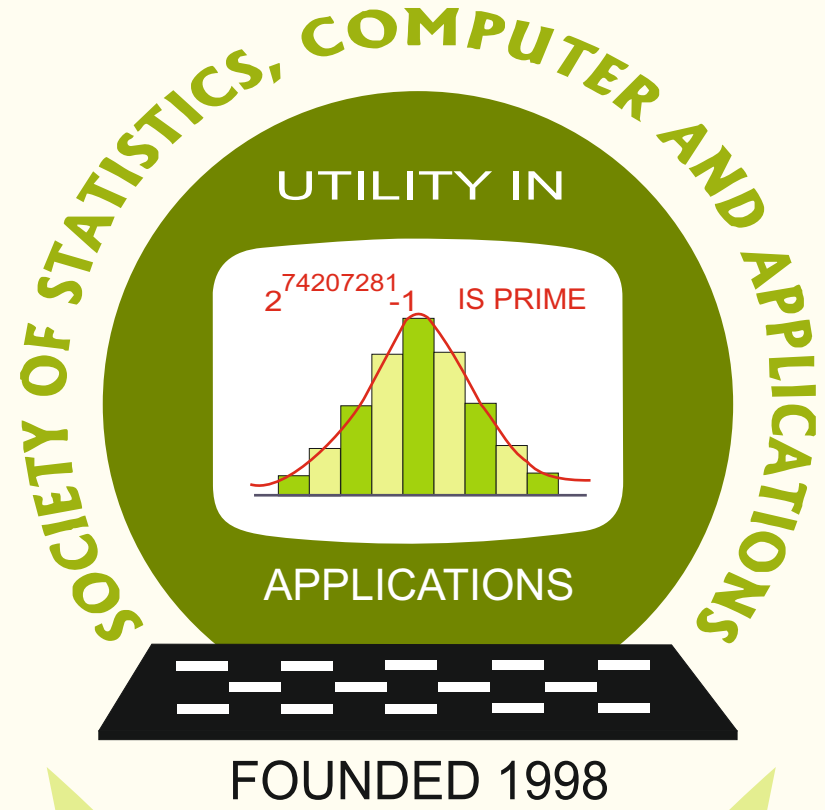
ISBN No.978-81-950383-0-5

Special Proceedings
23rd Annual Conference of SSCA held at ICAR National Academy of Agricultural Research Management, Hyderabad, Telangana

24-28 February 2021



Society of Statistics, Computer and Applications
https://ssca.org.in/
2021

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

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

Secretary
D. Roy Choudhury

Foreign Secretary
Sudhir Gupta

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri
Shibani Roy Choudhury

Council Members

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

Ex-Officio Members (By Designation)

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

Society of Statistics, Computer and Applications, B - 133, Ground Floor, C.R. Park, New Delhi - 110019

ISBN No.978-81-950383-0-5

Special Proceedings 23rd Annual Conference of SSCA held at ICAR National Academy of Agricultural Research Management, Hyderabad, Telangana

24-28 February 2021

Editors

V.K. Gupta

Baidya Nath Mandal

R. Vishnu Vardhan

Rajender Parsad

Ranjit Kumar Paul

Dipak Roy Choudhury

Date of Publication: September 2021

Published By:
Institutional Publisher: Society of Statistics, Computer and Applications
B-133, Ground Floor, C.R. Park, New Delhi-110019
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2021

Printed by: Galaxy Studio & Graphics
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 #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021

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 keeninterest in the applications of Statistics to varied fields like agriculture, biological sciences, medical sciences, financial statistics, and industrial statistics. Since then, the Society has beenperforming several activities and promoting development of theoretical and applied research work in Statistics and Informatics.

One of the major activities of SSCA is to organize national/international conferences annuallyacross 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-third Annual Conference of the SSCA was organized during 24-28 February 2021 at the ICAR National Academy of Agricultural Research Management, Hyderabad, Telangana. This was the first web conference organized online to observe appropriate COVID – 19 protocols. The theme of the conference was *Visionary Innovations in Statistical Theory and Applications (VISTA* 2021*)*. 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, Symbiosis Institute of Management Studies, Pune, Savitri Phule Pune University and Reserve Bank of India made presentations. Two special memorial sessions were organized in the memory of Late Professor Aloke Dey. The speakers in this session were close associates, collaborators, friends and students of Late Prof Aloke Dey. We are pleased to share that the special proceedings have got ISBN #: 978-81-950383-0-5.

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, Rajender Parsad, Ranjit Kumar Paul and Dipak Roy Choudhury as Guest Editors for bringing out these special proceedings. The GuestEditors 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, 16 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. S.K. Soam, Dr. A. Dhandapani, and Dr. S. Ravichandran for organizing such an academically enriching first web conference of the SSCA.

Guest Editors

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

New Delhi
September 2021

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 1- 13

Big Data and Agriculture

B.L.S. PRAKASA RAO
CR Rao Advanced Institute of Research in Mathematics, Statistics and Computer Science, Hyderabad 500046, India

Received: 25 March 2021; Revised: 25 May 2021; Accepted: 28 May 2021

Abstract

This article is based on the presentation "Big Data and Agriculture" delivered as the M.N. Das Memorial Lecture at the 23rd Annual Conference of the Society of Statistics, Computers and Applications "Visionary Innovations in Statistical Theory and Applications (VISTA-2021), Hyderabad. We present several aspects of what is now known as BIG DATA with its advantages as well as pitfalls in general with special reference for agriculture.

Key words: Big data in Agriculture; Massive data; Handling of Big data; Dimension reduction; Computing issues for Big data.

1. Introduction

This article is based on the presentation "Big Data-Agriculture" delivered as the M.N. Das Memorial Lecture at the 23rd Annual Conference of the Society of Statistics, Computers and Applications "Visionary Innovations in Statistical Theory and Applications (VISTA-2021)" held at Hyderabad, India. We present several aspects of what is now known as BIG DATA with its advantages as well as pitfalls. Far from being an exhaustive review of this emerging field, this is an overview from the point of view of a statistician and it is a compilation of ideas of many researchers, organizations and from online sources. For recent surveys on various aspects of Big data as applied to several other areas beside agriculture, the reader is referred to Govindarajulu *et al.* (2015), Prakasa Rao (2015, 2017), Pyne *et al.* (2016), Kitchin (2015), Struiijs *et al.* (2014) and a comprehensive report on the current trends and the future challenges in statistics of Big data published by the London workshop on the future of Statistical Sciences (2014).

Without any doubt, the most discussed current trend in statistics is BIG DATA. Different people think of different things when they hear about Big Data. For statisticians, how to get usable information out of data bases that are so huge and complex that many of the traditional or classical methods cannot handle? For computer scientists, Big Data poses problems of data storage and management, communication and computation. For citizens, Big Data brings up questions of privacy and confidentiality.

Big Data is *relentless*. It is continuously generated on a massive scale. It is generated by online interactions among people, by transactions between people and systems and by sensor-

Corresponding Author: B.L.S. PRAKASA RAO
Email: blsprao@gmail.com

enabled equipment such as aerial sensing technologies (remote sensing), information-sensing mobile devices, wireless sensor networks etc. Big Data is relatable. It can be related, linked and integrated to provide highly detailed information. Such a detail makes it possible, for instance, for banks to introduce individually tailored services and for health care providers to offer personalized medicines. Big data is a class of data sets so large that it becomes difficult to process it using standard methods of data processing. The problems of such data include capture or collection, cleaning, storage, search, sharing, transfer, visualization and analysis. Big data is difficult to work with using most relational data base management systems, desktop statistics and visualization packages. Big Data usually includes data sets with size beyond the ability of commonly used software tools.

When do we say that a data is a Big Data? Is there a way of quantifying the data? We will come back to answering this question later in this article.

Advantage of studying Big Data is that additional information can be derived from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found. For instance, analysis of a large data in marketing a product will lead to information on business trend for that product. Big Data can make important contributions to international development. Analysis of Big Data leads to a cost-effective way to improve decision making in important areas such as health care, economic productivity, crime and security, natural disaster and resource management. Large data sets are encountered in meteorology, genomics, biological and environmental research. They are also present in other areas such as internet search, finance and business informatics. Data sets are big as they are gathered using sensor technologies. There are also examples of Big Data in areas which we can call Big Science and in Science for research. For Government, Big Data is present for climate simulation and analysis and for national security areas. Decoding human genome which took ten years to process earlier can now be done in a week. This is also an example of a Big Data. Human genome data base is another example of a Big Data. A single human genome contains more than 3 billion base pairs. The 1000 Genomes project has 200 terabytes (200 trillion bytes) of data. Human brain data is an example of a Big Data. A single human brain scan consists of data on more than 200,000 voxel locations which could be measured repeatedly at 300 time points.

Big Data sizes are a constantly moving target. It involves increasing volume (amount of data), velocity (speed of data in and out) and variety (range of data types and sources). Big Data are high volume, high velocity and/or high variety information assets. It requires new forms of processing to enable enhanced decision making, insight discovery and process optimization. During the last twenty years, several companies abroad such as Monsanto dealing with agricultural products are adopting to data-driven approach to conduct more targeted services to reduce risks and to improve performance. They are implementing specialized data analytics to collect, store, manage and analyze large data sets. Analyzing these massive data sets help measuring firms risks as well as systemic risks. Analysis of such data requires people who are familiar with sophisticated statistical techniques dealing with productivity and risk management. Big Data are of various types and sizes. Massive amounts of data are hidden in social networks such as Google, Facebook, LinkedIn, Youtube and Twitter. These data reveal numerous individual characteristics and have been exploited by government or official statistics and other agencies.

2. Practice of Big Data in Agriculture

In order to tackle the increasing challenges of agricultural production, the complex agricultural ecosystems need to be better understood. This can happen by means of modern digital technologies that monitor continuously the physical environment producing large quantities of data in an unprecedented pace. The analysis of the big data would enable farmers and companies to extract values from it improving their productivity. Our discussion here is based on the excellent reviews by Carolan (2016), Kamilaris *et al.* (2017) and Coble *et al.* (2018).

Population growth along with socio-economic factors have historically been associated to food shortage. In the last fifty years, the world population has grown from three billion to more than six creating a high demand for food. The global population would increase by more than thirty percent by the year 2050 according to UN which means that a seventy percent increase on food production must be achieved. Land degradation and water contamination, climate change, socio-cultural development, government policies, market fluctuations add uncertainties to food security. These uncertainties challenge agriculture to improve productivity. Advancements in crop growth modeling and yield monitoring together with global navigation satellite systems (GPS) have enabled precise localization of point measurements in the field so that spatial variability maps can be created, a concept known as "precision agriculture".

Agricultural practices are now supported by biotechnology and emerging digital technologies such as remote sensing, cloud computing and internet of things leading to the notion of "smart farming". Smart farming is important for tackling the challenges of agricultural production in terms of productivity, environmental impact, food security and sustainability. To address the challenges of smart farming and sustainable agriculture, the complex multivariate and unpredictable agricultural ecosystems need to be better analyzed and understood. Emerging digital technologies contribute to this understanding by monitoring and measuring continuously various aspects of the physical environment producing large quantities of data in unprecedented pace. This implies the need for large scale collection, storage, pre-processing, modeling and analysis of large amount of data coming from various heterogeneous sources.

Agricultural "big data" creates the necessity for large investments in infrastructure for data storage and processing which need to operate in real-time for some applications such as weather forecasting, monitoring for crop pests and animal diseases. "Big Data Analytics" is the term used to describe a new generation of practices designed so that the farmers and related organizations can extract economic value from very large volumes of data by enabling high-velocity capture, discovery and analysis.

Big data can be used in several agricultural applications. Applications related to estimations of crop production and yields, land mapping, weather forecasting and food security require large amounts of data. Recognition of animals' diseases, plants' poor nutrition require high velocity decision making as well decisions on farmers' productivity, weather forecasting which need to be taken in real time. In these cases, the time horizon for the decisions involved require operational and tactical planning in a short term.

Some applications such as computation of crop insurance premiums, farmers' produc-

tivity improvement require wide variety of data from heterogeneous sources. Furthermore some applications require high reliability of the data involved for studies on diseases of plants/animals.

2.1. Some possible area of applications of Big Data analytics

(1) Platforms enabling supply chain actors to have access to high-quality products and processes enabling crops to be integrated to the international supply chain according to the global needs.

(2) As farmers are sometimes not able to sell harvests due to oversupply and not getting the expected returns from harvests, tools for better predictions for yield and demand have to be developed.

(3) Providing advice and guidance to farmers based on their crops' responsiveness to fertilizers is likely to lead to a more appropriate management of fertilizers use. This could also apply to better use of herbicides and pesticides by the farmers.

(4) Scanning equipment in plants, shipment tracking and retail monitoring of consumers' purchases creates the potential to enhance products' traceability through the supply chain increasing food safety. Since the agricultural production is prone to deterioration over time after harvesting, optimization procedures are essential to minimize losses and maximize quality.

(5) Remote-sensing for large scale land/crop mapping will be critical for monitoring the impact in respect of measuring and achieving the productivity and environmental sustainability targets.

(6) Advanced scientific models and simulations for environmental phenomena could provide a basis for establishing platforms for policy makers assisting in decision making towards sustainability of physical ecosystems.

(7) High throughput screening methods that can offer quantitative analysis of the interaction between plants and their environment with high precision and accuracy.

(8) In advanced countries, self-operating agricultural robots could revolutionize agriculture and its over-all productivity as they may automatically identify and remove weeds and identify and fight pests in the agricultural farms.

(9) Fully automated and data-intensive production systems (such as greenhouses and other indoor led illuminated places) would be on the rise with less use of pesticides, water etc.

Majority of the applications discussed would produce large amounts of big data which could be used by policy makers to balance supply and demand, reduce the negative impact of agriculture on the environment, raise food safety and security and increase productivity. The potential open access of the data to the public can create large number of opportunities for research and development towards smooth and more sustainable farming.

A variety of indicators suggest that the availability of sensors, mapping technology and tracking technologies have changed many farming systems and the management of the

food system as it flows from producers to consumers abroad. Big data has significant potential to address the issues of modern societies including the needs of consumers, financial analysts, bankers, marketing agents, producers and decision makers. Sensor technology and data analytics are now applied abroad to agricultural applications. Agricultural companies are investing in robotics, mechanization and other hardware along with farm management software, sensors and internet of things (IoT).

2.2. Applications in agriculture involving Big Data

(i) **Weather forecasting**: Environmental factors like weather influence crop growth and development as well as demand for agricultural and non-agricultural lands. Agricultural production has spatial yield variability partly because of spatial variability in soil and interaction with weather which is also spatially varied.

(ii) **Crop yield production and crop selection**: Machine learning provides many effective algorithms which can identify input and output relationship in crop selection and yield prediction. Using techniques such as artificial neural networks (ANN), one can do crop selection based on factors such as climate, soil, natural calamities, famine and other inputs.

(iii) **Irrigation systems**: Agriculture consumes a major portion of water. Variability in rainfall, climate change and dropping of the water table creates problems. Using smart irrigation systems, the data collected by the sensors can be used to make better decisions regarding water shortages.

(iv) **Crop disease prediction**: ANN is helpful in predicting pest attacks in advance. Such models deal with noisy and wide variety of data and account for possible factors such as historical data, sattelite/sensor data, field conditions, images of leaves etcetra to effectively learn and predict crop diseases.

(v) **Agricultural policy and trade**: A large quantity of data on production output of crops, changes in import and export costs, market demand and supply, market price trends, cultivation costs, payments for workers, transportation costs and marketing costs could be used by ANN algorithms to predict support prices for different commodities produced by farmers.

Big Data has been celebrated as a "revolution", as "the end of theory" and predictive analytics based on the big data are "crucial" for the end of poverty everywhere. Big Data is considered as the next "Big Thing" in many agricultural circles due to the expanding amount of information collected relating to crop production (big sil data) combined with extensive weather data (big climate data) which together form the backbone of precision agriculture technology. Precision agriculture can play a substantial role in meeting the increasing demand for food and raw materials while ensuring sustainable use of natural resources and the environment.

"Big Data" is still undefined. When do we term a data as Big Data? What was "big" data fifty or sixty years ago is "small" by today's standards. What is "big" today might be considered "small" several years from now.

Bio-technology can also be considered as "big data" as locating the genes for favourable and profitable traits in plants in order to create new seed varieties involves sifting through

billions of base pairs in genomes which is a big data!! Analyzing large amount of data is a difficult issue. One important tool for such analysis is the vector auto-regressive model involving generally at most ten variables and the number of parameters grows quadratically with the size of the model. One solution is to incorporate what is called a sparsity assumption. What it essentially implies that even though there may be infinite number of unknown parameters present in the stochastic model, except for finite number of them, most of the others are negligible and have very little influence on the estimation issues. There are concerns dealing with Big Data such as privacy. We will not discuss this issue in this article.

2.3. When is a data a BIG DATA? (*c.f.* Fokoue (2015); Report of London Workshop (2014))

Big Data comes in various ways, types, shapes, forms and sizes. The dimensionality p of the input space (number of parameters) and the sample size n are usually the main ingredients in characterization of data bigness. Large p small n data sets will require different set of tools from the large n small p sets. Here n is the data size and p the number of unknown parameters/varaibles/covariates. There is no method which performs well on all types of data.

Let us consider a data set $\mathcal{D} = \{(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), \ldots, (\mathbf{x_n}, \mathbf{y_n})\}$ where $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of characteristics/covariates from the input space $\mathcal{X}$ and y_i is the corresponding response. The matrix $\mathbf{X}$ of order $n \times p$ given by

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np}) \end{pmatrix}$$

is the data matrix.

Five aspects of the **data matrix** are important:

(i) the dimension p representing the number of explanatory variables measured; (ii) the sample size n representing the number of observations/sites at which the variables are measured or collected; (iii) The relationship between p and n measured through the ratio of them; (iv) The type of variables measured (categorical, interval, count, ordinal, real-valued, vector-valued, function-valued) and the indication of scales/units of measurement; and (v) The relationship among the columns of the data matrix to check multi-collinearity in the explanatory variables.

2.4. What is meant by "Massive or Big Data" as a function of p?

Suppose we are dealing with a multiple linear regression problem with p covariates or explanatory variables under a Gaussian noise/error. For a model space search for variable selection, we have to find the best subset from among $2^p - 1$ models/sub-models. If $p = 20$, then $2^p - 1$ is about a million; if $p = 30$, then $2^p - 1$ is about a billion; and if $p = 40$, then $2^p - 1$ is about a trillion. Hence any problem with more than $p = 50$ variables is a massive data problem. It involves searching a thousand trillion models which is a huge task even for modern computers. Hence any problem with more than 50 predictor variables can be

called BIG DATA. If the number of predictor variables is more than 100, then it is called a MASSIVE DATA problem.

2.5. What is meant by "Massive or Big Data" as a function of n?

We generally believe that the larger the sample from a population, the better is the inference, due to the law of large numbers. However the computational and statistical complexity in using methods of regression analysis involves inversion of $n \times n$ matrices which is computationally intensive when n is large. It takes $O(n^3)$ number of operations to invert an $n \times n$ matrix. Based on this observation, we might say that the data is observation-massive if $n > 1000$.

2.6. What is meant by "Massive or Big Data" as a function of n/p?

Suppose that we are in a situation with a data where $p > 50$ or $n > 1000$. We have seen that the data can be considered massive in both cases. However the ratio n/p is even more important than n and p taken separately. Let us suppose that we have at least ten observations for each one of the p variables. Hence we have $n > 10p$. Let us also suppose that the information in the data is an increasing function of n. We have the following scenario (c.f. Fokoue (2015)).

(A) $n/p < 1$	Ip	$n << p, n > 1000$	Large p, Large n
(D) $n/p < 1$	Ip	$n << p, n \leq 1000$	Large p, Small n
(B) $1 \leq n/p < 10$	Is	$n > 1000$	Small p, Large n
(E) $1 \leq n/p < 10$	Is	$n \leq 1000$	Small p, Smaller n
(C) $n/p \geq 10$	Ia	$n >> p, n > 1000$	Smaller p, Large n
(F) $n/p \geq 10$	Ia	$n >> p, n \leq 1000$	Smaller p, Small n

(Ip=Information poverty; Is= Information scarcity; Ia= Information abundance)

The BIG DATA problem is with the cases (A) and (D).

For statisticians, Big Data challenges some basic paradigms. The aim is to develop a model that describes how the response variable is related to p other variables or covariates and to determine which variables are important to characterize or explain the relationship. Fitting the model to data involves estimating the parameters from the data and assessing the evidence that they are different from zero indicating the importance of the variable. When $p >> n$, the number of parameters is huge relative to the information about them in the data. Thousands of irrelevant parameters will appear to be statistically significant if one uses small data statistics. Big Data has special features that are not present in the classical data sets. Big Data are characterized by massive sample size and high-dimensionality. Massive sample size allows one to discover hidden patterns associated with small sub-populations. Modeling the intrinsic heterogeneity of Big Data needs better statistical methods. The problems of high-dimensionality in data are noise accumulation, spurious correlation and incidental endogeny. Big Data is often a consequence of aggregation of many data sources corresponding to different sub-populations. Each sub-population might have a unique feature which is

not shared by others. A large sample size enables one to better understand heterogeneity. A mixture model for the population may be appropriate for a Big data. Analyzing Big Data requires simultaneous estimation or testing of a large number of parameters. Errors in inferring on these parameters accumulate when a decision on inference from the data depends on these parameters. Such a noise accumulation is severe in high-dimensional data and it may even dominate the true signal. This is handled by the sparsity assumption. High-dimensionality brings in spurious correlation due to the fact that many uncorrelated random variables may have high sample correlation coefficient in high dimensions. Spurious correlation leads to wrong inferences and hence false results. Unlike spurious correlation, incidental endogeny may be present in Big Data. It is the existence of correlation between variable "unintentionally" as well as due to "high-dimensionality". The former is analogous to finding two persons who look alike but have no genetic relationship where as the latter is similar to meeting an acquaintance by chance in a big city. Endogeny happens due to selection bias, measurement errors and omitted variables. With the advantage of high-tech measurement techniques, it is now possible to collect as many features as possible. This increases the possibility that some of them might be correlated to the residual noise leading to incidental endogeny.

Another reason for incidental endogeny is the following. Big Data are usually aggregated from multiple sources with possibly different data generating schemes. This increase the possibility of selection bias and measurement errors which also leads to possible incidental endogeny. Some statistical methods have been proposed to handle such issues such as penalized quasi-likelihood to handle noise accumulation issue.

Big Data are massive and very high-dimensional and involve large-scale optimization if one wants to use a likelihood or quasi-likelihood approach directly. Optimization with a large number of variables is not only expensive due to computational costs but also suffers from slow numerical rates of convergence and instability. It is also computationally infeasible to apply optimization methods on the raw data. To handle the data both from statistical and computational views, **dimension-reduction** techniques have to be adopted.

2.7. Some issues with Big Data (*c.f.* Fokoue (2015); Buelens *et al.* (2014))

(i) **Batch data against incremental data production :** Big Data is delivered generally in a sequential and incremental manner leading to online learning methods. Online algorithms have the important advantage that the data does not have to be stored in memory. All that is required is in the storage of the built model at the given time in the sense that the stored model is akin to the underlying model. If the sample size n is very large, the data cannot fit into the computer memory and one can consider building a learning method that receives the data sequentially or incrementally rather than trying to load the complete data set into memory. This can be termed as sequentialization. Sequentialization is useful for streaming data and for massive data that is too large to be loaded into memory all at once.

(ii) **Missing values and Imputation schemes:** In most of the cases of massive data, it is quite common to be faced with missing values. One should check at first whether they are missing systematically, that is in a pattern, or if they are missing at random and the rate at which they are missing. Three approaches are suggested to take care of this problem: (a) Deletion which consists of deleting all the rows in the Data matrix that contain any missing

values (b) Central imputation which consists of filling the missing cells of the Data matrix with central tendencies like mean, mode or median and (c) Model-based imputation methods such as EM-algorithm.

(iii) **Inherent lack of structure and importance of pre-processing:** Most of the Big Data is unstructured and needs pre-processing. With the inherently unstructured data like text data, the pre-processing of data leads to data matrices, whose entries are frequencies of terms in the case of text data, that contain too many zeroes leading to the sparsity problem. The sparsity problem in turn leads to modeling issues.

