

Q1	–3.14	–2.56	2.90	2.69	91	94
Mean	–0.60	–0.40	4.62	4.28	93	94
Median	–0.35	–0.24	4.27	3.75	98	99
Q3	0.78	0.83	5.72	5.28	99	99
Maximum	37.61	34.28	39.94	36.31	100	100

Table A2: Summary statistics of percentage relative biases (%RB), percentage relative root mean squared errors (%RRMSE) and percentage coverage rates (%CR) for HBP and HBNSP methods in model-based simulations for different values of η under spatial nonstationary data generation process for *D*= 64 small areas

Criterion	RB		RRMSE		CR	
	HBP	HBNSP	HBP	HBNSP	HBP	HBNSP
η=0.5						
Minimum	–8.21	–6.54	1.61	1.62	64	79
Q1	–2.16	–1.52	3.19	3.12	97	98
Mean	–0.34	0.00	4.18	4.10	97	98
Median	0.23	0.08	4.32	4.05	99	99
Q3	1.34	1.86	4.86	4.71	100	100
Maximum	5.76	5.66	8.63	7.13	100	100
η=1						
Minimum	–7.79	–6.35	1.39	1.49	68	81
Q1	–1.99	–1.45	2.98	2.37	98	99
Mean	–0.26	0.08	3.91	3.83	98	98
Median	0.35	0.29	4.05	3.83	100	99
Q3	1.33	1.86	4.61	4.61	100	100
Maximum	5.30	5.19	8.19	7.29	100	100
η=2						
Minimum	–7.35	–6.25	1.42	1.45	67	77
Q1	–1.76	–1.29	2.87	2.09	99	99
Mean	–0.20	0.12	3.86	3.83	98	98
Median	0.37	0.28	4.08	3.69	100	100
Q3	1.26	1.73	4.72	4.54	100	100
Maximum	4.92	4.54	7.77	7.28	100	100
η=4						
Minimum	–7.02	–5.88	1.58	1.55	68	72
Q1	–1.58	–1.09	3.26	3.42	97	98
Mean	0.30	0.28	4.40	4.36	98	98
Median	0.40	0.36	4.32	4.11	99	99
Q3	1.39	1.44	5.24	4.94	100	100
Maximum	6.25	5.80	9.48	9.45	100	100

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
2022

Printed by : Galaxy Studio & Graphics

Mob: +91 9818 35 2203, +91 9582 94 1203