(iv) **Homogeneity versus heterogeneity:** There are massive data sets which have input space homogeneous, that is, all the variables are of the same type.There are other types of Big Data where the input space consists of variables of different types. Such types of data arise when the variables can be categorical, ordinal, interval, count and real-valued.

(v) **Differences in measurement:** It is generally observed that the variables involved are measured on different scales leading to modeling problems. One way to take care of this problem is to perform transformations that project the variables onto the same scale. This is done either by standardization which leads all the variables to have mean zero and variance one or by unitization which consists in transform the variables so that the support for all of them is the unit interval [0,1].

(vi) **Selection bias and quality:** When Big Data are discussed in relation to official statistics, one point of criticism is that Big Data are collected by mechanisms unrelated to probability sampling and are therefore not suitable for production of official statistics. This is mainly because **Big Data sets are not representative of a population of interest**. In other words, they are selective by nature and therefore yield biased results. We will not discuss this issue here. Big Data is often a by-product of some process not primarily aimed at data collection.

Analysis of Big Data is data-driven and not hypothesis-driven. **For Big Data, the coverage is large but incomplete and selective.** It may be unclear what the relevant target population is.

(vii) **Comparison of data sources:** Let us look at a comparison of different data sources for official statistics as compared to Big Data.

Comparison between Sample Survey and big data:

Data Source	Sample Survey	Big Data
Volume	Small	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Sample	Various Organizations
Coverage	Small fraction	Large/Incomplete

(Ref: Buelens *et al.* (2014))

Comparison between Census and Big Data:

Data Source	Census	Big Data
Volume	Large	Big
Velocity	Slow	Fast
Variety	Narrow	Wide
Records	Units	Events or Units
Generator	Administration	Various Organizations
Coverage	Large/Complete	Large/Incomplete

(Ref: Buelens *et al.* (2014))

Big Data can be the single source of data for the production of some statistic about a population of interest. Assessing selectivity of the data is important. Correcting for selectivity can some times be achieved by choosing suitable method of model-based inference (Breiman (2001)). These methods are aimed at predicting values for missing/unobserved units. The results will be biased if specific sub-populations are missing from the Big Data set. Big Data set can be used as auxiliary data set in a procedure mainly based on a sample survey. The possible gain of such an application for the sample survey is likely reduction in sample size and the associated cost. Using small area models, the Big Data can be used as a predictor for survey based measurement.Big Data mechanism can be used as a data collection strategy for sample surveys. Big Data may be used irrespective of selectivity issues as a preliminary survey. Findings obtained from Big Data can be further checked and investigated through sample surveys.

3. Methods of Handling Big Data (*c.f.* Fokue (2015))

(i) **Dimension reduction:** Dimensionality reduction involves the determination of intrinsic dimensionality q of the input space where $q << p$. This can be done by orthogonalization techniques on the input space which reduces the problem to a lower dimension and orthogonal input space leading to variance reduction for the estimator. Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are the methods for dimensionality reduction. However if $p >> n$, then most of these techniques cannot be used directly.

(ii) **Bagging:** As it was observed earlier, it is common in massive data that a single model selected does not lead to optimal prediction. If there is a multicollinearity between the variables which is bound to happen when p is very large, the estimators are unstable and of large variances. Bootstrap aggregation (also called bagging) reduces the variance of the estimators by aggregation of bootstrapped versions of the base estimators.

(iii) **Parelellization:** When the computational complexity for building the base learner is high, the method of bagging becomes inefficient and not practical. One way to avoid this problem is to use parallel processing. Big Data analytics will need parallel processing or parallelization for speeding up computation or to handle massive data that cannot fit into a single computer memory.One way to make statistical procedures more efficient in analysis of Big Data is to paralellize them, that is, to write many algorithms that can run on many computers or many processors at the same time. The method of "Bootstrap" is a standard method for inferring the probability distribution from a sample. It is computationally

intensive. It is ideally suitable for paralellization because it involves generating numerous independent rounds of simulated data. One can use "Bag of Little Bootstraps" (BLB) which generates results comparable to the regular bootstrap but much faster.

(v) **Regularization:** With large p and small n, there exist a multiplicity of solutions for any optimization problem involving Big Data and hence the problem becomes ill-posed. Regularization methods are used to find a feasible optimal solution and one method of regularization is Lagrangian formulation of a constrained version of the problem. LASSO (Tibshirani (1996)) is one such method in high-dimensional data analysis.

(vi) **Assumption of sparsity:** As we noted earlier, thousands of irrelevant parameters will appear to be statistically significant if we use small data statistics for Big Data. In classical statistics, if the data implies occurrence of an event that has one-in-a million chance of occurring, then we are sure it is not by chance and hence consider it statistically significant. But if we are considering a Big Data with a large number of parameters, it is possible for the event to occur by chance and not due to significance of the relationship. Most data sets have only a few strong relationships between variables and everything else is noise. Thus most of the parameters do not matter. This leads to sparsity assumption which is to assume that all but a few parameters are negligible.

(viii) **The problem of "Big n, Big p, Little t":** The speed at which one can process is an important element in analyzing Big Data. Classical statistics was always done in an off-line mode, the size was small and the the time for analysis was essentially unlimited. However, in the era of Big Data things are different.

(ix) **Computing issues for Big Data (Fan *et al.* (2013)):** As was mentioned earlier, the massive or very large sample size of Big data is a challenge for traditional computing infrastructure. Big Data is highly dynamic and not feasible or possible to store in a centralized data base. **The fundamental approach to store and process such data is to "divide and conquer".** The idea is to partition a large problem into more tractable and independent sub-problems. Each sub-problem is tackled in parallel by different processing units. Results from individual sub-problems are then combined to get the final result. "Hadoop" is an example of basic software and programming infrastructure for Big Data processing. "MapReduce" is a programming model for processing large data sets in a parallel fashion. "Cloud Computing" is suitable for storing and processing of Big Data. We are not discussing the problems involved in storage and computation connected with Big Data in this article.

3.1. Why Big Data is in trouble?

Answer: They forgot about Applied Statistics (Jeff Leak, May 7, 2014 "Simply Statistics"). There were articles with titles such as "The parable of Google Flu: traps in big data analysis", "Big data: are we making a big mistake ?", "Google Flu trends: the limits of big data", "Eight (No, Nine!) problems with Big Data" *etc.* All of the articles mentioned here and on-line point out the problems of Big Data such as sampling populations, multiple testing, selection bias and over-fitting beside others.

"There is a tendency for Big Data researcher and more traditional applied statistician to live in two different realms. Big Data offers enormous possibilities for understanding human

interactions at a societal scale with rich spatial and temporal dynamics and for detecting complex interactions and nonlinearities among variables. However traditional "small data" often offer information that is not contained in Big Data" (Lazer *et al.* (2014)).

Acknowledgements

The author thanks all the colleagues whose ideas were used in preparing this overview. Work of the author was supported under the scheme "INSA Senior Scientist" of the Indian National Science Academy (INSA) at the CR Rao Advanced Institute of Mathematics, Statistics and Computer science, Hyderabad, India. The author thanks Prof. V.K. Gupta for inviting him to deliver the M.N. Das Memorial Lecture at the 23rd Annual conference of SSCA held in Hyderabad, India during February 24-28, 2021 where this paper was presented.

References

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, **16(3)**, 199-231.

Buelens, D., Daas, P., Burger, J. and van den Brakel, J. (2014). Selectivity of Big Data. *Discussion paper, Statistics Netherlands.*

Carolan, M. (2016). Publicizing food: Big data, precision agriculture and co-experimental techniques of addition. *Sociologia Ruralis*, DOI:10.1111/soru.12120.

Coble, K. H., Mishra, A. K., Ferell, S. and Griffin, T. (2018). Big data in agriculture: A challenge for the future. *Applied Economic Perspectives and Policy*, **40**, 79-96.

Fakoue, E. (2015). A taxonomy of Big Data for optimal predictive machine learning and data mining. *arXiv.1501.0060v1 [Stat. ML] 3 Jan 2015.*

Fan, J., Han, F., and Liu, H. (2013). Challenges of Big Data analytics. *arXiv.1308.1479v1 [stat.ML] 7 Aug 2013.*

Govindarajulu, V., Raghavan, V. V. and Rao, C. R. (2015). Big Data Analytics. *Handbook of Statistics*, Vol. 33, Elsevier, Amsterdam.

Kamilaris, A., Karatakoullis, A. and Prenafeta-Boldu, F. X. (2017). A review on the practice of big data in agriculture. *Computers and Electronics in Agriculture*, **143**, 23-37.

Kitchin, R. (2015). Big data and official statistics: Opportunities, challenges and risks. *The Programmable City working Paper 9* .

Lazer, D., Kennedy, R., King, G. and Vespignani,, A. (2014). The parable of Google Flu traps in Big Data analysis. *Science.* **343**, 1203-1205.

Leak, J. (2014). Why big data is in trouble; they forgot about applied statistics. *"Simply Statistics"*, May 7, 2014.

Prakasa Rao, B.L.S. (2015). Brief Notes on Big Data: A Cursory Look. *Lecture Notes*, C R Rao Advanced Institute for Mathematics, Statistics and Computer Science, Hyderabad, India, pp.1.

Prakasa Rao, B.L.S. (2017) Brief Notes on Big Data. *Visleshana*, Computer Society of India, Special Interest Group - Big Data Analytics, **1** (3), April-June 2017, pp.9-12.

Pyne, S., Prakasa Rao, B.L.S. and Rao, S.B. (2016). *Big Data Analytics.* Edited by Saumyadipta Pyne, B.L.S. Prakasa Rao and S.B. Rao, Springer (India) Pvt. Ltd., New Delhi.

Struijs, P., Braaksma, B. and jH daas, P. (2014). Official statistics and Big data. *BIG DATA & Society*, DOI: 10.1177/2053951714538417, SAGE (2014).

Tibshirani, R. (1996). Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society*, **Series B 58**, 267-288.

Current trends and future challenges in statistics: Big Data. (2014). *Statistics and Science: A Report of the London Workshop on future of the Statistical Sciences*, pp. 20-25.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 15-29

Cost Minimization under Sporadic Shocks and Healing Impetus when the Healing Stage is Subdivided

Debolina Chatterjee and Jyotirmoy Sarkar
Department of Mathematical Sciences
Indiana University-Purdue University Indianapolis, Indiana, USA

Received: 18 May 2021; Revised: 28 May 2021; Accepted: 31 May 2021

Abstract

A system experiences two kinds of sporadic impacts: valid shocks (VS) that cause damage, and positive interventions (PI) that induce partial healing. The system lifetime is divided into two stages: In Stage 1, healing can occur; but in Stage 2, no healing is possible. Stage 1 is further subdivided into two parts: In the early part, called Stage 1A, healing happens faster than in the later stage, called Stage 1B. The system stays in Stage 1A until the net count of impacts (VS registered *minus* VS nullified) reaches a predetermined threshold m_A; then the system enters Stage 1B and stays there until the net count reaches another predetermined threshold m_1 ($> m_A$). Thereafter, the system enters Stage 2. The system fails when the net count of valid shocks reaches another predetermined higher threshold m_2 ($> m_1$).

We assume that the inter-arrival times between successive VS and those between PI are independent and follow *arbitrary* distributions F and G, respectively. We compute the distributions of the sojourn time in Stage 1 and the failure time of the system using two approaches. We calculate the percentage improvement in the system lifetime after subdividing Stage 1. Finally, we make optimal choices which minimize the expected maintenance cost per unit time for two maintenance policies.

Key words: Reliability; Counting process; Weighted convolution; Mean time to failure; Replacement cost

AMS Subject Classifications: 90B25, 62N05, 60K10

1. Introduction

In machine maintenance and reliability engineering, it is often necessary to study the impacts of external shocks. Other than degradation due to aging, system lifetime is affected by the accumulated damage due to shocks. Because a system failure causes a severe loss, it is preferable to replace a system before it fails, but only after utilizing its potential life to the extent possible. Therefore, we seek optimal replacement policies before the system fails.

Shock models have been studied extensively in the past four decades. They can be

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

classified according to types of shocks, arrival processes of shocks, types of systems, nature of system degradation and so on. A comprehensive literature review is given in Chatterjee and Sarkar (2021); it will not be repeated here, except to draw attention to papers dealing with self-healing such as Dong *et al.* (2020) and Shen *et al.* (2018), together with applications of self-healing mentioned in Lafont *et al.* (2012) (Solid-state lighting devices), Keedy and Feng (2013) (bio-medical application), and Bhuyan and Dewanji (2017) (secondary email account system and battery life of cell phones). Let us mention in some detail only one paper Zhao *et al.* (2018) , which is instrumental in developing our model, though we modify some of the assumptions therein to make our model more realistic.

In Zhao *et al.* (2018), the authors studied a two-stage shock model with a self healing mechanism. A shock that results in a certain degree of damage is called valid; otherwise, it is deemed invalid. A δ-invalid shock is one whose time lag with the preceding shock exceeds a given time-threshold δ. A change point splits the failure process into two stages: In Stage 1, δ-invalid shocks trigger self healing by essentially reducing by one the number of valid shocks. When the cumulative number of net valid shocks exceeds a critical level d, the system enters Stage 2, where it loses its healing ability. The cumulative number of valid shocks increases by one whenever a valid shock arrives; but the self-healing behaviour is triggered only when there is a running trail of k **consecutive** δ**-invalid** shocks, reducing the cumulative number of damage by one. A system fails when the cumulative number of valid shocks reaches a threshold n ($n > d$). Further, the authors considered three preventive maintenance policies and the optimal replacement strategies for each policy after considering the associated costs. In their illustrations, inter-arrival times between successive valid shocks and δ-invalid shocks are exponentially distributed.

In Chatterjee and Sarkar (2021), two types of impacts — valid shocks (VS) and positive interventions (PI) — are considered, with their inter-arrival times having *arbitrary* distributions, and the system lifetime is split into two stages — Stage 1 where it can heal, and Stage 2 where it can not heal. However, healing occurs when the **cumulative** effect of k PIs (not necessarily consecutive) nullify one VS. Furthermore, the PIs need not be δ-invalid. This continues until the system reaches a "change point" beyond which it can no longer heal.

The main focus of the current work is to extend the two-stage model by splitting Stage 1 further into two parts. In the earlier part of Stage 1, called Stage 1A, k_A PIs nullify the damaging effect of one VS. In the later part of Stage 1, called Stage 1B, k_B ($> k_A$) PIs can heal one VS. The system is in Stage 1A until the net VS reaches a threshold m_A; thereafter, it enters Stage 1B. Next, the system reaches the "change point" and enters Stage 2 when the net VS reaches m_1. Therefore, $m_1 - m_A = m_B$ is the net number of VS allowed in Stage 1B.

Previous research considered either healing or degradation, or they have assumed that the shocks/impacts have inter-arrival times exponentially distributed. Although some works mention non-exponential inter-arrival times, they illustrate only exponential examples. As in Chatterjee and Sarkar (2021), here we illustrate with several non-exponential inter-arrival time distributions. As long as we can count the number of VS and PI, we can figure out the distributions of duration of Stage 1 and system lifetime.

Section 2 describes the evolution of the system under shocks and healing; Section 3

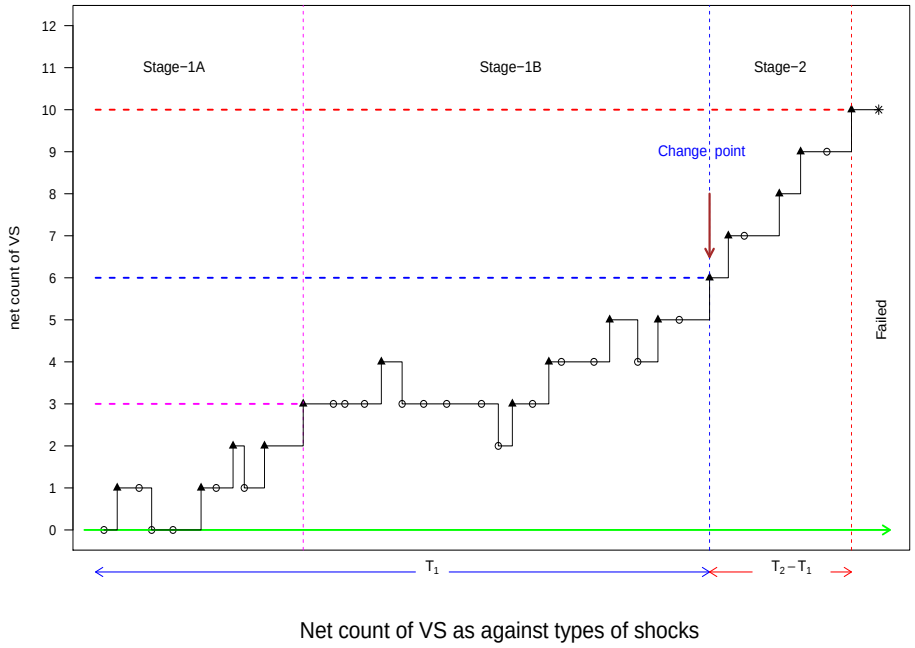


Figure 1: Stages of the system based on net count of VS

illustrates two approaches to compute the distributions of Stage 1 duration and lifetime; Section 4 compares Stage 1 duration and system lifetime between divided Versus undivided Stage 1; Section 5 obtains optimal decisions for two maintenance policies; and finally Section 6 summarizes the main findings of this research.

2. The System Set-up

External impacts to the system are of two types: **Valid Shocks (VS)** that cause damage to the system and **Positive Interventions (PI)** that do not have any damaging effect; on the contrary, the accumulation of a certain (predetermined) number of PIs nullify the effect of one VS. This behaviour is what we call **healing**, which means the net number of VS (VS arrived minus VS nullified by PIs) reduces by one. For simplicity, we assume each VS causes an equal amount of damage. Hence, leaving for future the study of magnitude of damages, here we focus on counting the net number of VS to the system.

The system lifetime is divided into two stages depending on the net VS it receives. In Stage 1, the system has healing ability as described above and the system remains in this stage until the net VS reaches a certain predetermined threshold m_1. Thereafter, the system moves to Stage 2 where it can no longer heal; that is, new PIs do not reduce the net VS anymore. The system fails when net VS reaches another higher threshold m_2. Furthermore, Stage 1 is subdivided into two parts: Stage 1A requires fewer and Stage 1B requires larger number of PI's to nullify one VS.

In Figure 1, the arrival processes of VS (denoted by ▲) and PI (denoted by ∘) illustrate the net count of VS, and hence the stages. Here, $k_A = 2, k_B = 4, m_A = 3, m_1 = 6$, and

$m_2 = 10$. Do not start counting until the first VS arrives. Stop counting PI if the net number of VS, drops to 0. Resume counting once the next VS arrives. The **change point** T_1 defines the transition from Stage 1 (Stages 1A and 1B combined) to Stage 2.

The inter-arrival times of VS, denoted by $X_1, X_2, X_3, \ldots$ are independently and identically distributed (IID) with an *arbitrary* cumulative distribution function (CDF) F. Likewise, $Y_1, Y_2, Y_3, \ldots$, the inter-arrival times of PIs are IID with another *arbitrary* CDF G. Arrival processes of PIs and VS are stochastically independent. Let the duration of system in Stage 1 be denoted as T_1 and the system lifetime be denoted as T_2. The total number of VS in Stage 1 be N_1 and that until failure be N_2. Note that $m_2 - m_1 = N_2 - N_1$ since in Stage 2 there is no healing. Let r denote the number of PIs rendered unused towards healing in Stage 1. Further, let D_1 and D_2 denote the total number of impacts (VS+PI) in Stage 1 and until failure, respectively.

3. Distributions of T_1 and T_2

We describe the underlying stochastic process in terms of two approaches: a counting process approach and a convolution process approach.

3.1. The counting process

Given the constant integers k_A, k_B, m_A, m_1 (hence, $m_B = m_1 - m_A$) and m_2, we describe a simulation of the system status as follows.

Generate a sequence of inter-arrival times of VS $X_1, X_2, X_3, \ldots$ from F, and another sequence of inter-arrival times of PIs $Y_1, Y_2, Y_3, \ldots$ from G. Take the cumulative sums of the two sequences, to obtain the arrival times. Sort these arrival times of the impacts (VS and PI), and associate with each arrival time an indicator **1** to denote VS and **0** to denote PI. Begin counting as soon as the first VS arrives. Ignore all PIs (0) before this moment.

Stage 1A: Count the VS. Arrival of k_A PIs nullify one VS. Compute the net VS as the VS arrived minus VS nullified. If the net VS ever drops to 0, stop counting; resume counting when again another VS arrives. When net VS reaches m_A, the system enters Stage 1B. We keep record of the total number of VS that arrived in Stage 1A, namely N_A.

Stage 1B: In this later part of Stage 1, arrival of a VS increases its count by one, but now to nullify one VS we need k_B PIs ($k_B > k_A$). Again, we stop counting if the net count ever drops down to 0; and resume counting when a new VS arrives. When the net VS reaches m_1, the system enters Stage 2. Let N_B denote the total number of VS that arrive Stage 1B, and let r denote the number of PIs that have arrived in Stage 1, including those that arrived before the very first VS, those which were used for nullifying VS, those which arrived after the net count of VS dropped down to 0 and those that were unused towards healing. Similarly let s denote the total number of PI that arrived in Stage 2.

Stage 2: In this stage the system does not heal. The VS keeps accumulating one by one without being nullified since the PIs have no effect. The system fails when the net VS reaches a threshold m_2.

Thus, in one iteration of the simulation, we obtain as outputs the following quantities:

$N_1 = N_A + N_B$, N_2, T_1, T_2, r; $D_1 = N_1 + r$ which is the total number of impacts in Stage 1 and $D_2 = N_2 + r + s$ which is total number of impacts in Stage 1 and Stage 2 combined. Next, we repeat the above steps for a total of 10^4 iterations.

We will approximate the probability mass functions (PMF) of N_1 and N_2 from the relative frequencies observed in the simulation. Also, based on the simulation, we will directly approximate the probability density function (PDF) of T_1 and T_2. Alternatively, we will reconstruct these PDFs using the PMFs of N_1 and N_2, respectively, through a convolution process as explained below.

3.2. The convolution process

The underlying stochastic process is described below:

In Stage 1A, $(N_A - m_A)$ VS have been nullified by the arrival of $(N_A - m_A) * k_A$ PIs. Similarly in Stage 1B, $(N_B - m_B)$ VS have been nullified by the arrival of $(N_B - m_B) * k_B$ PI and there may have arrived h more PIs, where $0 \leq h \leq k_B - 1$, which are insufficient to nullify another VS. Therefore, in total, the arrival of $Q = (N_A - m_A) * k_A + (N_B - m_B) * k_B$ PIs has contributed towards nullifying $(N_1 - m_1)$ PIs.

Let us denote $S_j = \sum_{i=1}^{j} X_i$ as arrival time of the j-th VS, and $U_j = \sum_{i=1}^{j} Y_i$ as arrival time of the j-th PI. We describe how Stage 1 duration T_1 depends on the number of VS N_1.

(1) The system receives N_1 VS in Stage 1. The arrival time of the $(N_1 - 1)$-st VS is S_{N_1-1} and that of the N_1-th VS is S_{N_1}.

(2) Let U_Q be the arrival time of a PI which causes the $(N_1 - m_1)$-th nullification of a VS, and let U_{Q+h} be the arrival time of the $(Q + h)$-th PI, which do not nullify any VS (where $h = 1, 2, \ldots, k_2 - 1$), since the count of unused PI has not reached k_B yet.

(3) Before the $(N_1 - 1)$-st VS arrives, the Q-th PI has already arrived. Hence, $U_Q < S_{N_1-1}$. Thereafter, until the N_1-th VS arrives, fewer than k_B PI have arrived in Stage 1B. Therefore, the arrival times satisfy the inequality

$$U_Q < S_{N_1-1} < U_{Q+h} < S_{N_1} < U_{Q+k_2} \tag{1}$$

and the sojourn time in Stage 1 is

$$T_1 = S_{N_1} \tag{2}$$

Note that S_j has a CDF given by $F * F * \cdots * F$, the j-fold convolution of F. Moreover, since N_1 is a random stopping time (that determines the end of Stage 1), by Wald's first identity, we have $E(T_1) = E(N_1) \times E(X)$. However, N_1 and T_1 are not independent. Therefore, using a second-order approximation (by matching the mean and the mean squared deviation from the mean), we model

$$T_1 = S_j + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P(N_1 = j) \tag{3}$$

for $j = m_1, m_1 + 1, \ldots$; where $\lambda \in [0, 1]$ depends on F and G. That is, the distribution of T_1 is modeled as a weighted average of adjusted j-fold convolutions of F, where the adjustment equals a suitable fraction of the departure of N_1 from its expectation *times* the expected inter-arrival time between shocks, with weights given by the probability masses $P(N_1 = j)$ for $j = m_1, m_1 + 1, \ldots$.

The above explanations justify the following theorem:

Theorem 1: The distribution of Stage 1 duration is a weighted average of j-fold convolutions of F shifted by $\lambda\,(j - E[N_1])\,E[X]$, where $\lambda \in [0, 1]$ is described below, with weights given by $P\{N_1 = j\}$, the probability that N_1 VS arrive in Stage 1, for $j = m_1, m_1 + 1, \ldots$.

Description of λ*:* For several combinations of inter-arrival time distributions F and G, the fraction λ is numerically obtained via a grid search (with increment 0.01) to match the standard deviations of the distribution of T_1 obtained from the point process and the convolution process.

The system lifetime equals the duration of Stage 1 *plus* $(m_2 - m_1)$ additional inter-arrival times of VS (which is the duration of Stage 2), since the system can no longer heal in Stage 2. Hence, the system lifetime is

$$T_2 = T_1 + \sum_{i=N_1+1}^{N_1+m_2-m_1} X_i = S_{N_2}$$

where $N_2 = N_1 + m_2 - m_1$. Using Theorem 1, we can describe

$$T_2 = S_{j+m_2-m_1} + \lambda\,(j - E[N_1])\,E[X] \text{ with probability } P\{N_1 = j\} \tag{4}$$

for $j = m_1, m_1 + 1, \ldots$.

Corollary 1: The distribution of time to failure T_2 is a weighted average of $(j + m_2 - m_1)$-fold convolution of F shifted by $\lambda\,(j - E[N_1])\,E[X]$, with weights given by $P\{N_1 = j\}$, for $j = m_1, m_1 + 1, \ldots$.

It is noteworthy that the above theorem and corollary are exactly the same as in Chatterjee and Sarkar (2021). The intuition behind this agreement is that both the theorem and the corollary involve the PMF of $N_1 = N_A + N_B$, the total number of VS in stages 1A and 1B combined. Exactly how Stage 1 is subdivided (based on the requirements of healing) is irrelevant to describe the Stage 1 duration or the lifetime.

3.3. Computation and comparison

We shall consider different inter-arrival time distributions for X and Y satisfying $E(X) = 1$ and $E(Y) = 2/3$. The distribution of sojourn time in Stage 1 is found directly by repeating the point process 10^4 times. In Chatterjee and Sarkar (2021), we had considered $k = 3$ and $m_1 = 10$ for illustration. For comparability, here we choose $k_A = 2, k_B = 4$ to keep the overall average number of PIs required to nullify one VS roughly the same and we choose $m_A = 5, m_B = 5$, so that $m_1 = 10$.

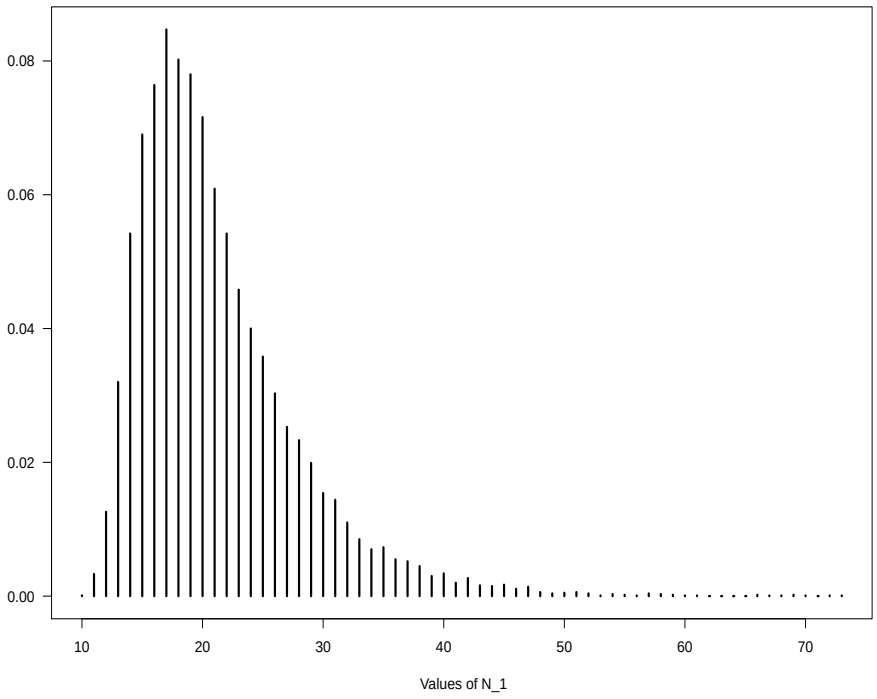


Figure 2: Probability Distribution of $\mathbf{N_1}$ is unimodal with mode 17, $\mathbf{E(N_1) \approx 21}$, $\mathbf{SD(N_1) = 6.52}$, $\mathbf{Q_1 = 17}$, $\mathbf{Q_2 = 20}$, $\mathbf{Q_3 = 24}$, 99-th percentile $\mathbf{(N_1)_{0.99} = 36, P(N_1 > 40) = 0.0057}$.

We emphasize that while similar results hold for all combinations of inter-arrival times, to save space, we will show detailed results for one particular combination of inter-arrival times: $F \equiv$ Weibull (shape $= 2$, scale $= 2/\sqrt{\pi}$) and $G \equiv$ Gamma (shape $= 2$, scale $= 1/3$), such that $E(X) = 1$ and $E(Y) = 2/3$. Figure 3 shows the simulated PMF of N_1.

Figure 3 shows the PDFs of T_1 and T_2 obtained both directly from point process and from adjusted convolution. For T_1, the mean Kullback-Leibler divergence of the adjusted convolution PDF from the point process PDF measures 0.001466, which is very small (with simulated p-value 0.997) and supports Theorem 1. Similarly, for T_2, the mean Kullback-Leibler divergence of 0.00102 (with simulated p-value 0.999) supports Corollary 1. The discrepancies between the two PDFs is at most 0.00035, establishing that the PDFs arising from the point process and the adjusted convolution process are the same,

Let us now consider all combinations of F and G simultaneously. In the Table 1, we show the mean and the standard deviations of T_1 obtained from the two processes for various choices of F and G. Similarly, in Table 2, we show the mean and the standard deviations of T_2. We show the corresponding λ's for each combination of F and G in Table 3.

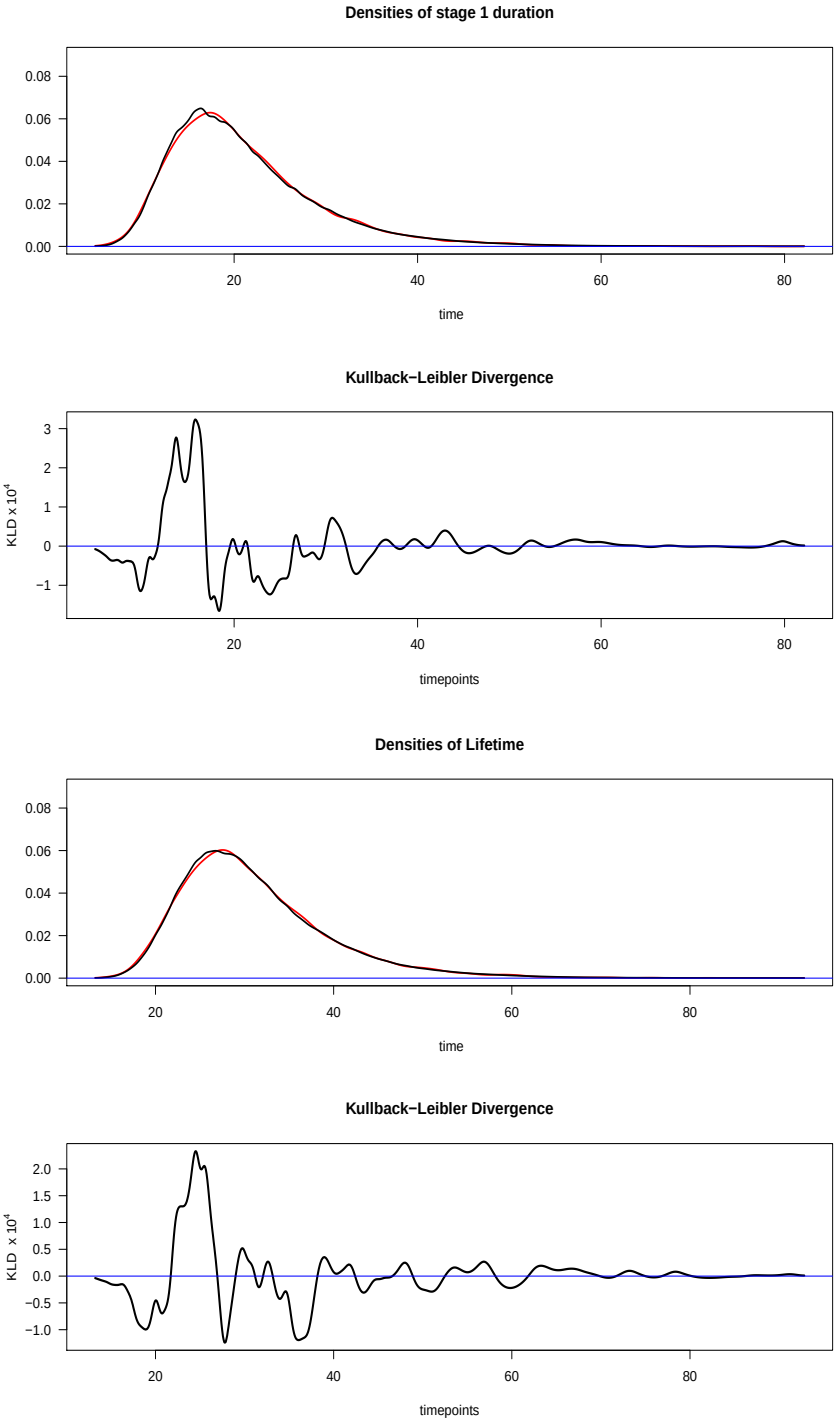


Figure 3: Densities of T_1 (upper) and T_2 (lower) estimated from a point process (red) and an adjusted convolution process (black), with their difference being within 3.5×10^{-4} of 0.

Table 1: For various inter-arrival time distributions satisfying $\mathbf{E(X) = 1}$ and $\mathbf{E(Y) = 2/3}$ the top entries give mean (standard deviation) of Stage 1 duration $\mathbf{T_1}$ according to a point process, and the bottom entries (in *italics*) show the same quantities according to an adjusted convolution process.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $SD \approx 0.12$	Gamma $(2, \frac{1}{3})$ $SD \approx 0.22$	Inv-Gauss $(2/3)$ $SD \approx 0.29$	Exponential $(3/2)$ $SD \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $SD \approx 0.27$	21.51 (8.12) *21.50 (8.15)*	21.17 (8.35) *21.15 (8.38)*	21.01 (8.50) *21.01 (8.53)*	20.34 (8.39) *20.31 (8.39)*
Gamma $(2, \frac{1}{2})$ $SD \approx 0.50$	20.77 (9.13) *20.76 (9.12)*	20.42 (9.07) *20.44 (9.07)*	20.43 (9.21) *20.43 (9.20)*	19.82 (9.05) *19.81 (9.01)*
Inv-Gauss (1) $SD \approx 1.00$	19.53 (10.16) *19.61 (10.16)*	19.26 (10.01) *(19.36 (10.01)*	19.18 (10.11) *19.27 (10.07)*	18.88 (9.97) *18.98 (9.97)*
Exponential (1) $SD \approx 1.00$	19.59 (10.36) *19.59 (10.37)*	19.41 (10.36) *19.37 (10.39)*	19.46 (10.59) *19.43 (10.58)*	18.91 (10.04) *18.89 (10.06)*

4. Comparison with Undivided Stage 1

In Table 3, we compare the means of the Stage 1 duration, showing the percentage change, between the divided Stage 1 studied here and the undivided Stage 1 studied in Chatterjee and Sarkar (2021). We also report the λ's obtained in the current research and compare them to the λ's reported in Chatterjee and Sarkar (2021).

From Table 3, we identify a trend in the values of λ as we scan through the rows and the columns. For a particular choice of F in a row, as we look from left to right across the columns, we see that λ decreases. A closer look at the corresponding standard deviations reveals that λ decreases as the standard deviation of G increases. Similarly, for a fixed choice of G in a particular column, as we go from top to bottom down the rows, we see that λ increases as the standard deviation of F increases. This led us to believe that λ is a function of the ratio of the standard deviations of F and G. When we plotted λ against σ_F/σ_G, we noticed a non-linear relationship. Thereafter, we fitted a linear regression of λ on $\log(\sigma_F/\sigma_G)$ with slope $= 0.11833$, intercept $= 0.20644$ and adjusted coefficient of determination of 0.832.

Let us look at the change in λ before and after the subdivision of Stage 1. When the SD of F is ≈ 0.25 or ≈ 0.50, the λ's for the divided Stage 1 is about one-half to three-fifths of the λ's from the undivided Stage 1; but when the SD of F is ≈ 1, the λ's for the divided Stage 1 is about two-thirds of the λ's from the undivided Stage 1.

In Table 4, we compare the means of the lifetimes, showing the percentage change, between the divided Stage 1 studied here and the undivided Stage 1 studied in Chatterjee and Sarkar (2021).

Table 2: For various inter-arrival time distributions satisfying $\mathbf{E(X) = 1}$ and $\mathbf{E(Y) = 2/3}$ the top entries give mean (standard deviation) of system lifetime $\mathbf{T_2}$ according to a point process, and the bottom entries (in *italics*) show the same quantities according to an adjusted convolution process.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $SD \approx 0.12$	Gamma $(2, \frac{1}{3})$ $SD \approx 0.22$	Inv-Gauss $(2/3)$ $SD \approx 0.29$	Exponential $(3/2)$ $SD \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $SD \approx 0.27$	31.48 (8.27) *31.50 (8.31)*	31.14 (8.51) *31.15 (8.55)*	31.01 (8.65) *31.03 (8.69)*	30.32 (8.58) *30.33 (8.55)*
Gamma $(2, \frac{1}{2})$ $SD \approx 0.50$	30.77 (9.44) *30.76 (9.39)*	30.44 (9.34) *30.45 (9.33)*	30.42 (9.48) *30.43 (9.47)*	29.82 (9.34) *29.80 (9.28)*
Inv-Gauss (1) $SD \approx 1.00$	29.45 (10.69) *29.55 (10.67)*	29.23 (10.53) *29.32 (10.51)*	29.17 (10.66) *29.23 (10.57)*	28.83 (10.54) *28.91 (10.49)*
Exponential (1) $SD \approx 1.00$	29.62 (10.80) *29.57 (10.89)*	29.42 (10.82) *29.34 (10.88)*	29.51 (11.07) *29.43 (11.08)*	28.95 (10.53) *28.90 (10.58)*

5. Preventive Maintenance Policies

System failure being disruptive to the production process and too expensive to recover from, oftentimes a maintenance engineer must intervene to replace a functioning unit. Clearly there is a tension between utilizing the remaining lifetime of the system and the prevention of failure. We consider here two types of preventive maintenance policies. Let c_{p_A} be the cost of replacement in Stage 1A, c_{p_B} in Stage 1B, c_{p_2} in Stage 2, and c_f after failure. Furthermore, we assume that the costs of replacement is the same throughout Stage 1, because the healing rate ought not affect the cost of replacement. We consider $c_{p_A} = c_{p_B} < c_{p_2} << c_f$ with justification as follows: In Stage 1, the system is young, and so an early replacement will incur a loss; if we replace in early part of Stage 2, we are not utilizing the system lifetime sufficiently, but the system has already aged, and so maintenance/repair at this stage will cost more. Therefore, we find it logical to consider that replacement cost in Stage 2 is higher than that in Stage 1. Finally, a system failure is highly expensive. Furthermore, there is an initial cost c_0 of setting up a new system. For illustration, we choose $c_0 = 100, c_{p_A} = 10, c_{p_B} = 10, c_{p_2} = 15, c_f = 200$. In Figures 4 and 5, we illustrate decision making when $F \equiv$ Weibull $(2, 2/\sqrt{\pi})$ and $G \equiv$ Gamma $(2, 1/3)$, such that $E(X) = 1$ and $E(Y) = 2/3$.

5.1. Maintenance Policy 1

Suppose that a monitoring equipment can detect the arrival of an impetus, but cannot distinguish between a VS and a PI, nor can it identify whether the system is in Stage 1 or Stage 2. The system will be replaced when it has failed or has experienced a specified number of impetus N (the sum of VS and PI).

Within one cycle (between two successive replacements of the system), the total cost of replacement C_1 under Policy 1 is a random variable taking three possible values:

Table 3: For various inter-arrival time distributions satisfying $\mathbf{E(X) = 1}$ and $\mathbf{E(Y) = 2/3}$, the top entries give the mean duration of $\mathbf{T_1}$ for divided Stage 1 (undivided Stage 1), the middle row gives the % increase in $\mathbf{T_1}$ after subdividing Stage 1, and the third row gives the multiplier λ of adjusted convolution for the divided Stage 1 (undivided Stage 1).

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $SD \approx 0.12$	Gamma $(2, \frac{1}{3})$ $SD \approx 0.22$	Inv-Gauss $(2/3)$ $SD \approx 0.29$	Exponential $(3/2)$ $SD \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $SD \approx 0.27$	21.51 (17.96) $\approx 19.76\%$ $\lambda = 0.25(0.50)$	21.17 (17.97) $\approx 17.8\%$ $\lambda = 0.21(0.40)$	21.01 (17.98) $\approx 16.85\%$ $\lambda = 0.18(0.35)$	20.34 (17.93) $\approx 13.44\%$ $\lambda = 0.16(0.29)$
Gamma $(2, \frac{1}{2})$ $SD \approx 0.50$	20.77 (17.92) $\approx 15.90\%$ $\lambda = 0.31(0.59)$	20.42 (17.91) $\approx 14.00\%$ $\lambda = 0.27(0.50)$	20.43 (17.93) $\approx 13.94\%$ $\lambda = 0.25(0.46)$	19.82 (17.85) $\approx 11.03\%$ $\lambda = 0.22(0.38)$
Inv-Gauss (1) $SD \approx 1.00$	19.53 (17.63) $\approx 10.78\%$ $\lambda = 0.43(0.68)$	19.26 (17.61) $\approx 9.37\%$ $\lambda = 0.40(0.61)$	19.18 (17.67) $\approx 8.55\%$ $\lambda = 0.37(0.58)$	18.88 (17.61) $\approx 7.21\%$ $\lambda = 0.36(0.52)$
Exponential (1) $SD \approx 1.00$	19.59 (17.85) $\approx 9.75\%$ $\lambda = 0.45(0.69)$	19.41 (17.87) $\approx 8.62\%$ $\lambda = 0.43(0.63)$	19.46 (17.92) $\approx 8.59\%$ $\lambda = 0.39(0.60)$	18.91 (17.82) $\approx 6.12\%$ $\lambda = 0.37(0.54)$

Table 4: For various inter-arrival time distributions satisfying $\mathbf{E(X) = 1}$ and $\mathbf{E(Y) = 2/3}$, the top row gives the mean duration of $\mathbf{T_2}$ for divided Stage 1 (undivided Stage 1), the bottom row gives approximate % increase in mean $\mathbf{T_2}$ after subdividing Stage 1.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $SD \approx 0.12$	Gamma $(2, \frac{1}{3})$ $SD \approx 0.22$	Inv-Gauss $(2/3)$ $SD \approx 0.29$	Exponential $(3/2)$ $SD \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $SD \approx 0.27$	31.48 (27.94) $\approx 12.67\%$	31.14 (27.96) $\approx 11.37\%$	31.01 (27.96) $\approx 10.91\%$	30.02 (27.92) $\approx 7.52\%$
Gamma $(2, \frac{1}{2})$ $SD \approx 0.50$	30.77 (27.91) $\approx 10.25\%$	30.44 (27.92) $\approx 9.03\%$	30.42 (27.93) $\approx 8.92\%$	29.82 (27.86) $\approx 7.03\%$
Inv-Gauss (1) $SD \approx 1.00$	29.45 (27.57) $\approx 6.82\%$	29.23 (27.58) $\approx 5.98\%$	29.17 (27.60) $\approx 5.69\%$	29.83 (27.54) $\approx 8.32\%$
Exponential (1) $SD \approx 1.00$	29.62 (27.86) $\approx 6.32\%$	29.42 (27.90) $\approx 5.45\%$	29.51 (27.93) $\approx 5.66\%$	28.95 (27.82) $\approx 4.06\%$

(1) c_{p_1}, if N impetus arrive while the system is still in Stage 1, with an associated probability of $P(D_1 > N)$.

(2) c_{p_2}, if the system has already moved from Stage 1 to Stage 2 and the N impetus have arrived before system failure, with an associated probability of $P(D_1 \leq N < D_2)$.

(3) c_f, if the system has already failed before the arrival of N impetus, with an associated probability of $P(D_2 \leq N)$.

Hence, the expected cost (C) under Policy 1, is given by

$$E(C|\text{Policy 1}) = c_{p_1}\, P(D_1 > N) + c_{p_2}\, P(D_1 \leq N < D_2) + c_f\, P(D_2 \leq N) \tag{5}$$

and, writing W_j as the arrival time of the j-th impact (either VS or PI), the expected cycle time (CT) is given by

$$\begin{aligned} E(CT|\text{Policy 1}) = E[\min\{W_N, W_{D_2}\}] &= E[W_N|D_1 > N]\ P(D_1 > N) \\ &+ E[W_N|D_1 \leq N < D_2]\ P(D_1 \leq N < D_2) \\ &+ E[W_{D_2}|D_2 \leq N]\ P(D_2 \leq N). \end{aligned} \tag{6}$$

Therefore, the expected cost per unit time is the ratio

$$E(C|\text{Policy 1})/E(CT|\text{Policy 1}) \tag{7}$$

which we must minimize by choosing N.

For the example considered, Figure 4 shows that the expected cost per unit time is minimized when we choose $N = 56$. Moreover, note that for any other choice of N in the vicinity of the optimal value 56, say between 50 and 60, the expected cost per unit time increases only slightly (no more than 3%). Such a robustness result allows us to rely on the optimal value even when the inter-arrival time distributions deviate slightly from the stated ones. Table 5 documents the optimal choices for other combinations of inter-arrival times.

5.2. Maintenance Policy 2

Suppose that the monitoring equipment can identify the stages of the system. If the system is in Stage 1, we do not replace it at all. After the system enters Stage 2, if the system is still functioning for an additional t units of time, we replace it immediately at epoch $T_1 + t$; otherwise, we replace the system immediately on failure during $[T_1, T_1 + t)$. Our objective is to determine an optimum additional time t in Stage 2. To do so, we minimize the expected cost per unit time, where the expected cost (C) is given by

$$E(C|\text{Policy 2}) = c_{p_2} P(T_2 > T_1 + t) + c_f P(T_2 \leq T_1 + t) \tag{8}$$

and the expected length of the cycle time (CT) is

$$E(CT|\text{Policy 2}) = E(min(T_2, T_1 + t)) = E(T_1) + E(min(T_2 - T_1, t)) \tag{9}$$

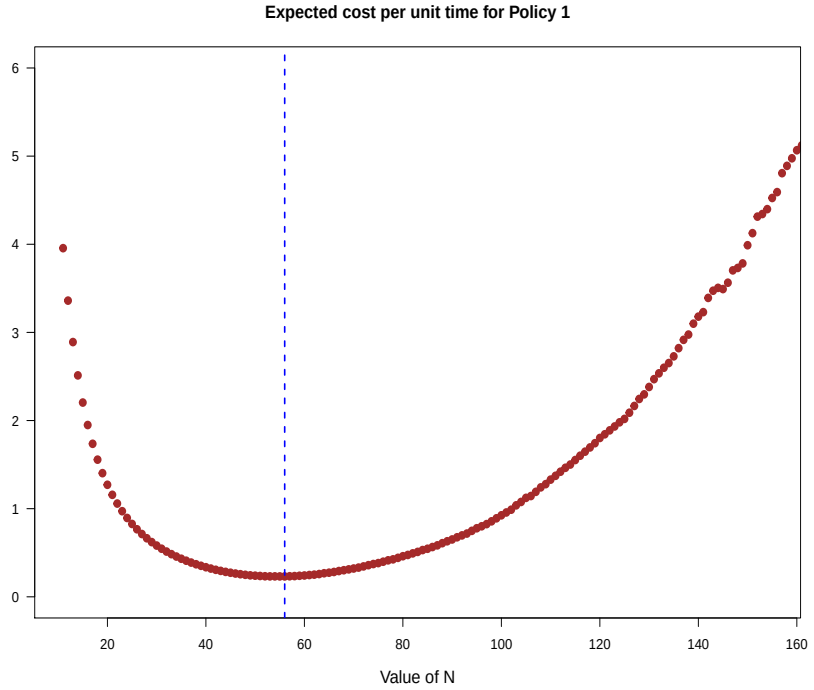


Figure 4: Under Policy 1, when F ≡ Weibull (2, $2/\sqrt{\pi}$) and G ≡ Gamma (2, 1/3), and cost parameters are $c_0 = 100$, $c_{p_{1a}} = 10$,$c_{p_{1b}} = 10$, $c_{p_2} = 15$, $c_f = 200$, the maintenance cost per unit time is minimized when we choose N = 56.

We wish to minimize the expected cost per unit time

$$E(C|\text{Policy 2})/E(CT|\text{Policy 2}) \tag{10}$$

by choosing t. Under Policy 2, the assumed cost parameters, and $F \equiv$ Weibull $(2, 2/\sqrt{\pi})$ and $G \equiv$ Gamma $(2, 1/3)$, Figure 5 shows that the expected cost per unit time is minimized at $t = 6.6$. In fact, we identified this optimal t value via a grid search between the first and the 99-th percentiles of system lifetime with an increment of 0.05. This choice suffices because any other choice of t in the interval $[6, 7]$ increases the cost per unit time only marginally.

Table 5 gives the summary of the optimal choices of N and t for Policy 1 and Policy 2, respectively, for different choices of F and G and for cost parameters $c_0 = 100, c_{p_{1a}} = 10, c_{p_{1b}} = 10, c_{p_2} = 15, c_f = 200$.

We see that for policy 1, the total number of impacts N is only 0-3 impacts more than the optimal values of N when Stage 1 was not divided. This close agreement should not come as a surprise because, for our choice of (k_A, k_B) and (m_A, m_B), the average impact throughout the entire undivided Stage 1 is comparable to that in the undivided Stage 1 case. Hence, there is only a negligible amount of change in N due to subdivision. Similarly, for policy 2, there is no significant change in t, because the choice of t depends only on the arrival rate of VS in Stage 2 and not at all on the subdivision of Stage 1 to accommodate varying rates of healing.

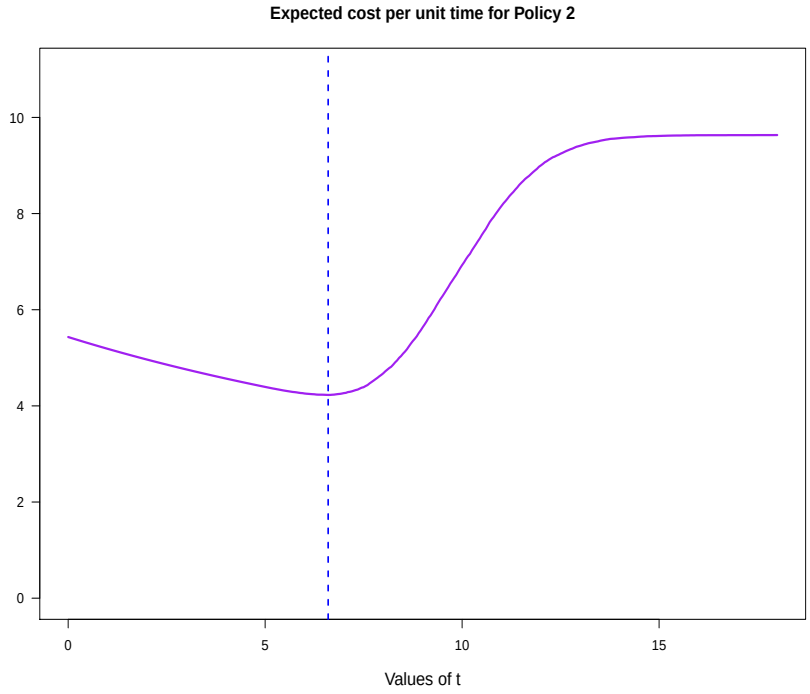


Figure 5: Under Policy 2, when F ≡ Weibull (2, $2/\sqrt{\pi}$), G ≡ Gamma (2, 1/3), and cost parameters are $c_0 = 100$, $c_{p_{1a}} = 10$,$c_{p_{1b}} = 10$, $c_{p_2} = 15$, $c_f = 200$, the maintenance cost per unit time is minimized when we choose t = 6.6.

Table 5: For various inter-arrival time distributions F and G satisfying E(X) = 1 and E(Y) = 2/3, to minimize the maintenance cost per unit time, the optimal N for Policy 1 is shown in the first row, and the optimal t for Policy 2 is shown in the second row.

VS \ PI	Weibull $(2, \frac{4}{3\sqrt{\pi}})$ $SD \approx 0.12$	Gamma $(2, \frac{1}{3})$ $SD \approx 0.22$	Inv-Gauss $(2/3)$ $SD \approx 0.29$	Exponential $(3/2)$ $SD \approx 0.40$
Weibull $(2, \frac{2}{\sqrt{\pi}})$ $SD \approx 0.27$	$N = 55$ $t = 6.45$	$N = 56$ $t = 6.60$	$N = 53$ $t = 6.55$	$N = 52$ $t = 6.60$
Gamma $(2, \frac{1}{2})$ $SD \approx 0.50$	$N = 54$ $t = 5.70$	$N = 51$ $t = 5.75$	$N = 52$ $t = 5.80$	$N = 50$ $t = 5.85$
Inv-Gauss (1) $SD \approx 1.00$	$N = 51$ $t = 4.85$	$N = 50$ $t = 4.65$	$N = 49$ $t = 5.05$	$N = 49$ $t = 4.90$
Exponential (1) $SD \approx 1.00$	$N = 50$ $t = 4.75$	$N = 50$ $t = 4.65$	$N = 50$ $t = 4.65$	$N = 49$ $t = 4.80$

6. Summary and Future Work

In this paper, we subdivided Stage 1 into two parts: initially the system heals at a faster rate requiring fewer PIs to nullify one VS; but once enough net VS have accumulated, more PIs are needed to nullify one VS. We derived the distributions of Stage 1 duration and the system lifetime for any *arbitrary* inter-arrival time distributions of VS and PI, in contrast to only exponential distributions commonly assumed in the literature. Given a prefixed net number of shocks that the system can withstand in various stages, we can work out the distributions of Stage 1 duration T_1 and lifetime T_2 using a point process or an adjusted convolution process with an adjustment factor λ that is approximately a linear function of the logarithm of the ratio of standard deviations of F and G. The theoretical investigation of λ remains an open problem. Moreover, in this research we found that subdivision of Stage 1 leads to an increase in the Stage 1 duration, and hence the system lifetime.

In future, we like to consider varying magnitudes of VS and PI and allow natural system degradation according to some stochastic process. Our simple counting approach suffices to make optimal decisions that minimize maintenance costs per unit time for various inter-arrival distributions and various maintenance policies with different costs of replacement in different stages. This realistic set up promotes an utmost utilization of system lifetime.

Acknowledgements

We thank our colleagues for some discussions and feedback.

References

Bhuyan, P. and Dewanji, A. (2017). Estimation of reliability with cumulative stress and strength degradation. *Statistics*, **51(4)**, 766–781. DOI:10.1080/02331888.2016.1277224

Chatterjee, D. and Sarkar, J. (2021). Optimal replacement policies for systems under sporadic impacts that inflict damage or trigger healing. *Submitted*

Dong, W., Liu, S., Cao, Y., Javed, S. A. and Du, Y. (2020). Reliability modeling and optimal random preventive maintenance policy for parallel systems with damage self-healing. *Computers & Industrial Engineering*, **142**. DOI:10.1016/j.cie.2020.106359

Keedy, E. and Feng, Q. (2013). Reliability analysis and customized preventive maintenance policies for stents with stochastic dependent competing risk processes. *IEEE Transactions on Reliability*, **62(4)**, 887–897. DOI:10.1109/TR.2013.2285045

Lafont, U., van Zeijl, H. and van der Zwaag, S. (2012). Increasing the reliability of solid state lighting systems via self-healing approaches: A review. *Microelectronics Reliability*, **52(1)**, 71–89. DOI:10.1016/j.microrel.2011.08.013

Shen, J., Cui, L. and Yi, H. (2018). System performance of damage self-healing systems under random shocks by using discrete state method. *Computers & Industrial Engineering*, **125**, 124–134. DOI:10.1016/j.cie.2018.08.013

Zhao, X., Guo, X. and Wang, X. (2018). Reliability and maintenance policies for a two-stage shock model with self-healing mechanism. *Reliability Engineering & System Safety*, **172**, 185–194. DOI:10.1016/j.ress.2017.12.013

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 31-47

Predictive Modelling of Lapsation of Life Insurance Policies in India

Gurprit Grover[1], Vajala Ravi[2], Richa Saini[1] and Manoj Kumar Varshney[3]

[1]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, India*
[2]*Department of Statistics, Lady Shri Ram College for Women, University of Delhi, India.*
[3]*Department of Statistics, Hindu College, University of Delhi, India*

Received: 06 April 2021; Revised: 29 May 2021; Accepted: 03 June 2021

Abstract

Retaining customers is the biggest challenge in Indian insurance sector. Customer Relationship Management (CRM) department in every company plays a role of a platform from customer to company which informs insurance company about the needs of customers to be satisfied. Moreover, all the insurance companies are facing the problem of low persistency or high lapsation. Predictive modelling helps in classification of any policy as in force or lapsed. Predictive classification may be of help to insurer in identifying the groups of insured with various characteristic which may not frequently experience the lapsation. Such information can be found as useful for targeting the segment of society which is insurance minded and afford to keep policy in force. We have applied four different techniques for predictive modelling namely Cox PH model, Gompertz law of mortality, Bayesian networks using naïve Bayesian technique and random forests technique and the best fitted model is identified based on confusion matrix and error rate of misclassification. On applying the best fitted model to the data set of 3663 policies, it is found that nearly 54 per cent policies are classified as lapsed and 46 per cent policies are classified as in force. The major contributors in classifying the policies are age of the policyholders, sum assured (SA), policy term, occupation and income of the policyholders.

Key words: Predictive modelling; Naïve Bayesian model; Random forest; Gompertz; Cox PH model; Persistency.

1. Introduction

1.1. Outline

Loss of customers' confidence in company is distressing for any type of business. From a management perspective, quality of sales is to be ensured highly. For life insurance Company, quality of sales is ensured if its customers are persistent or retained with the same company till the end of the term of insurance contract. If policies are being terminated due to non-payment of premiums then it costs high to all who are involved in a contract including insured, insurer and the agent. Over the decade, the severe problem of low persistency has been experienced by every life insurance company in Indian sector. CRM understands the need of customers and provides the same information to the company to retain its customers. The services provided by CRM mainly involve consultation, execution, subcontracting and

[1]Disclaimer: The views expressed in the paper are those of the authors and not necessarily those of the institution to which they belong. Moreover, data have also been obtained solely by surveys of different types and do not pertain to any specific institution.

Corresponding Author: Richa Saini
Email Id: richarawat55@gmail.com

training. But still the insurance companies are facing the challenge of customers' churn. For handling the problem of low persistency, a large sample of policies that are procured in a particular financial year, is observed every year until the fifth year after inception. For a deep dive in to the problem, it is important to identify the contribution of each working factor and better techniques for predictive modelling. In such a case, data mining techniques have evidently established the reliable results with great accuracy. These techniques are valid in our case too. In the present study we have made a comparison of different techniques used for predictive modelling and have applied the famous Cox PH model, Gompertz Curve, Naïve Bayesian model and Random forest technique. These four models are chosen to compare the convention with the advancement. Former two models are the conventional survival and actuarial models which have been established as most popular for modelling of survival data and often they are used for predictions also. For classification purpose, these two models have utility but the algorithm is complex. Over the last two decades, the latter two models have gained much attention by researchers for predictive modelling and classification. These models are the results of advancement in machine learning algorithms. These models do not assume any form of underlying distribution or specific structure but learn the features of the best fit model from the data itself. The analysis is performed using the statistical software *R*.

1.2. Review of literature

Previously many pioneering investigations have been conducted for predictive modelling of customers' churn in various fields including insurance. (Tirenni *et al*., 2007) drew out the general methodology for categorization of customers according to their life time value and also forecasted their lifetime values based on demographic and behavioral characteristics. They addressed the problem of lifetime values of airline customers using decision trees and classified the customers as long, medium and short term customers. Their predictions are based on limited information. (Jasek *et al*., 2018) provided the life time value models in the field of E commerce and discussed their forecasting abilities. They applied extended Pareto, Markov chain and status quo models and compared their results. (Huigevoort, 2015) carried out the predictive modelling of customer churn for a health insurance company. He utilized the four data mining techniques namely logistic regression, decision trees, neural network and support vector machine (SVM) to identify important churning variables and characteristics. (Adebiyi *et al*., 2016) explored the blend of two models namely Analytic Hierarchy Process (AHP) and Markov chains (MCM) for tackling the problem of customers' churn. Their study recommended the organizational strategies that reverse the churn alternatives with high priority and improves service delivery. (Zhang *et al*., 2017) employed the Deep and Shallow model for predictive modelling of insurance churn. Their proposed model yields enhanced performance as compared to Deep models or Shallow models, if applied individually. (Lariviere and Van Den Poel, 2005) predicted customers' withholding and profitability using Random forest and Regression forest techniques. In their study they considered three important measures of customer outcome *viz.* subsequent buy, fractional defection and customers' prosperity by employing random forest and regression techniques. They concluded that the results will be improved if behavioral outcome variables are also taken into consideration along with demographic factors. (Bandyopadhyay *et al*., 2014) offered a machine learning approach based on Bayesian networks (BN) with an application on Electronic health data (EHD) to forecast the probability of encountering a cardiovascular incident within next five years. In EHD, censoring exists as an inevitable feature and they described how to alter both the modelling and estimation techniques to account for censoring. Their proposed model is an improvement over Cox PH model or Bayesian networks with informal approach to right censoring. (Onisko *et al*., 2001) proposed a method that utilizes the Noisy OR gates to minimize the data requirements in learning

conditional probabilities. Their proposed method is proved to be better than the simple multiple disorders model and a single disorder diagnosis model. (Rudolph, 2002) presented in first of his paper, an analysis of part of the long-term care insurance portfolio using Cox Proportional Hazard model to estimate transition intensities. Rudolph showed that the approach allowed the inclusion of censored as well as time dependent factors at risk. In his second part of the paper, it was shown how the evaluated intensities can be utilized in a multiple state model to calculate premiums. (Gustafsson, 2009) applied survival analysis to predict low retention in non-life insurance industry. He applied statistical models to estimate the survival probabilities on customer level in a competing risk framework, where low retention could be the consequences of different types of causes. (Richmond and Roehner, 2016) used the famous Gompertz law for prediction of human mortality beyond the age of 100 years and established that the law also works well for ages over 100 years. They carried out the transversal analysis for a sample of industrialized and developed countries. All these revolutionary researches inspired the authors to carry out the present study in Indian scenario.

The paper ahead is organized as follows: Section 2 deals with the brief discussion of the evaluated and applied techniques for predictive modelling. Section 3 provides the description of data and methodology. In section 4, the results are discussed followed by conclusions in section 5.

2. Survival, Actuarial and Machine Learning Techniques for Predictive Modelling

2.1. Cox proportional hazard (PH) model (Lee and Wang, 2003)

Cox PH model is a popular tool in survival analysis for modelling the relationship between survival time and covariates. This is a semi parametric approach and assumes that the form of hazard function for the random lifetime T be given by,

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \le T \le t + dt \mid T > t)}{dt} \tag{1}$$

which is the product of a baseline hazard function $\lambda_0(t)$ and a specific scaling factor depending upon covariates and it is of the form given below,

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp(\beta^t Z) \tag{2}$$

where, $z \in R^p$ denotes the observed vector of covariates and $\beta \in R^p$ denotes the unknown regression coefficient.

Assuming no tied survival times, the estimation procedure is as follows:

Suppose that the k ($< n$) survival times of n individuals are distinct and uncensored, and the survival times of remaining $n - k$ individuals are right-censored. Let $t_{(1)} < t_{(2)} < t_{(3)} \ldots < t_{(k)}$ be the k distinct ordered survival times with corresponding predictors $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Let the risk set at time $t_{(i)}$ be denoted by $\boldsymbol{R(t_{(i)})}$. Then, $\boldsymbol{R(t_{(i)})}$ consists of all persons whose survival times are at least up to $t_{(i)}$. Then, the conditional probability of failure at time $t_{(i)}$ is given by

$$\frac{\exp\left(\underset{\sim}{b}' X_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\underset{\sim}{b}' X_{(l)}\right)} \tag{3}$$

And hence the partial likelihood function is given by,

$$L(\underset{\sim}{b}) = \prod_{i=1}^{k} \frac{\exp\left(\underset{\sim}{b}' X_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\underset{\sim}{b}' X_{(l)}\right)} \tag{4}$$

Then, (5) will give the maximum partial likelihood estimates (MPLEs) $\hat{\underset{\sim}{b}}$ by employing Newton Raphson iterated procedure.

$$\frac{\partial\left(log\, L\left(\underset{\sim}{b}\right)\right)}{\partial \underset{\sim}{b}} = 0 \tag{5}$$

2.2. Gompertz mortality law

The probability density function (pdf) of Gompertz distribution with location parameter *a* and shape parameter *b* is given by,

$$f(x) = ae^{bx - \frac{a}{b}(e^{bx}-1)}, \; x[0, \infty) \tag{6}$$

with the distribution function *F(x)*and the hazard function μ:

$$F(x) = 1 - e^{-\frac{a}{b}(e^{bx}-1)} \tag{7}$$

$$\mu(x) = ae^{bx}; a, b > 0 \tag{8}$$

Then, the revised maximum likelihood estimators (Lenart, 2012) of parameters *a* and *b* are given by (9) and (10).

$$\hat{a} = \frac{\sum_{x} D_x}{\sum_{x} E_x e^{bx}} \tag{9}$$

Estimator of *b* can be obtained solving numerically the following equality,

$$\frac{\sum_{x} D_x x}{\sum_{x} D_x} = \frac{\sum_{x} E_x e^{\hat{b}x} x}{\sum_{x} E_x e^{\hat{b}x}} \tag{10}$$

They provided the estimators for both the discrete and continuous ages. For discrete ages, if the number of deaths and the number of person years exposed to the risk of dying should be available.

2.3. Bayesian networks using Naïve Bayes classifier (Rao and Rao, 2014)

The Bayesian networks are probabilistic in nature and can model the joint probability density function (pdf) over a finite number of random variables. Bayesian network is represented as directed acyclic graph and it is often abbreviated as DAG. Its nodes are the random variables and the directed arcs express the dependencies among the random variables. A conditional probability table (CPT) of a variable *X* consists of probability distributions over the different states of random variable *X* for all possible combinations of

the states of *X's* parents. The joint probability distribution over all the random variables under the network can also be calculated by taking the product of all the priors and conditional pdfs. These networks deal with the complexity in the model and offer great flexibility. Therefore, these are more comprehensive than the other conventional survival models. Bayesian networks can be used in situations where little or no data are available. These models are based on the Bayes' theorem which is mathematically stated as,

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} \tag{11}$$

where A and B are events and $P(B) \neq 0$.

For an instance, the simple Bayesian network in life insurance can be represented as follows:

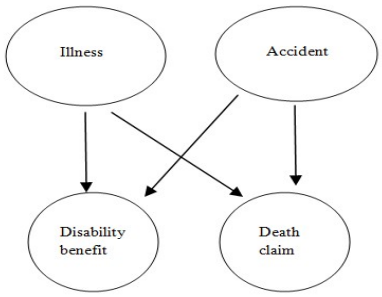


Figure 1: Simple Bayesian network

Figure 1 above shows the simple cause and effect form of Bayesian network which has two causes *viz.* illness or accident and two effects as claim for disability benefit or death. Suppose that the following probabilities are known:

P(accident), P(illness), P(disability benefit | accident), P(disability benefit | illness), P(death claim | accident) and P(death claim | illness), P(disability benefit | no accident), P(disability benefit | no illness), P(death claim | no accident) and P(death claim | no illness). Then, the conditional probability table (CPT) corresponding to any effect or node (disability benefit or death claim) given the cause or parent node information (illness or accident) can be constructed. One of such CPTs can be evaluated corresponding to disability benefit which is cited below in Table 1:

Table 1: Conditional probability table for disability benefit

Accident	Illness	Disability benefit
True	True	P_1
True	False	P_2
False	True	P_3
False	False	P_4

In Table 1 above the conditional probabilities for disability benefit have been denoted by P_1, P_2, P_3 and P_4. These probabilities are conditionally dependent upon their parent nodes, accident and illness. Other such CPTs are also evaluated for rest of the effect nodes and then any type of joint or conditional probabilities can be evaluated.

The best Bayesian network is generally fitted by learning from the data and maximizing the entropy scoring function (Cheng and Greiner, 2013). In the present study, we have

utilized the Bayesian networks (BN) Naïve Bayes classifier. Naïve Bayes classifiers have proved to be influential tools for solving classification problems in every field. Naïve Bayes classifier is the simplest probabilistic model based on the Bayes' theorem with strong independence assumptions between the characteristics (also known as predictors in case of regression).

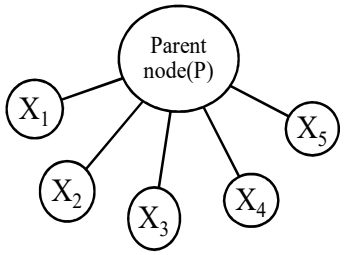


Figure 2: Structure of Naïve Bayes classifier

The Naïve Bayes classifier is the easy structure that has only one parent node of all the other next generation nodes. Graphically its structure is represented above in Figure 2.

In Figure 2 above, there is only one parent node denoted by *P* and all the other nodes X_1, X_2, X_3, X_4 and $X_5$are the next generation nodes (Cheng and Greiner, 2013).

For instance, consider the above Bayesian network with disability benefit and death claim as next generation nodes and accident as their only parent. Then, the structure of Naïve Bayes classifier is as follows (Figure 3):

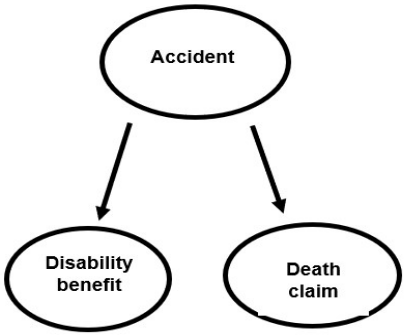
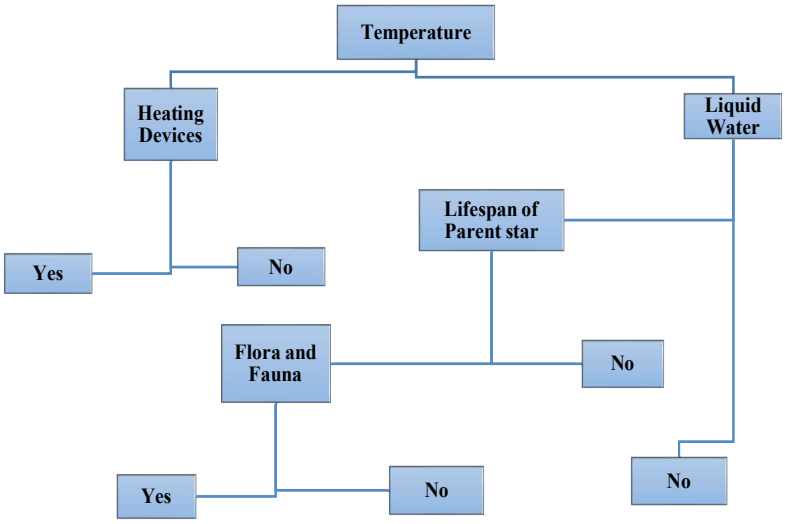


Figure 3: Structure of Naïve Bayes classifier

2.4. Random forest technique

The random forest technique is first introduced by Bell lab researcher (Ho, 1995). Random forest or random decision forest is a collaborative learning method for classification, regression and other jobs, that function by building a multitude of decision trees at training time and producing the class that is the type of the classes (classification) or mean forecasting (regression) of the individual tree. Random forest technique is more accurate for prediction, avoid over fitting and include bagging also. (Breiman, 2001) has compared empirically the results of the random forest technique with two different random features of selection from

the original inputs and the random linear combinations of inputs. When we make a prediction, the new observation gets strapped down each decision tree and allocated a predicted value. For example, suppose we want to know whether the planet is habitable or not. Then, the following type of decision tree helps in predicting. If the temperature on a planet is 150 Kelvin and there are no heating devices, then can it be predicted as habitable or not habitable? Figure 4 below shows the decision tree.



(Source: http://www.machinelearningtutorial.net/2017/01/23/randomforest-basics/)

Figure 4: Decision tree

Once each of the possible trees as shown above (in Figure 4) in the forest has reported its forecasted value, the predictions are matched up for the final prediction. In the case of **decision trees**, a simple majority vote (mode) controls the output, whereas in the case of a **regression trees**, the mean predictions of all individual trees form the final prediction.

3. Data and Employed Methodology

Data has been obtained by conducting a survey around Delhi NCR and from commercial sources of business. Data remains confidential and unpublished as it has been obtained from different surveys and is not available in the public domain. Data is spatial too, as it has been collected over the northern region of India. Data consists of 3363 policies which were procured in 2014 - 2015 with a balanced mix of areas of habitat; various products offered by insurance companies like Term, Savings, ULIP, Health and Pension; various ages of policy holders, income levels, sum assured (SA) *etc*. The categorical variables are assigned appropriate codes according to each level of such variables.

As there is a facility of payment of premium during grace period of one month thus, policies with the date of first unpaid premium (FUP) in the months of May and June of any year (time of assessment of policies) are assumed to be in force policies.

The methodology for implementation of the four models discussed above is given below:

3.1. Cox PH model

Let the predictor age be denoted by Z_{1t}, area by Z_2, income band by Z_{3t}, occupation by Z_4, plan type by Z_5, mode by Z_6, channel by Z_7, sum assured (SA) by Z_{SA}, gender by Z_{Gen} and policy term by Z_{8t}.

We first obtained the pair wise correlations among the predictors and then fitted the following Cox PH model:

$$\lambda(t \mid Z = z) = \lambda_0(t)\exp(\beta^t Z) \tag{12}$$

where, $\beta^t = (\beta_1, \beta_2, \beta_3, ..., \beta_n)$ and $Z = (Z_{1t}, Z_2, Z_{3t}, Z_4, Z_5, Z_6, Z_7, Z_{8t}, Z_{1t}Z_4Z_{8t}, Z_{1t}Z_4Z_7Z_{8t}, Z_{1t}Z_2Z_{3t}Z_4Z_5, Z_{1t}Z_2Z_{3t}Z_5Z_6, Z_{1t}Z_2Z_4Z_5Z_6, Z_{1t}Z_2Z_{3t}Z_5Z_{8t}, Z_{1t}Z_4Z_5Z_7Z_{8t})$.

Equation (12) is the form of hazard function under the Cox PH model which is expressed as the product of baseline hazard $\lambda_0(t)$ and the exponential function of the prognostic factors Z_i's which is denoted by $exp(\beta^t Z)$. Elements of vector Z include main predictors and various interactions of among these predictors like age:occupation:Policyterm is the third order interaction among age, occupation of policyholder and policy term and it is denoted by $Z_{1t}Z_4Z_{8t}$. Similarly, other higher order interaction terms have been used. We have also obtained the residuals using Deviance approach and predicted values of the type "expected" for the fitted model. Deviance residuals are symmetric unlike Martingale residuals. Using deviance residuals and predicted values, we have obtained the plot of Residuals *v/s* Predicted values. Such a plot is used diagnostic checking of the Cox PH model. That means we can assess whether the assumption of linearity is fulfilled or not.

3.2. Gompertz mortality law

Gompertz law of mortality for various predictors namely age, SA, area, gender and certain combinations of these predictors to estimate the probability of death which is lapsation in our case is defined below in four different forms.

Form 1: When mortality depends upon age only –

$$\mu(x) = ae^{bx}; a,b > 0 \text{ where, } x \text{ denotes age only.} \tag{13}$$

Form 2: When mortality depends upon SA only –

$$\mu(x) = ae^{bx}; a,b > 0 \text{ where, } x \text{ denotes SA only.} \tag{14}$$

Form 3: When mortality depends upon Area, Age and Gender –

$$\mu(x) = ae^{bx}; a,b > 0 \text{ where } x \text{ denotes interactive effect area*age*Gender} \tag{15}$$

Form 4: When mortality depends upon Area and SA:

$$\mu(x) = ae^{bx}; a,b > 0 \text{ where } x \text{ denotes the interactive effect area*SA} \tag{16}$$

The validity of these forms is checked by standardized residuals plot.

3.3. Bayesian networks using Naïve Bayes classifier

We first obtained the pairs and panel graphs before applying Naïve Bayes model which helps in ascertaining whether the predictors are linearly independent among themselves. This assumption has been earlier verified using the pair wise correlations among the predictors before fitting the Cox PH model. We then split the data in to two datasets namely training and testing in the ratio70:30. After splitting, we fitted the Naïve Bayes to the training data set with all the predictors. We further evaluated confusion matrix and error rate of misclassification for the model and compared them with the predicted results for testing data set.

3.4. Random forest technique

The entire dataset was divided into two parts in the ratio 70:30. Here, the first 70 *per cent* of the data are treated as training set and remaining 30 *per cent* treated as testing data set. The predicted variable is lapsation status of a policy which is coded as a factor variable with two levels *viz.,* 0, if policy is in force and 1, if it is lapsed. We then fitted the random forests model with various numbers of decision trees but with 20 trees, the model is found to be fitted well. We also obtained the plots of error rate for the model, proportion of predicted values for testing data and training data both. We also evaluated confusion matrix for the model to assess the goodness of fit of the model for predictive modelling.

4. Results and Discussions

4.1. Cox PH model

Pair wise correlations have been first calculated and Table 2 below shows the results:

Table 2: Pair wise Correlations

	Area	Plan type	Age	Mode	Channel	SA	Policy term	Gender	Occu.	Income Band
Area	1									
Plan type	0	1								
Age	–0.01	–0.02	1							
Mode	–0.02	–0.09	–0.01	1						
Channel	0	–0.02	0	0	1					
SA	0.07	–0.02	0.02	–0.08	–0.03	1				
Policy Term	–0.01	0.13	0.03	–0.01	–0.02	-0.01	1			
Gender	0.04	–0.03	0.03	–0.02	0.08	0.04	0	1		
Occupation (Occu.)	0.01	0.01	0.01	0	–0.01	0	0.02	0.02	1	
Income Band	0	–0.01	0	0.01	0.02	0.01	0.00	0	–0.01	1

The Table 2 above shows that all the pair wise correlations are very low, ensuring the linear independence among all the predictors and now we can proceed to fit the Cox PH model. Table 3 below presents the coefficients, exponential and *SE* of coefficients and *P* values:

Table 3:Results of Cox PH model

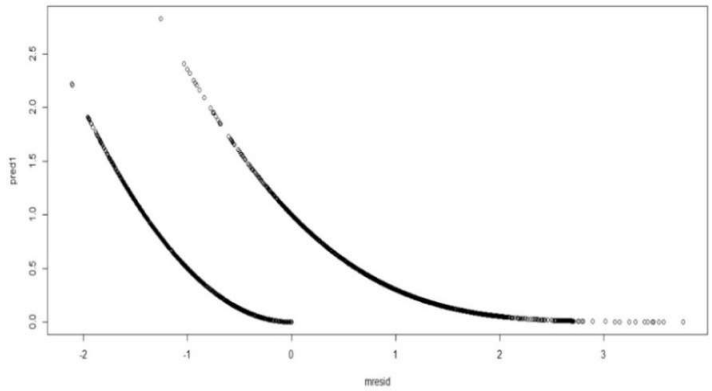
Effect	Notations	Coefficient	*SE* (Coeff)	*p*-value
Age	Z_{1t}	-2.828	64.58	0.9651
Area	Z_2	-10.860	3443.00	0.9975
Income_band	Z_{3t}	93.150	923.30	0.9196
Occupation	Z_4	-202.900	1034.00	0.8445
Plan_type	Z_5	56.790	3427.00	0.9868
Mode	Z_6	-56.170	1523.00	0.9706
Channel	Z_7	-320.800	3440.00	0.9257
Policy_term	Z_{8t}	-19.700	136.60	0.8853
age:occupation:Policy_term	$Z_{1t}Z_4Z_{8t}$	-0.305	0.168	*0.0702
age:occupation:channel:policy_term	$Z_{1t}Z_4Z_7Z_{8t}$	0.362	0.158	**0.0218
age:area:income_band:occupation:plan_type	$Z_{1t}Z_2Z_{3t}Z_4Z_5$	0.864	0.353	**0.0143
age:area:income_band:plan_type:mode	$Z_{1t}Z_2Z_{3t}Z_5Z_6$	1.326	0.472	***0.0050
age:area:occupation:plan_type:mode	$Z_{1t}Z_2Z_4Z_5Z_6$	1.866	0.676	***0.0058
age:area:income_band:plan_type:policy_term	$Z_{1t}Z_2Z_{3t}Z_5Z_{8t}$	0.113	0.036	***0.0017
age:occupation:plan_type:channel:policy_term	$Z_{1t}Z_4Z_5Z_7Z_{8t}$	-0.201	0.065	***0.0019

*denotes significance at 10*per cent* level of significance (los), **denotes significance at 5*per cent* level of significance, ***denotes significance at 1*per cent* level of significance (los).

The Table 3 above evidently exhibits that the main factors like age, area, income band *etc.* are all insignificant with probability values (*p*-values) more than 0.8 and all the interaction effects are significant at 10, 5 and 1 *per cent* level of significance. We observe that age has the coefficient – 2.828 and *p*-value is 0.9651 (> 0.05) which means that age is insignificant at 5 *per cent* level of significance for classifying the policies. Similarly, we can observe that the other main factors are also insignificant. On the other hand, the interaction effect of, say, age; occupation; and policy term is significant at 10 *per cent* level of significance with *p*-value 0.0702. The other interaction effects are also significant at different level of significance, like the interaction between age; occupation; channel; and policy term is significant at 5 *per cent* level of significance with *p*-value 0.0218 *etc*. This implies that the factors individually have no influence on the status of policy but taking together three or more factors is effective and helps in classification. In fact, as the order of interaction increases we observe that *p*-value is kept on decreasing and the effect is found to be significant.

The concordance in survival analysis is defined as measure of extent of agreement between risk score and time until failure. It was first popularized by (Harrell, Lee and Mark, 1996) in survival analysis and it is a statistic for Cox PH model. Now days, it is most widely used as a measure of goodness of fit in survival models. Not only in survival models but it has utility in logistic and ordinary linear regressions as well, (Therneau and Atkinson, 2020). The model with value of concordance more than 80 *per cent* is considered as a good fit but in our case the concordance value is evaluated as 62.8 *per cent* implying that the model is not a good fit. Other tests like Wald test, Likelihood ratio test and Score rank test with *p*-values 0.0001 also yield the same result. R^2 is obtained as 0.175 which seems to be inadequate. Figure 5 shows that the plot of deviance residuals also supports the same argument against the Cox PH model.

It can be apparently observed in Figure 5that residuals are following pattern which is sloping downwards from left to right. This curvilinear pattern in residuals leads us to conclude that Cox PH model is not a good fit in insurance phenomenon.



Please note that both the curves show the plot of deviance residuals against predicted values of response variable Y. Some coordinates lie on the first curve and others lie on the second curve.

Figure 5: Deviance residuals under Cox PH model

4.2. Gompertz Curve

We employed four different forms of Gompertz law of mortality for which results are tabulated below in Table 4. The *AIC* values for form 1 (with age only) and form 3 (with SA only) are shown as they are generated by the system. But the *AIC* values for form 2 (with age, area and gender) and form 4 (with area and SA) are obtained using the weighted average of the *AIC* values under the different combinations of factors that are affecting the status of policies. These various possible combinations under each form of Gompertz law are also displayed in column 2 of Table 4. For instance, to compute the *AIC* value for form 2, we calculated the weighted average four *AIC* values obtained by fitting the Gompertz law to the four combinations given below:

Male lives in rural area with all ages; male lives in urban area with all ages; female lives in rural area with all ages; and female lives in urban area with all ages.

We can also calculate the probability of lapsation for these models as weighted average of the probabilities obtained under each combination using respective estimates of *a* and *b*.

From Table 4, we observe that all the *AIC* values are very low. Form 1 and form 3 have the lowest *AIC* values –12.1 and –12.17. These forms do not involve any combination like forms 2 and 4. It is also observed that the estimate of parameter *b* is too small for all the forms.

Figure 6 shows the standardized residuals (*y*-axis) plotted against the predicted values of hazard (*x*-axis).The plot is observed to be random without any fixed pattern or cycle. This is the plot obtained for the applied form 1 of Gompertz law of mortality. The other three plots also have the similar random plots. These aforesaid random plots of standardized residuals validate the Gompertz law of mortality as the good fit. We may also conclude that each form of Gompertz curve fits well to the data as compared to Cox PH model for which the similar plot is non-random supporting the fact. The same conclusion has been drawn by (Ravi *et al*., 2020). They have also recognized the robustness of the actuarial laws for modelling the survival time in insurance phenomenon.

Table 4: Results under Gompertz curve

Model Form	Combinations	Value of *a*	Value of *b*	Weights	*AIC*
Form 1 - With age Only	None	0.5594	0.0003	1	–12.1
Form 2 - With age, Area and Gender	Male lives in rural area with all ages	0.547	0.0002	0.2037	–10.5724
	Male lives in urban area with all ages	0.5673	0.0003	0.7158	
	Female lives in rural area with all ages	0.3737	0.028	0.0131	
	Female lives in urban area with all ages	0.4695	0.0087	0.0674	
Form 3 - With SA only	None	0.5744	0.0001	1	–12.17
Form 4 - With area and SA	Rural area with all SA	0.5478	0.0005	0.2168	–11.127
	Urban area with all SA	0.5747	0.0008	0.7832	

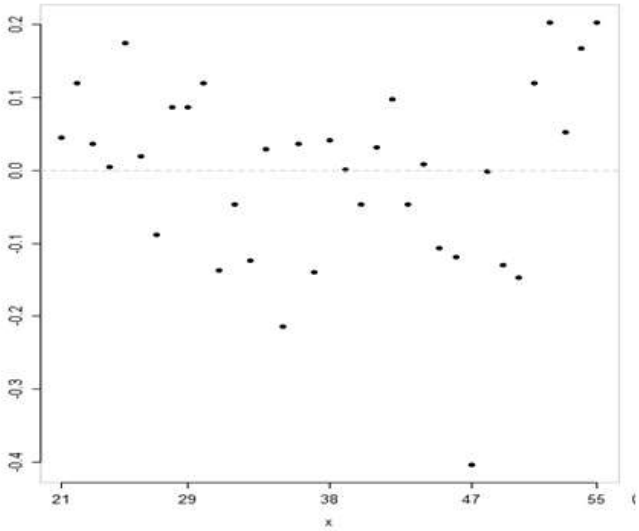


Figure 6: Residual plot under Gompertz law

Although the Gompertz law of mortality can be efficiently used for predicting lapsation probability but it is not suitable for classification purpose as it may require threshold value which in turn requires the complex algorithm to be employed. It predicts the probabilities of lapsation which may be categorized from low to high. To use this law for classification purpose, we are further required to find out a threshold value, such that if the predicted probabilities fall below it, then, the policies may be classified as lapsed and if the probabilities fall above it, then, the policies may be classified as in force. This makes the classification job more tedious. One more demerit of this law is that all the factors at role cannot be keyed in together to assess their effect on mortality (lapsation in our case). We must go on using various forms of the model one by one to assess their impact on the event of interest. So, we move ahead to apply other machine learning techniques of classification for predictive modelling.

4.3. Naïve Bayes classifier

We use the Kernel approach for fitting the Naïve Bayes classifier which is considered as more appropriate.

The a priori probabilities are evaluated as: 0.438 for in force policies and 0.562 for lapsed policies. We predicted few cases where certain combinations of values are given as input for factors at work like gender, SA, age, plan type *etc*. and model classified the policies as lapsed or in force. Then on comparing the predicted and observed values, the results are found to be consistent.

Table 5: Confusion matrix for training data

Status of policy (Lapsed: 1 and Inforce: 0) ↓ →	0	1
0	1130	2
1	0	1447

Table 5 above shows that out of 2579 policies in training data set, 1130 policies are correctly classified as in force and 1447 policies are correctly classified as lapsed whereas 2 policies are classified as lapsed while they were actually in force. The error rate for the model is calculated as 0.000775.

Further the confusion matrix for testing data set which contains 1084 policiesis exhibited below in Table 6.

Table 6: Confusion matrix for testing data

Status of policy (Lapsed: 1 and Inforce:0) ↓ →	0	1
0	462	3
1	0	619

Table 6 above shows out of 1084 policies; 462 policies are correctly classified as in force and 619 policies are correctly classified as lapsed which are consistent with their observed values and the remaining 3 policies are misclassified. Error rate of the model for testing data for the model is 0.002767.

In both the cases it is observed that misclassification probability is negligibly small. Therefore, the Naïve Bayesian model is a better fit as compared to the other two models discussed above.

4.4. Random Forest technique

Figure 7 below shows the error rates of misclassification for the model (y - axis) plotted against the chosen number of decision trees (x - axis). We observe that the decline in error rates become stagnant between 15 and 20 numbers of decision trees. If we choose decision tree beyond 20 then there will be no significant reduction in error rate but the number of parameters to be estimated will redundantly increase.

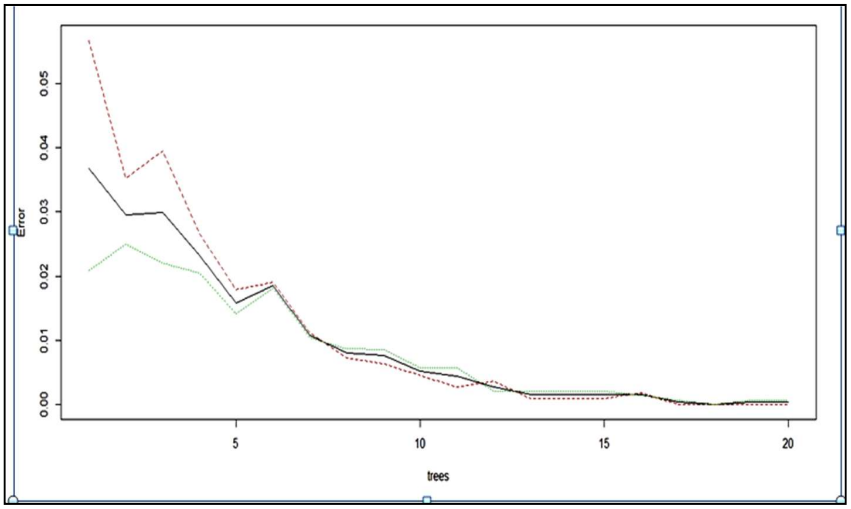


Figure 7: Error rate for fitted random forest

Following results are obtained by fitting the random forest with 20 decision trees:

(i) The Out of Bag (*OOB*) estimate of error rate is 0.04 *per cent.*

(ii) The confusion matrix (Table 7 below) for the fitted random forest on training data set.

Table 7: Confusion matrix on training data

Status of policy (Lapsed: 1 and In force: 0) → ↓	0	1
0	1111	1
1	0	1431

It is apparent that out of total 2543 policies in training data set, random forest correctly classified the 1111 policies as in force when they are actually in force and 1431 policies as lapsed when they are actually lapsed. The model misclassified only in one case where the policy is predicted as lapsed but it is actually an in-force policy.

(iii) The confusion matrix (Table 8 below) for the fitted random forest on testing data set is given below and the *OOB* error rate estimated as 0.067 *per cent*.

Table 8: Confusion matrix on testing data

Status of policy (Lapsed: 1 and In force: 0) ↓ →	0	1
0	480	2
1	0	638

The above confusion matrix in Table 8 gives correct classification for total 1118 policies out of 1120 policies in testing data set and misclassified only 2 policies. All these above results show that the model is a best fit to the data.

(iv) The proportions of the policies with the predicted status as lapsed (coded as 1) or in force (coded as 0), for both the training and testing data sets are shown below in Figure 8.

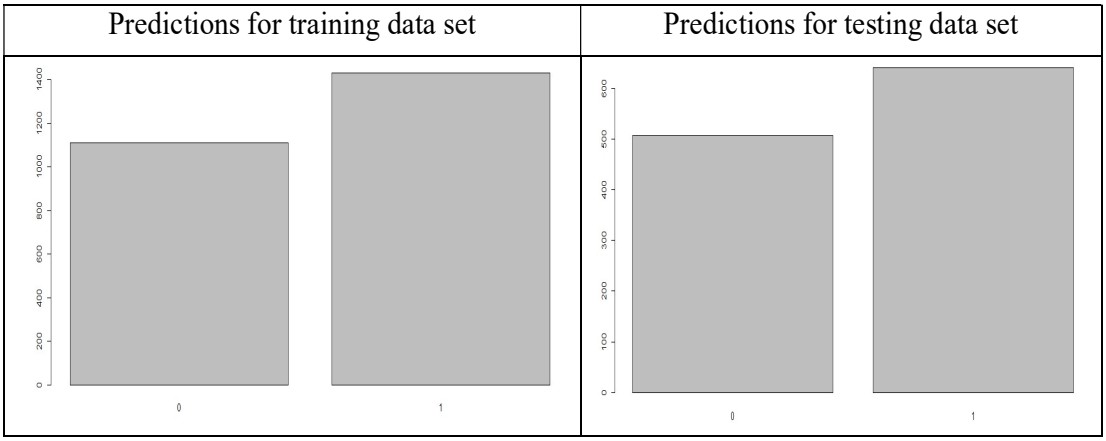


Figure 8: Prediction proportions for both training and testing data

The proportions for both the data sets are almost equal for lapsation as well as in force policies. For both training and testing data set the model predicted same proportion around 54 *per cent* for lapsed policies. The proportion of in force policies in both the data sets is approximately obtained as 46 *per cent*. Both data sets having almost same proportions of predictions for in force and lapsed policies specify that the model works well.

5. Conclusions

A comparison of all the four models for predictive modelling and classifying the policies as lapsed or in force is important before concluding and therefore, it is shown below in Table 9.

Table 9: Comparison of employed four models

Points for comparison	Cox PH model	Gompertz law	Naïve bayes classifier	Random forests
Does model fit well to the survival data of life insurance policies	No	Yes	Yes	Yes
Basis of assessment of a model as a good fit	Standardized Residuals plot and Concordance	Standardized residuals plot and AIC values	Confusion matrix and error rate of misclassification	Confusion matrix and error rate of misclassification
Is the model appropriate to classify the life insurance policies	No Not appropriate as it does not fit the data well	No It is tedious to use the law for classification which requires complex algorithm	Yes It is easy to use and interpret	Yes It is easy to use and interpret
Does model involve complex algorithm to classify	Yes	Yes	No	No

The conventional survival model (Cox PH model) and the actuarial law (Gompertz law) are not found to be suitable for the classification of policies as lapsed or in force and the reasons have been tabulated above in Table 9. Thus, we require some better predictive modelling techniques which can classify the policies with certain set of values of predictors/factors at work, as lapsed or in force. These techniques are also capable of taking into consideration all the factors affecting lapsation together. We see that Naïve Bayesian technique and Random forest technique both performed very well. Both the techniques have the almost same error rate of prediction. Under both the models, confusion matrix is depicting the same proportions of correctly classified policies. We conclude that both the techniques performed equivalently well on insurance data. Classification is important in predictive modelling to assess whether the policy holder with certain features like income band, SA opted, policy term opted, age etc. will be lapsing the policy or will continue the policy with certain probabilities attached to the outcome. It is found that nearly 54 per cent policies are classified as lapsed in the class of policyholders who are businessmen up to the age 45 and with income up to 5 lakhs, insured with low SA up to 5 lakhs and policy term between 25 to 30 years and they purchased savings/health/ULIP insurance products. This classification helps insurer to identify the segments in the society with the different combinations of various features where more prospects can be canvassed which will be retained for long time or up to the maturity. Insurers may also identify the segments where the prospects or policyholders are not being canvassed or retained for long time to analyze for probable reasons responsible for problem of low retention. The issues may be related to after sales service, low awareness of insurance in the segment or products which do not appropriately serve the needs in the segment. Some issues are also beyond the control of an insurer. But some of the issues like mentioned above may be improved by working upon them.

References

Adebiyi, S. O., Oyatoye, E.O. and Amole B. B. (2016). Improved customer churn and retention decision management using operations research approach. *Emerging Markets Journal*, **6**(**2**). DOI 10.5195/emaj.2016.101, http://emaj.pitt.edu

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E. and O'Connor, P. J. (2014). Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data.http://arxiv.org/abs/1404.2189v1.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32. Kluwer Academic Publishers. Manufactured in The Netherlands.

Cheng, J. and Greiner, R. (2013). *Comparing Bayesian Network Classifier*. Department of Computing Science University of Alberta Edmonton, Email: {jcheng, greiner} @cs.ualberta.ca.

Gustafsson, E. (2009). Customer duration in non-life insurance industry. *Mathematical Statistics Stockholm University*, Examensarbete, **3**.http://www.math.su.se/matstat.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

Huigevoort, C. (2015). *Customer Churn Prediction for an Insurance Company*. TUE School of Industrial Engineering. Series Master Theses, Operations Management and Logistics.

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z. and Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics*, **5**, 2. doi:10.3390/informatics5010002.

Lariviere, B. and Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, **29**, 472-484. 10.1016/j.eswa.2005.04.043.

Lee, E. T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*. Third Edition. Wiley.

Lenart, A. (2012). *The Gompertz Distribution and Maximum Likelihood Estimation of its Parameters - a Revision*". Max Planck Institute for Demographic Research, Working Paper 2012-008.

Onisko, A., Druzdzel, M. J. and Wasyluk, H. (2001). Learning Bayesian network parameters from small data sets: application of noisy OR gates. *International Journal of Approximate* Reasoning, **27**,165-182.

Rao, M. B. and Rao, C. R. (2014). *Handbook of Statistics*. Elsevier (ISBN: 9780444634313).

Ravi, V., Saini, R., Varshney, M. K. and Grover, G. (2020). Modelling of survival time of life insurance policies in India: A Comparative study. *International Journal of System Assurance Engineering and Management.* DOI: 10.1007/s13198-020-01026-2.

Richmond, P. and Roehner, B. M. (2016). Predictive implications of Gompertz law. *Physics A*: *Statistical Mechanics and its Applications*. http://arxiv.org/abs/1509.07271v1.

Rudolph, C. (2002). *Application of Survival Analysis Methods to Long Term Care Insurance.* Sonderforschungsbereich386, Paper 268. http://epub.ub.uni-muenchen.de/.

Therneau, T. and Atkinson, E. (2020). *Concordance*. www.CRAN.r-project.org.

Tirenni, G., Kaiserand C. and Herrmann, A. (2007). Applying decision trees for value-based customer relations management: Predicting airline customers "future values". *Database Marketing and Customer Strategy* Management, **14,** 130–142.

Zhang, R., Li, W., Mo and, T. and Tan, W. (2017). Deep and shallow model for insurance churn prediction service. *IEEE Computer Society*, 2474-2473/17.

DOI 10.1109/SCC.2017.51

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 49-58

Modified Exponential Product Type Estimators for Estimating Population Mean Using Auxiliary Information

Sajad Hussain[1], Manish Sharma[1] and Hukum Chandra[2]

[1]*Division of Statistics and Computer Science, FBSc, SKUAST-Jammu, Chatha-180009*
[2]*Formerly at ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012*

Received: 30 May 2021; Revised: 09 June 2021; Accepted: 12 June 2021

Abstract

In this paper two exponential product type estimators of population mean of the study variable have been proposed in case of simple random sampling without replacement (SRSWOR) sampling scheme. The large sample properties of the proposed estimators have been evaluated to the first order of approximation. The estimators proposed are found more efficient than the mean per unit estimator, product type estimator of Robson (1957), exponential product type estimator of Bahl and Tuteja (1991) and Onyeka (2013). The theoretical findings of the study have been evaluated and verified empirically using data of two real populations.

Key words: Exponential product estimator; Auxiliary information; Optimum value; Efficiency.

1. Introduction

Sampling methods are used to get an overview of the universe by studying a subset. Even if the subset is chosen sufficiently large, it may not fully represent the whole universe meaning thereby that the estimates obtained from this subset may be far away from the true estimates of the universe. To get these sample estimates close and close to the actual parameters of the universe, one may define a second variable called auxiliary variable having high correlation with the variable under study and use some known parameters such as mean, coefficient of variation, median, skewness, kurtosis, *etc.* of this auxiliary variable for the said purpose. This auxiliary variable may have a positive or negative correlation with the study variable. In case of positive correlation, the estimators of Cochran (1940), Kadilar and Cingi (2004), Miasra *et al.* (2017), Hussain *et al.* (2021) *etc.* known as ratio estimators are used while as in case of negative correlation, the estimators of Robson (1957), Murthy (1964), Shukla (1976), Vishwakarma *et al.* (2016) *etc*. known as product type estimators are used. The pioneer work of Bahl and Tuteja (1991) proposed exponential ratio and product type estimators. The significance of exponential estimators lies in estimating the population mean precisely even at low degree of correlation. However, the precision of an estimate may be increased by modifying the conventional/classical estimators. Onyeka (2013) proposed a class of modified exponential product type estimators of population parameter by extending the work of Singh *et al.* (2009). Later Zaman and Kadilar (2019) and Zaman (2020) also contributed to this effort. This paper extends the work of Hussain *et al.* (2021) and proposes modified exponential product type estimators.

Corresponding Author: Manish Sharma
E-mail: manshstat@gmail.com
Manish Sharma had several close interactions with Hukum Chandra while preparing this article for presentation during the conference. Unfortunately, Hukum passed away on 26 April 2021. Our deepest condolences to the bereaved family. In view of his significant contributions, Hukum has been included as a co-author. This paper is a tribute to Hukum.

Consider a population of N units. A sample of size n is drawn from this population by simple random sampling without replacement (srswor). Let Y_i and X_i denote the study and the auxiliary variables respectively, corresponding to the i^{th} (i =1, 2, …N) unit of population and y_i and x_i denote the corresponding study and auxiliary variables respectively, for the i^{th} (i =1, 2, … n) unit in sample. The formulae and notations used in the paper (See Haq and Shabir, 2014 and John and Inyang, 2015) are as follows:

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ are the population means, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means, $C_y = \frac{S_y}{\bar{Y}}$ and $C_x = \frac{S_x}{\bar{X}}$ are the population coefficient of variation, $S_{yy} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ and $S_{xx} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ are the population mean squares, $s_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ are the sample mean squares of study and auxiliary variable respectively. $\rho = \frac{S_{xy}}{S_{xx}S_{yy}}$ is the correlation coefficient between the auxiliary and study variable, where $S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})$.

Further, $\theta = \frac{a\bar{X}}{2(a\bar{X}+b)}$, $\gamma = \frac{1-f}{n}$, where $f = \frac{n}{N}$ is the sampling fraction.

2. Existing Estimators of Population Mean

The usual sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ provides an unbiased estimator of the population mean. The Bias and MSE of $\bar{y}$ are as

$$\text{Bias } (\bar{y}) = 0.$$
$$\text{V } (\bar{y}) = \gamma C_y^2 \bar{Y}^2. \quad (1)$$

When study and auxiliary variables are negatively correlated, Robson (1957) proposed product type estimator as

$$\bar{y}_{RB} = \bar{y}\,\frac{\bar{x}}{\bar{X}}.$$

The estimator $\bar{y}_{RB}$ is biased and is more efficient than the estimator $\bar{y}$, if $\rho < -\frac{C_x}{2C_y}$. The Bias and MSE of the estimator $\bar{y}_{RB}$ are as

$$\text{Bias } (\bar{y}_{RB}) = \gamma \bar{Y} C_{yx}.$$
$$\text{MSE } (\bar{y}_{RB}) = \gamma \bar{Y}^2 (C_y^2 + C_x^2 + 2C_{yx}). \quad (2)$$

Bahl and Tuteja (1991) were the pioneer to propose exponential product type estimator as a precise estimator of population mean as

$$\bar{y}_{BT} = \bar{y} \exp\left(\frac{\bar{x} - \bar{X}}{\bar{X} + \bar{x}}\right).$$

The Bias and MSE of the estimator $\bar{y}_{BT}$ are as

$$\text{Bias } (\bar{y}_{BT}) = \gamma \bar{Y}\left(\frac{1}{2}C_{yx} - \frac{1}{8}C_x^2\right).$$
$$\text{MSE } (\bar{y}_{BT}) = \gamma \bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} + C_{yx}\right). \quad (3)$$

Onyeka (2013) extended the work which was carried out by Singh *et al.* (2009) and proposed a class of product type estimators as

$$\bar{y}_{NK} = \bar{y} \exp\left[\frac{(a\bar{x}+b)-(a\bar{X}+b)}{(a\bar{x}+b)+(a\bar{X}+b)}\right].$$

with the Bias and MSE as

$$\text{Bias}\,(\bar{y}_{NK}) = \gamma\bar{Y}\left(\frac{1}{2}\theta C_{yx} - \frac{1}{8}\theta^2 C_x^2\right).$$

$$\text{MSE}\,(\bar{y}_{NK}) = \gamma\bar{Y}^2\left(C_y^2 + \frac{1}{4}\theta^2 C_x^2 + \theta C_{yx}\right). \quad (4)$$

3. Proposed Exponential product type Estimators of population mean

The modified exponential product type estimators of population mean proposed are as

$$\bar{y}_{\alpha_1} = \bar{y} \exp\left(\frac{\bar{x}-\bar{X}}{\alpha_1\bar{x}}\right).$$

$$\bar{y}_{\alpha_2} = \bar{y} \exp\left(\frac{\bar{x}-\bar{X}}{\alpha_2\bar{X}}\right).$$

where $\alpha_1$and α_2 are the constants to be determined such that the proposed estimators $\bar{y}_{\alpha_1}$and $\bar{y}_{\alpha_2}$ estimate population mean precisely. The Bias and MSE of the proposed estimators $\bar{y}_{\alpha_1}$and $\bar{y}_{\alpha_2}$ to the first order of approximation are as

$$\text{Bias}\,(\bar{y}_{\alpha_1}) = \bar{Y}\,\frac{1}{\alpha_1}\left(\frac{1}{2\alpha_1}C_x^2 - C_x^2 + C_{yx}\right).$$

$$\text{Bias}\,(\bar{y}_{\alpha_2}) = \bar{Y}\,\frac{1}{\alpha_2}\left(\frac{1}{2\alpha_2}C_x^2 + C_{yx}\right).$$

$$\text{MSE}\,(\bar{y}_{\alpha_1}) = \gamma\bar{Y}^2\left(C_y^2 + \frac{1}{{\alpha_1}^2}C_x^2 + \frac{2}{\alpha_1}C_{yx}\right).$$

$$\text{MSE}\,(\bar{y}_{\alpha_2}) = \gamma\bar{Y}^2\left(C_y^2 + \frac{1}{{\alpha_2}^2}C_x^2 + \frac{2}{\alpha_2}C_{yx}\right).$$

In order to find out the expressions for Bias and MSE. Let us consider

$$e_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}} \text{ and } e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}.$$

We have,

$$\text{E}(e_0) = \text{E}(e_1) = 0,\ \text{E}\,(e_0^2) = \gamma C_y^2,\ \text{E}\,(e_1^2) = \gamma C_x^2,\ \text{E}\,(e_0e_1) = \gamma C_{yx}.$$

Writing the estimator $\bar{y}_{\alpha_1}$in terms of e_i's $(i = 1, 2)$, therefore

$$\bar{y}_{\alpha_1} = \bar{Y}(1+e_0)\exp\left[\frac{(1+e_1)\bar{X}-\bar{X}}{\alpha_1(1+e_1)\bar{X}}\right]$$

$$\Rightarrow\ \bar{y}_{\alpha_1} = \bar{Y}(1+e_0)\exp\left[\left(\frac{e_1}{\alpha_1}\right)(1+e_1)^{-1}\right]$$

$$\Rightarrow\ \bar{y}_{\alpha_1} = \bar{Y}(1+e_0)\exp\left[\left(\frac{e_1}{\alpha_1}\right)(1-e_1+e_1^2-e_1^3+\cdots)\right]. \quad (5)$$

Solving equation (5) and retaining the terms upto second degree only, we have the resulting expression as

$$\bar{y}_{\alpha_1} = \bar{Y}\left(1 + e_0 + \frac{e_1}{\alpha_1} - \frac{e_1^2}{\alpha_1} + \frac{e_1^2}{2\alpha_1{}^2} + \frac{e_0 e_1}{\alpha_1}\right)$$

$$\Rightarrow \bar{y}_{\alpha_1} - \bar{Y} = \left(e_0 + \frac{e_1}{\alpha_1} - \frac{e_1^2}{\alpha_1} + \frac{e_1^2}{2\alpha_1{}^2} + \frac{e_0 e_1}{\alpha_1}\right). \quad (6)$$

Taking expectation on both sides of equation of equation (6) for obtaining Bias $(\bar{y}_{\alpha_1})$ as

$$\mathrm{E}\,(\bar{y}_{\alpha_1} - \bar{Y}) = \mathrm{E}\left[e_0 + \frac{e_1}{\alpha_1} - \frac{e_1^2}{\alpha_1} + \frac{e_1^2}{2\alpha_1{}^2} + \frac{e_0 e_1}{\alpha_1}\right]$$

$$\Rightarrow \mathrm{E}\,(\bar{y}_{\alpha_1} - \bar{Y}) = \frac{1}{2\alpha_1{}^2}\mathrm{E}\,(e_1^2) - \frac{1}{\alpha_1}\mathrm{E}\,(e_1^2) + \frac{1}{\alpha_1}\mathrm{E}\,(e_0 e_1)$$

$$\Rightarrow \text{Bias}\,(\bar{y}_{\alpha_1}) = \bar{Y}\frac{1}{\alpha_1}\left(\frac{1}{2\alpha_1}C_x^2 - C_x^2 + C_{yx}\right). \quad (7)$$

Squaring equation (6) on both sides and then taking expectation for obtaining MSE $(\bar{y}_{\alpha_1})$ as

$$\mathrm{E}\,(\bar{y}_{\alpha_1} - \bar{Y})^2 = \mathrm{E}\left[e_0 + \frac{e_1}{\alpha_1} - \frac{e_1^2}{\alpha_1} + \frac{e_1^2}{2\alpha_1{}^2} + \frac{e_0 e_1}{\alpha_1}\right]^2. \quad (8)$$

Solving equation (8) and retaining the terms up to second degree only, the expression for MSE $(\bar{y}_{\alpha_1})$ is obtained as

$$\text{MSE}\,(\bar{y}_{\alpha_1}) = \gamma\bar{Y}^2\left(C_y^2 + \frac{1}{\alpha_1{}^2}C_x^2 + \frac{2}{\alpha_1}C_{yx}\right). \quad (9)$$

Now writing the estimator $\bar{y}_{\alpha_2}$ in terms of e_i's $(i = 1, 2)$, therefore

$$\bar{y}_{\alpha_2} = \bar{Y}(1 + e_0)\exp\left[\frac{(1+e_1)\bar{X} - \bar{X}}{\alpha_2\bar{X}}\right]$$

$$\Rightarrow \bar{y}_{\alpha_2} = \bar{Y}(1 + e_0)\exp\left(\frac{e_1}{\alpha_2}\right)$$

$$\Rightarrow \bar{y}_{\alpha_2} = \bar{Y}(1 + e_0)\left(1 + \frac{e_1}{\alpha_2} + \frac{e_1^2}{2\alpha_2^2} + \frac{e_1^3}{6\alpha_2^3} + \cdots\right). \quad (10)$$

Solving (10) and retaining the terms up to the second degree only, the resulting expression is as

$$\bar{y}_{\alpha_2} = \bar{Y}\left(1 + e_0 + \frac{e_1}{\alpha_2} + \frac{e_1^2}{2\alpha_2^2} + \frac{e_0 e_1}{\alpha_2}\right)$$

$$\Rightarrow \bar{y}_{\alpha_2} - \bar{Y} = \bar{Y}\left(e_0 + \frac{e_1}{\alpha_2} + \frac{e_1^2}{2\alpha_2^2} + \frac{e_0 e_1}{\alpha_2}\right). \quad (11)$$

Using the same procedure for finding Bias $(\bar{y}_{\alpha_2})$ and MSE $(\bar{y}_{\alpha_2})$ from equation (11) as applied to equation for finding Bias $(\bar{y}_{\alpha_1})$ and MSE $(\bar{y}_{\alpha_1})$, we have

$$\text{Bias}\,(\bar{y}_{\alpha_2}) = \bar{Y}\frac{1}{\alpha_2}\left(\frac{1}{2\alpha_2}C_x^2 + C_{yx}\right). \quad (12)$$

$$\text{MSE}\,(\bar{y}_{\alpha_2}) = \gamma\bar{Y}^2\left(C_y^2 + \frac{1}{\alpha_2{}^2}C_x^2 + \frac{2}{\alpha_2}C_{yx}\right). \quad (13)$$

The expressions obtained (7), (9), (12) and (13) are the required expressions.

Optimum value of α_1 and α_2

Now differentiating the equations (9) and (13) partially with respect to α_1 and α_2 respectively and equating to zero, the optimal value of α_1 and α_2 is found to be $\frac{-C_x}{\rho C_y} = \eta$ (say). The value of η can be obtained quite accurately from some previous survey or from the experience of the researcher (See Reddy (1973, 1974), Singh and Vishwakarma (2008), Singh and Kumar (2008), Singh and Kapre (2010)). Substituting the value of η value in equations (7) and (12), the expressions for Bias ($\bar{y}_{\alpha_1}$) and Bias ($\bar{y}_{\alpha_2}$) at the optimal condition are obtained as

$$\text{Bias } (\bar{y}_{\alpha_1}) = \bar{Y}\left(C_{yx} - \frac{1}{2}\rho^2 C_y^2\right). \tag{14}$$

$$\text{Bias } (\bar{y}_{\alpha_2}) = -\frac{1}{2}\gamma\bar{Y}\rho^2 C_y^2. \tag{15}$$

Now substituting the optimal value of α_1 and α_2 in equations (9) and (13), the minimum value of MSE ($\bar{y}_{\alpha_1}$) and MSE ($\bar{y}_{\alpha_2}$) is obtained as

$$\text{MSE}_{\min}(\bar{y}_{\alpha_i}) = \gamma\bar{Y}^2 C_y^2 (1-\rho^2): \qquad i = 1,2 \tag{16}$$

Special cases: The proposed product type estimators $\bar{y}_{\alpha_i}$ (i=1, 2) can be used as an alternative to product type estimator of Robson (1957), exponential product type estimator of Bahl and Tuteja (1991) under the conditions as

(i) $\alpha_1 = \alpha_2 = 1$, the MSE of the proposed estimators $\bar{y}_{\alpha_i}$ (i=1, 2) is same as that of the MSE of product type of Robson (1957).

(ii) $\alpha_1 = \alpha_2 = 2$, the MSE of the proposed estimators $\bar{y}_{\alpha_i}$ (i=1, 2) is same as the MSE of the exponential product type estimator of Bahl and Tuteja (1991).

4. Theoretical Efficiency Comparison

The efficiency comparisons of the study are done using the MSE of the proposed estimators $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ and of the existing estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$ considered.

(i) Efficiency comparison of $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ when the values of α_1 and α_2 coincide with its optimal value

Solving the expressions (1), (2), (3), (4) and (16), the conditions obtained are as

$$\text{MSE}_{\min}(\bar{y}_{\alpha_i}) < \text{V}(\bar{y})$$

$$\Rightarrow \gamma\bar{Y}^2 C_y^2 (1-\rho^2) < \gamma C_y^2 \bar{Y}^2, \text{ if } \rho^2\bar{Y}^2 > 0. \tag{17}$$

$$\text{MSE}_{\min}(\bar{y}_{\alpha_i}) < \text{MSE}(\bar{y}_{RB})$$

$$\Rightarrow \gamma\bar{Y}^2 C_y^2 (1-\rho^2) < \gamma\bar{Y}^2(C_y^2 + C_x^2 + 2C_{yx}), \text{ if } (\rho C_y + C_x)^2 > 0. \tag{18}$$

$$\text{MSE}_{\min}(\bar{y}_{\alpha_i}) < \text{MSE}(\bar{y}_{BT}), \text{ if } (2\rho C_y + C_x)^2 > 0.$$

$$\Rightarrow \gamma\bar{Y}^2 C_y^2 (1-\rho^2) < \gamma\bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} + C_{yx}\right), \text{ if } (2\rho C_y + C_x)^2 > 0. \tag{19}$$

$$\text{MSE}_{\min}(\bar{y}_{\alpha_i}) < \text{MSE}(\bar{y}_{NK})$$

$$\Rightarrow \gamma\bar{Y}^2C_y^2\,(1-\rho^2) < \gamma\bar{Y}^2\left(C_y^2+\frac{1}{4}\theta^2C_x^2+\theta C_{yx}\right),\ if\ (2\rho C_y+\theta\ C_x)^2>0. \tag{20}$$

Therefore, under the conditions (17) to (20), the proposed estimators $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ will be more efficient than the product type estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$ considered in this study.

(ii) Efficiency comparison of $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ when the value of α_1 and α_2 does not coincide with its optimal value

When the equations (1), (2), (3), (4) and (9), (13) were solved, the following conditions were obtained

$$\text{MSE}(\bar{y}_{\alpha_1}) < \text{V}(\bar{y}),\ if\ \alpha_1 > \left(\frac{-C_x^2}{2C_{yx}}\right) \tag{21}$$

$$\text{MSE}(\bar{y}_{\alpha_2}) < \text{V}(\bar{y}),\ if\ \alpha_2 > \left(\frac{-C_x^2}{2C_{yx}}\right) \tag{22}$$

$$\text{MSE}(\bar{y}_{\alpha_1}) < \text{V}(\bar{y}_{RB}),\ if$$

$$min\left(1,\frac{-C_x^2}{2C_{yx}+C_x^2}\right) < \alpha_1 < max\left(1,\frac{-C_x^2}{2C_{yx}+C_x^2}\right),\ \frac{C_{yx}}{C_x^2} < -\frac{1}{2} \tag{23}$$

$$Or\ \alpha_1 > 1,\ -\frac{1}{2} \le \frac{C_{yx}}{C_x^2} < 0.$$

$$\text{MSE}(\bar{y}_{\alpha_2}) < \text{V}(\bar{y}_{RB}),\ if$$

$$min\left(1,\frac{-C_x^2}{2C_{yx}+C_x^2}\right) < \alpha_2 < max\left(1,\frac{-C_x^2}{2C_{yx}+C_x^2}\right),\ \frac{C_{yx}}{C_x^2} < -\frac{1}{2} \tag{24}$$

$$Or\ \alpha_2 > 1,\ -\frac{1}{2} \le \frac{C_{yx}}{C_x^2} < 0.$$

$$\text{MSE}(\bar{y}_{\alpha_1}) < \text{V}(\bar{y}_{BT}),\ if$$

$$min\left(2,\frac{-2C_x^2}{4C_{yx}+C_x^2}\right) < \alpha_1 < max\left(2,\frac{-2C_x^2}{4C_{yx}+C_x^2}\right),\ \frac{C_{yx}}{C_x^2} < -\frac{1}{4} \tag{25}$$

$$Or\ \alpha_1 > 2,\ -\frac{1}{4} \le \frac{C_{yx}}{C_x^2} < 0\ .$$

$$\text{MSE}(\bar{y}_{\alpha_2}) < \text{V}(\bar{y}_{BT}),\ if$$

$$min\left(2,\frac{-2C_x^2}{4C_{yx}+C_x^2}\right) < \alpha_2 < max\left(2,\frac{-2C_x^2}{4C_{yx}+C_x^2}\right),\ \frac{C_{yx}}{C_x^2} < -\frac{1}{4} \tag{26}$$

$$Or\ \alpha_2 > 2,\ -\frac{1}{4} \le \frac{C_{yx}}{C_x^2} < 0.$$

$$\text{MSE}(\bar{y}_{\alpha_1}) < \text{V}(\bar{y}_{NK}),\ if$$

$$min\left(\frac{\theta}{2},\frac{-2C_x^2}{2C_{yx}+\theta C_x^2}\right) < \alpha_1 < max\left(\frac{\theta}{2},\frac{-2C_x^2}{2C_{yx}+\theta C_x^2}\right),\ \frac{C_{yx}}{C_x^2} < -\frac{\theta}{4}. \tag{27}$$

$$Or\ \alpha_1 > \frac{\theta}{2},\ -\frac{\theta}{4} \le \frac{C_{yx}}{C_x^2} < 0.$$

$$\text{MSE}(\bar{y}_{\alpha_2}) < \text{V}(\bar{y}_{NK}),\ if$$

$$min\left(\frac{\theta}{2}, \frac{-2C_x^2}{2C_{yx}+\theta C_x^2}\right) < \alpha_2 < max\left(\frac{\theta}{2}, \frac{-2C_x^2}{2C_{yx}+\theta C_x^2}\right), \frac{C_{yx}}{C_x^2} < -\frac{\theta}{4}. \quad (28)$$

$$or\ \alpha_2 > \frac{\theta}{2}, -\frac{\theta}{4} \leq \frac{C_{yx}}{C_x^2} < 0.$$

Under the conditions (21) to (28) the estimators $\bar{y}_{\alpha_1}$ ($\bar{y}_{\alpha_2}$) are more efficient than the estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$.

5. Numerical Study

The numerical study of the present work is done using data of two populations P1 and P2 where the study and auxiliary variable are negatively correlated. The population P1 is taken from Onyeka (2013) where the study variable (*Y*) is the Percentage of hives affected by disease and the auxiliary variable (*X*) is the date of flowering of a particular summer species (number of days from January 1). The population P2 has been taken from Gujarati (2004) where the study variable (*Y*) is the Average miles per gallon and the auxiliary variable (*X*) is the top speed, miles per hour. The data regarding the populations taken is given in Table 1. The performance of the product type estimators $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ has been compared with the sample mean estimator $\bar{y}$ and the product type estimators $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$.

Table-1 Summary statistics of the populations P1 and P2

Parameter		*N*	*n*	$\bar{Y}$	$\bar{X}$	ρ	C_y	C_x	C_{yx}
Population	**P1**	10	4	52	200	− 0.94	0.1562	0.0458	− 0.00673
	P2	81	13	33.83457	112.4568	− 0.69	0.2972	0.1256	− 0.02576

From Table 1 it can be seen that among P1 and P2, the population P1 has higher correlation than P2. For the population P2, the coefficient of variation of auxiliary and study variable is higher than the population P1.

Table 2: Range of α_1 and α_2 for $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ to be more efficient than the estimators considered

Estimator	**Range of α_1 and α_2 for Populations**			
	P1		**P2**	
$\bar{y}$	$\alpha_1 > 0.156$	$\alpha_2 > 0.156$	$\alpha_1 > 0.306$	$\alpha_2 > 0.306$
$\bar{y}_{RB}$	$\alpha_1 \in (0.185, 1)$	$\alpha_2 \in (0.185, 1)$	$\alpha_1 \in (0.441, 1)$	$\alpha_2 \in (0.441, 1)$
$\bar{y}_{BT}$	$\alpha_1 \in (0.169, 2)$	$\alpha_2 \in (0.169, 2)$	$\alpha_1 \in (0.362, 2)$	$\alpha_2 \in (0.362, 2)$
$\bar{y}_{NK}$	$\alpha_1 \in (0.169, 0.338)$	$\alpha_2 \in (0.169, 0.338)$	$\alpha_1 \in (0.250, 0.723)$	$\alpha_2 \in (0.250, 0.723)$
η **(optimum value)**	$\eta = 0.312$		$\eta = 0.613$	

Table 2 contains the range of α_1 and α_2 for the estimators $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ respectively, to be more precise. The optimum values for P1 and P2 are 0.312 and 0.613 respectively, at this optimum

value the proposed estimators are more efficient than all the estimators taken under consideration for study.

Table 3: MSE and Bias of the proposed and the estimators considered

Estimator	Population			
	P1		P2	
	MSE	|Bias|	MSE	|Bias|
$\bar{y}$	9.8960	0.0000	6.5298	0.0000
$\bar{y}_{RB}$	5.2917	0.0525	3.8878	0.0563
$\bar{y}_{BT}$	7.3812	0.0283	4.9176	0.0324
$\bar{y}_{NK}$	8.5851	0.0136	5.6493	0.0152
$\bar{y}_{\alpha_1}$	1.1519	0.1365	3.4209	0.1022
$\bar{y}_{\alpha_2}$	1.1519	0.0841	3.4209	0.0459

It can be observed from Table 3 that MSE of proposed estimators $\bar{y}_{\alpha_1}$ and $\bar{y}_{\alpha_2}$ is less than the MSE of all the estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$. The MSE of the proposed estimators for population P1 is less than that of population P2. Further the MSE of $\bar{y}_{RB}$ is less than $\bar{y}_{RB}$ and $\bar{y}_{BT}$ for both the populations P1 and P2. It can be observed on taking the modulus value of the Bias that the proposed estimator $\bar{y}_{\alpha_1}$ has less bias than the estimator $\bar{y}_{\alpha_2}$ for both the datasets.

Table 4: Percent relative efficiency w.r.t the estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$

Population	Estimator	Percent relative efficiency w.r.t			
		$\bar{y}$	$\bar{y}_{RB}$	$\bar{y}_{BT}$	$\bar{y}_{NK}$
P1	$\bar{y}$	100.0000	53.4731	74.5877	86.7532
	$\bar{y}_{RB}$	187.0094	100.0000	139.4864	162.2371
	$\bar{y}_{BT}$	134.0712	71.6916	100.0000	116.3104
	$\bar{y}_{NK}$	115.2695	61.6382	85.9769	100.0000
	$\bar{y}_{\alpha_i}$	859.1065	459.3888	640.7848	745.2991
P2	$\bar{y}$	100.0000	59.5393	75.3101	86.5157
	$\bar{y}_{RB}$	167.9562	100.0000	126.4879	145.3084
	$\bar{y}_{BT}$	132.7843	79.0589	100.0000	114.8792
	$\bar{y}_{NK}$	115.5860	68.8192	87.0479	100.0000
	$\bar{y}_{\alpha_i}$	190.8796	113.6485	143.7516	165.1406

The findings of Table 4 reveal that the proposed estimators $\bar{y}_{\alpha_i}$ (i = 1, 2) are more e fficient than the estimators $\bar{y}$, $\bar{y}_{RB}$, $\bar{y}_{BT}$ and $\bar{y}_{NK}$ considered in the study for both the population s P1 and P2.

6. Conclusion

The proposed exponential product type estimators $\bar{y}_{\alpha_i}$(i=1, 2) were found estimating the population mean precisely than the sample mean estimator, product type estimator of Robson (1957), exponential product type estimators of Bahl and Tuteja (1991) and Onyeka (2013) theoretically as well as empirically. Further as per the empirical study conducted, the optimum values of α_1 and α_2 for the proposed estimators were found 0.312 and 0.613 for the populations P1 and P2 respectively.

References

Bahl, S. and Tuteja, R. K. (1991). Ratio and product type exponential estimator. *Information and Optimization Sciences*, **12**(**1**), 159-163.

Cochran, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science,* **30**, 262-275.

Haq, A. and Shabir, J. (2014). An improved estimator of finite population mean when using two auxiliary attributes. *Applied Mathematics and Computation*, ***241***, 14-24.

Hussain, S., Sharma, M. and Bhat, M. I. J. (2021). Optimum exponential ratio type estimators for estimating the population mean. *Journal of Statistics Applications and Probability Letters*, **8**(**2**), 73-82.

John, E. E. and Inyang, E. E. (2015). Efficient exponential ratio estimators for estimating the population mean in simple random sampling. *Hacettepe Journal of Mathematics and Statistics*, **44**(**3**), 689-705.

Murthy, M. N. (1964). Product method of estimation. *Sankhya*, **26**, 69–74.

Gujarati, D. N. (2004). *Basic Econometrics*. McGraw-Hill Companies, New York.

Kadilar, C. and Cingi, H. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, **151**, 893–902.

Mishra, S., Kumar, B., Yadav, S. K., Yadav, D. K. and Shukla, A. K. (2017). An improved ratio type predictive estimator for estimating finite population mean using auxiliary information. *International Journal of Technology*, **6**(**6**), 524-530.

Onyeka, A. C. (2013). A class of product type exponential estimators of the population mean in simple random sampling. *Statistics in Transition*, **14**(**2**), 189-200.

Reddy, V. N. (1973). On ratio and product methods of estimation. *Sankhya*, **B35**(**3**), 307–316.

Reddy, V. N. (1974). On a transformed ratio method of estimation. *Sankhya*, **C36**, 59–70.

Robson, D. S. (1957). Applications of multivariate polykays to the theory of unbiased ratio type estimation. *Journal of the American Statistical Association*, **52**, 511–522.

Singh, H. P. and Karpe, N. (2010). Estimation of mean, ratio and product using auxiliary information in the presence of measurement errors in sample surveys. *Journal of Statistical Theory and Practice*, **4**(**1**), 111–136.

Singh, H. P. and Kumar, S. (2008). A general family of estimators of finite population ratio, product and mean using two phase sampling scheme in the presence of non-response. *Journal of Statistical Theory and Practice*, **2**(**4**), 677–692.

Singh, H. P. and Vishwakarma, G. K. (2008). Some families of estimators of variance of stratified random sample mean using auxiliary information. *Journal of Statistical Theory and Practice*, **2**(**1**), 21–43.

Shukla, N. D. (1976). Almost unbiased product type estimator. *Metrika*, **23**, 127.

Vishwakarma, G. K., Singh, R., Gupta, P. C. and Pareek, S. (2016). Improved ratio product type estimators of finite population mean in simple random sampling. *Revista Investigacion Operacional*, **36**(**3**), 70-76.

Zaman, T. (2020). Generalized exponential estimators for the population mean, *Statistics in Transition*, **21**(**1**), 159-168.

Zaman, T. and Kadilar, C. (2019). Novel family of exponential estimators using information of auxiliary attribute. *Journal of Statistics and Management Systems,* **22** (**8**),1499-1509.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 59-71

Population Growth in India in the Twenty-First Century- Huge Regional Imbalances Foreseen

Purushottam M. Kulkarni
Former Professor, *Jawaharlal Nehru University*, *New Delhi*

Received: 24 May 2021; Revised; 12 June 2021: Accepted: 15 June 2021

Abstract

During the second half of the last century, rapid population growth was a matter of serious concern in India as it was viewed as an impediment to the country's development. However, as India progressed well in demographic transition, fertility declined impressively and recent data show that India reached replacement fertility in 2017. Yet, population is bound increase further due to momentum of population growth and also on account of anticipated decline in mortality. On the other hand, fertility is expected to fall below replacement level offsetting some of the growth. In order to assess likely future population trends, population of India is projected up to 2101 on the basis of recent trends in fertility and mortality and global patterns. The results show that India's population is projected to reach a peak of 1.66 billion around 2060 and then decline slowly. During the process, the age structure of the population will change drastically with a steep rise in the share of the elderly and a drop in the share of the young population and a rise in the share of the working age population for some time yielding the demographic dividend. Projections for large states show that future growth will vary conspicuously across states, since the timing and tempo of the transition have varied regionally. This will result in changing shares of states in the national population substantially. The regional imbalances in population growth have social, economic, and political implications.

Key words: Population projections; Peak population; Inter-state variations; Growth imbalances.

1. Introduction

India experienced enormous population growth during the twentieth century. The population more than quadrupled over the century and the growth was particularly rapid during the second half. The growth occurred due to a welcome development, namely, mortality decline, which was notable after 1921 and quite impressive after 1951. As fertility remained high for some time after mortality decline had set in, rapid population growth naturally took place. This, in turn, raised Malthusian concerns about the impact of population increase on resources and the level of living and consequently neo-Malthusian programmes were introduced in India as in many developing countries. Over time, especially towards the later part of the century, impressive fertility decline was seen and the last decade did show a small fall in the population growth rate. India has further progressed in the demographic transition and the recent estimates of fertility and mortality obtained from the Sample Registration System (SRS) show that fertility has now reached the highly sought replacement level. Thus, the specter of population explosion is no longer seen.

Corresponding Author: Purushottam M. Kulkarni
Email: pmkulkarni.jnu@gmail.com

However, it is well recognised that due to momentum of population growth, population will continue to grow for some time even after reaching replacement level fertility (Keyfitz, 1971). On the other hand, fertility is widely predicted to fall below replacement level and in the long run a declining trend in population size is expected after the effect of population momentum wears off.

Some pertinent questions relevant for development planners and for the society as a whole are: How long will India's population continue to rise? Will it reach a peak and then begin to decline, and if so, what is likely to be the peak, and when would it be reached? Further, given that demographic transition in India has varied spatially, with some states well ahead in transition and some lagging far behind, the peaks are also likely to vary. Will this create regional imbalances in future population growth?

This paper seeks to address these questions on the basis of projections of population for India and for its large states. First, an overview of population growth since 1901 is presented to provide the background for the work followed by the methodology in brief. Results based on projections are presented for India as a whole and then for large states so as to understand variations. Implications of these findings are discussed at the end.

2. Population Growth in India, 1901-2011

The decennial census is the principal source of population size and structure in India. The first census organized around 1871 was not synchronous but since 1881 censuses have been conducted fairly regularly at intervals of 10 years up to 2011; the census scheduled for 2021 has been postponed in view of the COVID19 pandemic. The trends in population size based on censuses of 1901 to 2011 show that the population increased from 238 million in 1901 to 1029 million in 2001, and further to 1211 million in 2011 (see Table 1, Appendix). The population has been growing since 1921 and the growth has been particularly rapid during 1951 to 2001, with the decadal growth being over 20 percent and annual rate of growth above or close to two percent. The decade of 1991 to 2001 did show a small but perceptible fall in the growth rate and the rate declined notably, to well below two percent, during 2001-2011.

The growth of population was caused by mortality decline and as it happens in the process of classical demographic transition, fertility decline was lagged, and hence large transitional growth took place for some time [Table 2 (Appendix) shows trends in key indicators of mortality and fertility for India]. But towards the end of the twentieth century, fertility began to decline impressively curbing the population growth rate somewhat. Fertility decline has continued and recent estimates show that the Total Fertility Rate (TFR) declined to 2.18 in 2017 and 2.15 in 2018 (Registrar General, 2019; Registrar General, 2020a). In conjunction with the female life table constructed based on the age-specific death rates for these years, it can be inferred that the Net Reproduction Rate (NRR) has fallen to a shade below one fertility has achieved the replacement level.

3. Methodology

Reaching replacement level fertility does not imply that population growth will stop immediately. It has been well recognized that after a population with high fertility lowers fertility to replacement level, the population will continue to grow for some time before converging to the stationary state and size due to the phenomenon of population momentum. If an initially stable population attains replacement level fertility within a short time and

fertility and mortality remain constant after that, the ultimate size can be obtained using mathematical formulas. But these would not be applicable to India since fertility decline in India has been gradual and the population was not stable when the fertility decline began as mortality was declining for quite some time before the onset of fertility decline. Moreover, it has been observed that once fertility declines to replacement level, it does not remain constant but instead declines further, going below replacement. Therefore, to have an idea of future growth, it becomes necessary to project population. The standard procedure for this is component projection, which applies projected values of the three components, fertility, mortality, and migration to the baseline population and projects population in future. This allows projection of population along with the age-sex distribution. The technique is straight forward; it involves applying life table survival ratios and age-specific fertility rates to baseline population, making adjustments for migration as appropriate, and continuing the process in time steps. The projections would be as good as the projections of fertility, mortality and migration rates are. Generally, recent trends in these factors are extrapolated. However, there are variations in such projections and hence often alternate projections are made. The United Nations Population Division (UNPD) provides alternate projections called 'variants' with one labeled as 'medium'. This variant is commonly used as the most plausible projection. This is a deterministic approach. Recently, probabilistic projections have been attempted by the UNPD (U.N., 2019). These provide a set of projections at various percentiles. The median projection is equivalent to the 'medium' variant of the deterministic projection. In the present paper, we adopt the standard procedure of component projection using extrapolation of recent trends in fertility and mortality and provide a single projection rather than alternatives. The projections have been carried out up to 2101, that is, the beginning of the next century.

International migration is assumed to be negligible and hence no adjustment has been made for it. The sex ratio at birth, which has become more masculine than the natural level in India in the past two-three decades due to the practice of sex selection is projected to linearly decline to 1.06 by 2031 and remain at that level thereafter on the assumption that, with laws and campaigns to prevent sex selection and social change, this practice would gradually fade away. The baseline for the projections is the 2011 census, that is, March 1, 2011. The age-sex distribution from the census has been used. This has been smoothed following the method adopted by the Registrar General's office (for details, see Technical Group on Population Projections, 2006). From this baseline, population has been projected in intervals of five years up to 2101. The computer package MORTPAK has been used for this purpose.

The demographic transition in India has varied spatially in its timing and pace. Some parts experienced rapid fertility decline in the 1970s and 1980s and achieved replacement level low fertility quite early whereas some states have lagged in the process and are yet to reach low fertility. The latest data from the SRS show that among large states or union territories (those with population of over five million in 2011) of India, 13 (Delhi, West Bengal, Andhra Pradesh, Himachal Pradesh, Jammu & Kashmir, Punjab, Tamil Nadu, Telangana, Karnataka, Kerala, Maharashtra, Uttarakhand, and Odisha) had TFR of less than 2 in 2018 and thus were well below replacement level and three (Gujarat, Haryana, Assam) had TFR of 2.1-2.2 and thus were at replacement level (Registrar General, 2020a). Only Bihar, Uttar Pradesh, Madhya Pradesh, Rajasthan, Jharkhand, and Chhattisgarh had fertility above replacement level though these states have also shown a clear declining trend in fertility and are expected to reach replacement level in the near future. The level of mortality also varies across states with Kerala having very low mortality at present and the north-central states far behind though these have also shown notable declines in mortality. The

latest life tables based on the SRS show that during 2014-18 female life expectancy was high, 77.9 and 77.0 respectively, in Kerala and Delhi, and low, 65.8 and 66.6 in Uttar Pradesh and Chhattisgarh respectively, with other states showing values within this range (Registrar General, 2020b). A similar picture is seen for male life expectancy.

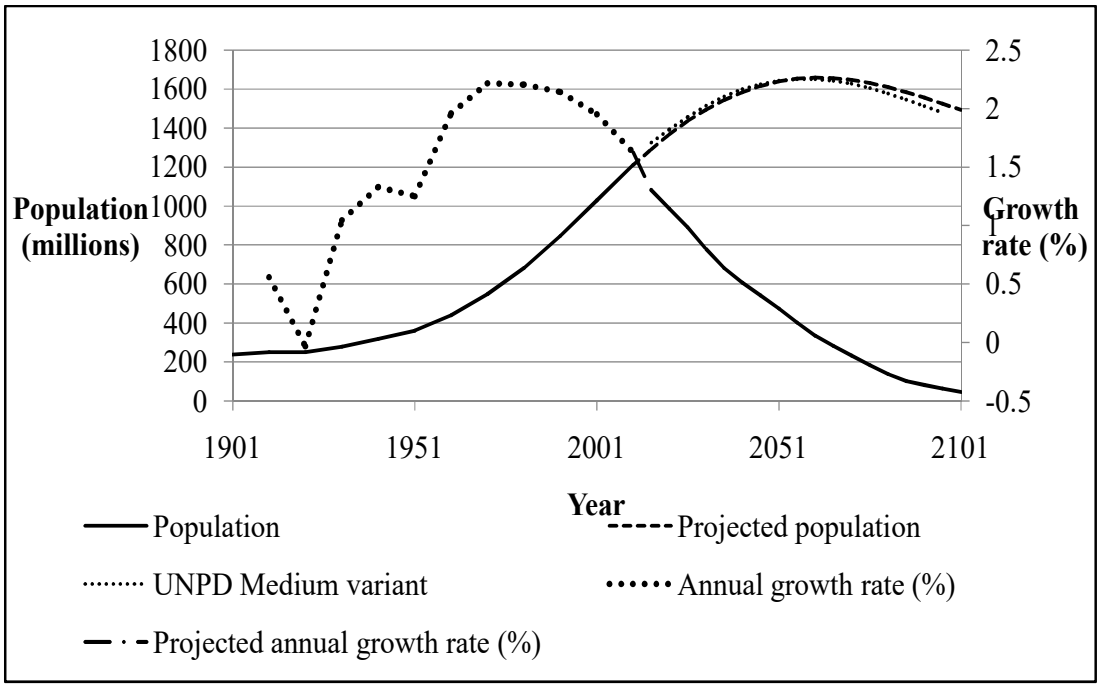
Given the inter-state variations in the initial parameters, future population trends are bound to vary across states. Therefore, it is necessary to make projections for each state individually. The method is identical to that adopted for India. For the national projection, migration was not factored in since international migration is negligible. But inter-state migration does take place and this is notable for some states. Ideally, therefore, inter-state migration should be accounted for in projections for states. On the other hand, projections without making adjustment for migration give an idea of natural increase and reveal expected inter-state imbalances in the absence of migration and thus give an indication of the population pressures for migration. Therefore, it was consciously decided to carry out the projections for states without factoring in inter-state migration. This has been done here for the states/union territories with 2011 population over five million each for which past trends in indicators of fertility and past life tables are available from the SRS there by allowing extrapolation of trends. These states together accounted for 98.5 percent of India's population in 2011. Projections for smaller states/union territories are not done because of non-availability of reliable data on past trends. A single projection has been made for Andhra Pradesh and Telangana combined since past trends in fertility and mortality were not available separately for these states prior to 2014. Similarly, a single projection has been made for Jammu and Kashmir and Ladakh.

4. Results

4.1. Projections for India

The projections show that India's population is expected to continue to grow though at a slower pace up to around 2060 and reach a size of 1659 million. After 2061, a gradual decline is expected and the population is projected to come down to 1500 million by the end of the century. The projected trends in population size and annual growth rate along with trends from 1901 to 2011 are shown in Figure 1. The figure also shows the medium variant in the latest projections by the UNPD in its report (U.N., 2019).The peak of the medium variant in the UNPD projections is 1652 million attained in 2059. Another independent projection by Vollset *et al.* (2020) shows a peak of 1600 million in 2048 in the 'reference' variant, which assumes a more rapid decline in fertility than that in the U.N. projections. It is seen that the U.N. projections agree closely with the projections performed following the extrapolations described above in this paper. This provides confidence to use the projections done in this paper for further discussion.

During the transition from high to low fertility, the age structure of the population also changes. As the growth rate of the population declines, birth cohorts become smaller and with a rise in longevity, population in old ages rises, that is, the populations ages. The remarkable change in the age structure of the population from 2011 to 2101 can be seen from the population pyramids shown in Figure 2 (Appendix). During the process, the share of the Middle Ages, or the working ages, rises substantially for some time leading to the well recognised 'demographic dividend' or the 'window of demographic opportunity' (Bloom *et al.,* 2003). In India, generally the age group 15-59 is considered as working ages and 60+ as old whereas the international convention is to label 15-64 as working ages and 65+ as old. As



Sources: 1901 to 2101*: Registrar General, Census of India* 2011*, Table A*-2.
After 2011: *Author's projections.*
UNDP Medium variant: U.N. (2019)

Figure 1: Past and projected trends in population size (in millions) and annual growth rate (in percent) for India, 1901 to 2011 and 2011 to 2101

longevity has increased considerably in India and is expected to rise further through this century, the international convention of categorization of age has been adopted here. Trends in projected shares of the three broad age groups, 0-14 (young), 15-64 (working) and 65+ (old) are shown in Table 3 (Appendix). A sharp fall in the share of the young ages is seen with a slow but steady rise in that of the old ages. The dependency ratio (ratio of population in young and old ages, considered as 'dependent', to that of the population in working ages) falls from a high of 67 percent in 2001 to 43 percent in the late 2030s.However, after that the rise in old age population becomes notable and the dependency ratio begins to rise. Thus, the advantage of the demographic dividend begins to decline, that is, the window begins to close after 2040, though the dependency ratio would remain below 50 percent till about 2060.

4.2. Projections for states

Projections for large states show that all would grow for some time but then begin to decline slowly. However, states ahead in demographic transition will reach a peak quite early and hence future growth for these would not be large. Table 4 presents the 2001 and 2011 census populations and the projected population in 2101 for large states. The table also shows projected peak populations and the year when the peak is expected to be reached. In Kerala and Tamil Nadu, leaders in the transition, the peak is expected by 2040. Several states like Himachal Pradesh, Punjab, West Bengal, Maharashtra, Delhi, Andhra Pradesh/Telangana and Karnataka are expected to reach peaks during the 2040s and then their populations would begin to decline. Uttarakhand, Jammu and Kashmir, Odisha, Haryana and Gujarat are expected to follow suit during the 2050s. States lagging in transition, Assam, Chhattisgarh, Jharkhand and Madhya Pradesh would grow for a longer time, up to 2060s and Uttar Pradesh

and Bihar even beyond it. The variations in the progress of transition result in corresponding variations in future growth. For the leaders, the 2101 population would not be much different from the 2001 population; in other words, there would be hardly any net growth over the century and for some states even a small decline [see the last column of Table 4 (Appendix)]. In these states, growth due to momentum would be offset by the decline caused by fertility being below replacement fertility for a long time. On the other hand, there would be enormous growth in Uttar Pradesh and Bihar over the twenty-first century, exceeding 100 million each and substantial growth in Rajasthan and Madhya Pradesh, over 40 million each. Such large variations in growth mean huge growth imbalances. Though the 2101 populations of all these states are also projected to be below their peaks, these would still be very large since the peak would be reached quite late in these. In 2101, the population of Uttar Pradesh is projected to be about 300 million, of Bihar 200 million, and of Rajasthan and Madhya Pradesh 100 million. Note that these projections do not factor in inter-state migration. If substantial migration does take place from these states, which are growing rapidly, to states ahead in transition which are expected to reach negative growth early, the imbalances would not be so large. However, for states which are projected to show very large increases over the period, this calls for massive out-migration, which seems unlikely.

High imbalances in growth imply large changes in shares of states in the national population. This can be seen from a comparison of the actual shares in 2001 and 2011 and projected shares at the middle of the century, in 2051, and at the beginning of the next century, in 2101, shown in Table 5 (Appendix). Over the century, Bihar and Uttar Pradesh would increase their share by about five percentage points and Rajasthan and Madhya Pradesh by over one point. On the other hand, Maharashtra, West Bengal, and Tamil Nadu would lose over two percentage points as would Andhra Pradesh and Telangana combined. Karnataka and Kerala would lose over one point. Such changes in share have major economic and political implications.

5. Conclusions

The findings of the study bring out several issues. First, the projections show that India's population will stop growing after some time, reaching a peak of about 1.66 billion most probably around 2060, and then begin to decline. The medium variant of the projections by the UNPD also gives fairly similar results. Thus, India is no longer in the alarming situation of population explosion that was often feared in the not too-distant past. India has completed demographic transition and is projected to move towards below replacement fertility and negative growth after some time. As a result of the transition, the age distribution is bound to change with a sharp decline in the share of young age population and rise in that of working age population. This will yield the well known demographic dividend that has drawn the attention of policy makers and the society as a whole in the recent years. The dependency ratio has by now fallen below 50 percent and it could be said that India has already entered the stage of drawing the dividend. The age distribution would be most favourable in the late 2030s, with the dependency ratio dropping to 43 percent and though the ratio would rise after 2040, as the old age dependency would increase, the degree of dependency would remain below 50 percent until 2060. Thus, India is in a position to benefit from an age structure conducive to economic development for quite some time. However, this is an opportunity and the dividend needs to be harvested by raising the quality of the labour force, creating employment, and increasing female labour force participation. With the passage of time, as the population ages, the share of elderly population would increase giving

rise to large old age dependency. This calls measures to address the needs of the elderly in the areas of health and financial and social security.

The results also show that there would be huge inter-state imbalances in population growth. This is a consequence of the differences in the process of demographic transition across states. While states ahead in the transition would soon reach their peak populations and their populations would then begin to decline, those lagging in transition would grow for a longer time and thus experience massive growth during this century. The growth imbalances would put intense population pressures on the states lagging behind in transition. This could induce large scale migration to states which would have begun to experience negative growth and face labour shortages. Such migration would drive economic growth as this would lead to redistribution of labour to meet the demands in various states. However, this has social implications as the resident population often resists large streams of in-migrants. Besides, migrants also face difficulties at the place of destination and these become critical at the time of crisis as seen recently in the lockdown on account of the COVID 19 pandemic.

In India's structure of governance, state population size is a crucial factor. The number of members to the *Lok Sabha* is based on state's share in the national population. If the shares of some states fall, as has been projected, the number of *Lok Sabha* members from the state would fall with the state losing its weight in the parliament. A constitutional amendment was made in 1976 to freeze the parliamentary seats for each state for some time and this provision has been extended to 2025. But this is a contentious matter in inter-state relations. Similarly, in financial transfers from the central government to states, population size is an important factor. Growth imbalances would impact such transfers. These are issues that need to be addressed in the coming years.

References

Bloom, D. E., David, C., and Jaypee, S. (2003). *The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change*. The Rand Corporation: Santa Monica Ca.

Jain, S. P. (1982). Mortality trends and differentials, *in Population of India, Country Monograph Series No.* 10. *ST/ESAP/220.* New York, 135-137.

Keyfitz, N. (1971). On the momentum of population growth. *Demography*, **8**(**1**),71–80.

Registrar General (2015). *Compendium of India's Fertility and Mortality Indicators 1971-*2013 *based on the Sample Registration System* (*SRS*), Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2016). *Sample Registration System Statistical Report* 2014. Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2017). *Sample Registration System Statistical Report* 2015. Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2018). *Sample Registration System Statistical Report* 2016. Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2019). *Sample Registration System Statistical Report* 2017. Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2020a). *Sample Registration System Statistical Report* 2018. Office of the Registrar General and Census Commissioner, India, New Delhi.

Registrar General (2020b). *SRS based Abridged Life Tables*2014-18.Office of the Registrar General and Census Commissioner, India, Delhi: Ministry of Home Affairs.

Registrar General and Census Commissioner. *Census of India* 2011,*Table A*-2. *Decadal variation in population since 1901*.**www.censusindia.gov.in**

Rele, J. R. (1982). Trends and differentials in fertility. In: *Population of India, Country Monograph Series No.* 10. *ST/ESAP/*220. New York, 91-108.

Technical Group on Population Projections (2006). *Census of India* 2001, *Population Projections for India and States, 2001-2026, Report of the Technical Group on Population Projections***,** National Commission on Population, Ministry of Health and Family Welfare, New Delhi.

Technical Group on Population Projections (2020). *Census of India* 2011, *Population Projections for India and States,* 2011-2036*, Report of the Technical Group on Population Projections,* National Commission on Population, Ministry of Health and Family Welfare, New Delhi.

United Nations (2019). *World Population Prospects* 2019.United Nations, Department of Economic and Social Affairs, Population Division 2019. https://population.un.org/wpp, accessed on August 10, 2019.

Vollset, S. E., Emily G., Chun-Wei, Y., Jackie C. *et al*. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the global burden of disease study. *Lancet* (Published online July 14, 2020).https://doi.org/10.1016/S0140-6736(20)30677-2.

APPENDIX

Table 1: Population growth in India since 1901

Year of Census	Population (in thousands)	Intercensal increase		
		Absolute (in thousands)	Percent	Annual exponential growth rate (%)
1901	238396			
1911	252093	13697	5.7	0.56
1921	251321	-772	-0.3	-0.03
1931	278977	27656	11.0	1.04
1941	318661	39683	14.2	1.33
1951	361088	42428	13.3	1.25
1961	439235	78147	21.6	1.96
1971	548160	108925	24.8	2.22
1981	683329	135169	24.7	2.20
1991	846421	163092	23.9	2.14
2001	1028737	182316	21.5	1.95
2011	1210570	181832	17.7	1.63

Source: *Extracted from Registrar General and Census Commissioner*, *India, Census of India*, *Table A*-2. *www.censusindia.gov.in*

Table 2: Trends in fertility and mortality, India, 1901-2015

Period	CDR (Crude Death Rate per 1000 population)	Life Expectancy (years)			CBR (Crude Birth Rate Per 1000 population)	TFR (Total Fertility Rate)
		Both sexes	Male	Female		
Census based estimates						
1901-10	42.6	22.9			49.2	
1911-20	48.6	20.1			48.1	
1921-30	36.3	26.8			46.4	
1931-40	31.2	31.8			45.2	
1941-50	27.4	32.1			39.9	
1951-60	22.8	41.3			41.7	
1961-70	19.0	45.6			41.1	
SRS (Sample Registration System) based estimates						
1971-75	15.5	49.7	50.5	49.0	35.6	5.02
1976-80	13.8	52.3	52.5	52.1	33.4	4.50
1981-85	12.1	55.4	55.4	55.7	33.6	4.46
1986-90	10.6	57.7	57.7	58.1	31.4	4.00
1991-95	9.5	60.3	59.7	60.9	28.9	3.54
1996-2000	8.8	61.9	61.2	62.7	26.6	3.27
2001-05	7.9	64.3	63.1	65.6	24.6	2.97
2006-10	7.4	66.1	64.6	67.7	22.8	2.63
2011-15	6.9	68.3	66.9	70.0	21.3	2.34

Sources:
Census based estimates for 1901-70:
CDR and life expectancy: *Jain* (1982), *CBR*: *Rele*(1982).
SRS based estimates:1971 *onwards*:
CDR and CBR: *Averages computed from Registrar* General (2015, 2016, 2017, 2018, 2019, 2020a);
Life expectancy: *Registrar General* (2020b).

Table 3: Trends in population size, broad age distribution and dependency ratio, 2001, 2011 census and 2021 onwards, projected

Year	Population (in millions)	Ages 0-14 years (%)	Ages 15-64 years (%)	Ages 65+ Years (%)	Dependency ratio (ages 0-14 plus 65+) /ages 15-64 years
2001	1029	35.4	59.8	4.8	67.3
2011	1211	30.9	63.7	5.5	57.1
2016	1292	28.3	65.7	5.9	52.2
2021	1368	26.3	67.3	6.4	48.6
2026	1437	24.9	68.2	7.0	46.7
2031	1496	23.1	69.0	7.9	44.9
2036	1544	21.4	69.6	9.0	43.8
2041	1584	19.9	69.7	10.3	43.4
2046	1616	18.8	69.5	11.7	43.9
2051	1640	18.0	68.7	13.2	45.5
2056	1654	17.3	67.7	15.0	47.7
2061	1659	16.6	66.4	17.0	50.6
2066	1656	16.0	64.9	19.1	54.2
2071	1647	15.5	63.5	21.0	57.5
2076	1631	15.2	62.5	22.3	60.1
2081	1609	14.9	61.3	23.8	63.2
2086	1583	14.6	60.2	25.2	66.2
2091	1555	14.4	59.1	26.5	69.2
2096	1524	14.2	58.3	27.6	71.6
2101	1492	14.0	57.7	28.3	73.5

Sources:2001*and* 2011, *Technical Group on Population Projections* (2006, 2020).
2021 *onwards; Author's projections*

Table 4: Projected population growth up to 2101 for large states/union territories, India
(*all populations in millions*)

State	2001 Census Population	2011 Census Population	2101 Projected population	Peak Population	Year of peak	Increase 2001-2101
India	1029	1211	1441	1658	2060	412
Andhra Pradesh and Telangana	76	85	76	99	2046	0
Assam	27	31	38	43	2060	12
Bihar	83	104	199	205	2082	116
Chhattisgarh	21	26	33	36	2064	12
Delhi	14	17	15	21	2046	1
Gujarat	51	60	71	80	2056	20
Haryana	21	25	28	33	2055	7
Himachal Pradesh	6	7	6	8	2043	0
Jammu & Kashmir$	10	13	13	16	2054	3
Jharkhand	27	33	45	50	2066	18
Karnataka	53	61	54	72	2046	2
Kerala	32	33	28	37	2040	-4
Madhya Pradesh	60	73	104	112	2068	44
Maharashtra	97	112	95	133	2045	-1
Odisha	37	42	45	52	2054	8
Punjab	24	28	22	32	2043	-2
Rajasthan	57	69	101	109	2070	44
Tamil Nadu	62	72	55	80	2038	-8
Uttar Pradesh	166	200	301	327	2070	135
Uttarakhand	8	10	10	13	2052	1
West Bengal	80	91	77	107	2043	-3

$: Including Ladakh.

Sources: 2001 *and* 2011, *Registrar General and Census Commissioner, Census of India* 2011, *Table A*-2.
Projected populations and peaks: Author's projections.

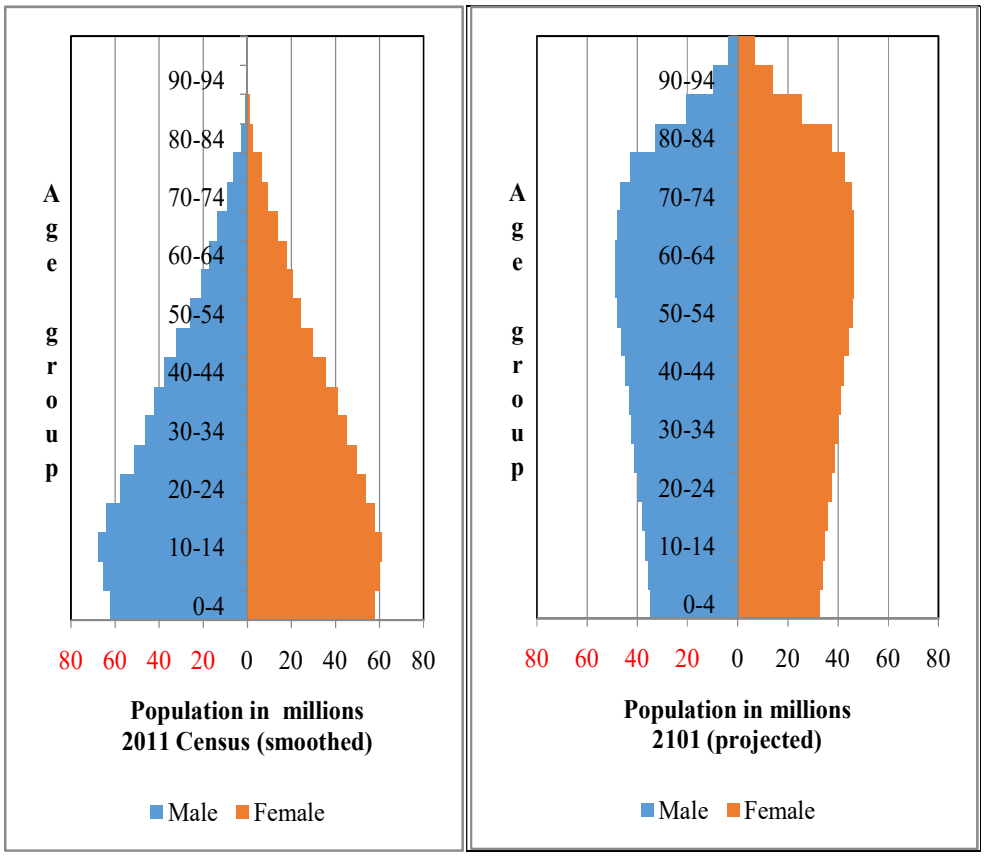
Table 5: Percentage population shares of large states/union territories, India, 2001 and 2011 census and 2051 and 2101 projected

State	2001	2011	2051	2101
Andhra Pradesh and Telangana	7.4	7.0	6.0	5.3
Assam	2.6	2.6	2.6	2.7
Bihar	8.1	8.6	11.2	13.8
Chhattisgarh	2.0	2.1	2.2	2.3
Delhi	1.3	1.4	1.2	1.0
Gujarat	4.9	5.0	4.8	4.9
Haryana	2.1	2.1	2.0	2.0
Himachal Pradesh	0.6	0.6	0.5	0.4
Jammu & Kashmir$	1.0	1.0	1.0	0.9
Jharkhand	2.6	2.7	2.9	3.1
Karnataka	5.1	5.0	4.4	3.8
Kerala	3.1	2.8	2.2	1.9
Madhya Pradesh	5.9	6.0	6.6	7.3
Maharashtra	9.4	9.3	8.0	6.6
Odisha	3.6	3.5	3.1	3.1
Punjab	2.4	2.3	1.9	1.5
Rajasthan	5.5	5.7	6.4	7.0
Tamil Nadu	6.1	6.0	4.7	3.8
Uttar Pradesh	16.2	16.5	19.1	21.0
Uttarakhand	0.8	0.8	0.8	0.7
West Bengal	7.8	7.5	6.4	5.3
Remaining states and union territories	1.6	1.6	1.9	1.9
India	100.0	100.0	100.0	100.0

$: Including Ladakh

Sources: 2001 *and* 2011, *computed from the data in Registrar General, Census of India, Table A-2.*

2021 *onwards; Author's projections.*



Sources: 2011,*Age distribution from Technical Group on Population Projections* (2020), *smoothed.* 2101 *onwards; Author's projections*

Figure2: Population pyramids,India, 2011 census and 2101 projected
(*populations inmillions*)

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 73- 85

Modern Statistical Methodologies in the COVID–19 Pandemic: A Review from a Statistical Perspective

Das, A. and Muralidharan, K.
Department of Statistics, Faculty of Science,
The Maharaja Sayajirao University of Baroda, Vadodara

Received: 28 May 2021; Revised: 14 June 2021; Accepted: 16 June 2021

Abstract

Ever since the world has been introduced to the novel corona virus, we have been privy to multiple statistical, epidemiological, and deep learning models to understand the biological prevalence of the disease. Neural Networks having played a pivotal part in this sphere, aimed at varied objectives. The most common among them have been (1) forecasting of transmission dynamics based on daily data, (2) classification of infected and non-infected patients based on various scans and images and (3) to analyze survival data to determine incubation period / time to death / duration of hospitalization, or to analyze competing risks, among others. This paper aims to showcase some of those Neural Networks that have been crucial in either handling or monitoring or in some cases, providing useful insight into the pandemic.

Key words: Corona virus; Pandemic; Artificial Neural network; Artificial Intelligence; Machine Learning; Deep Learning; Classification; Regression modeling.

1. Introduction

The advent of the novel corona virus, or technically SARS-CoV-2, or as it is more popularly known, COVID – 19, in Wuhan, Hubei province, People's Republic of China, as reported by the Wuhan Municipal Health Commission, on December 31, 2019, is a global phenomenon today, with more than 154 million infected, 90.3 million recovered and 3.22 million deaths to its name, as per the World Health Organization, as of May 4, 2021. During this time, a lot of time, effort, and money has been allotted by organizations and governments worldwide, to mitigate numerous micro-tragedies occurring in spurts across the globe.

A lot of research facilities around the world have invested their resources in developing epidemiological, statistical, and mathematical models to handle crises both at the micro as well as at the macro level, or to monitor transmission dynamics of the virus, or to gain some insight regarding how the virus might spread, the virus' mutation dynamics, patient's survivability in case of mild-to-moderate-to-severe lung damage, optimization of existing tests based on patient vitals, among others. An effort has also been made to monitor and subsequently, minimize the spread of false information in this time of crisis. The second wave of the disease has further boosted the necessity of model-based study to support the transactional issues of the pandemic worldwide.

The neural network model, especially the deep neural network model, since its inception, or more appropriately, since its computational barriers have subsided, has been in

Corresponding Author: K. Muralidharan
Email: lmv_murali@yahoo.com

the limelight for its prediction accuracy and applicability across varied data types. In practice, a deep neural network is any neural network having a large number of layers of neurons. Neural networks are a set of algorithms, modeled loosely after the human brain, that is designed to recognize patterns. They interpret sensory data through various kinds of machine perception, labeling, or clustering of raw inputs. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text, or time series, must be translated. This motivated us to make a review of neural network-based models available so far in statistical literature in understanding the behavior pattern of the COVID-19 pandemic and its mitigation potential.

2. Literature Review

The review is attempted in two statistical perspectives here. In the immediate section, we explore the modern-day techniques involving neural network-related studies and examine their potential in describing the virus patterns. We also extract those studies where the use of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) techniques are effectively used for modeling the pandemic. Section 2.2 will explore the statistical studies which are attempted using probability models and inferences therein.

It may be noted that most of the papers on models addressing transmission dynamics of Covid-19 are associated the reproduction number, R, which is a key epidemiological parameter that quantifies the average number of new infections caused by a single infected individual. When a pathogen emerges in an entirely susceptible population, this parameter is referred to as the basic reproduction number, R_0.

Utmost care is taken to establish the hierarchy of the published articles.

2.1. Neural network models in the pandemic

The paper by Chatterjee *et al*. (2020) addresses the potential of data science to assess the risk factors correlated with COVID-19, after analyzing existing data sets available in "ourworlddindata.org" and newly simulated datasets, following the analysis of different univariate Long Short-Term Memory(LSTM) models for forecasting new cases and resulting deaths. They discussed the findings related to the statistical analysis on simulated datasets.In a related paper, Pirouz *et al*. (2020) aim to study the possible correlation between the numbers of swab tests and the trend of confirmed cases of infection, while paying particular attention to the sickness level. The statistical analysis of this article provided the basis for an AI study by Artificial Neural Network (ANN). Also, they use multivariate linear regression approach to validate their findings.

Shams *et al*. (2020)have investigated the importance of using Generative Adversarial Networks (GAN) as a preprocessing stage to applied DNN (DotNetNuke) for image data augmentation. They presented a case study of using GAN networks for limited COVID-19 X-Ray Chest images. The results indicated that the proposed system based on using GAN-DNN is powerful with a minimum loss function for detecting COVID-19 X-Ray Chest images. Stochastic gradient descent and Improved Adam optimizers are used during the training process of the COVID-19 X-Ray images, and the evaluation results depend on loss function are determined to ensure the reliability of the proposed GAN architecture.

In the study by Yasar and Ceylan (2020), which aims at early diagnosis of Covid-19 disease using X-ray images, the deep-learning approach, a state-of-the-art artificial intelligence method, was used, and automatic classification of images was performed using Convolutional Neural Networks(CNNs). In the first training-test data set there were 230 X-ray images, of which 150 were Covid-19 and 80 were non-Covid-19, while in the second training-test data set there were 476 X-ray images, of which 150 were Covid-19 and 326 were non-Covid-19. Thus, classification results have been provided for two data sets, containing predominantly Covid-19 images and predominantly non-Covid-19 images. A23-layer CNN architecture and a 54-layer CNN architecture were developed, of which results were obtained in the form of chest X-ray images directly in the training-test procedures and the sub-band images obtained by applying dual-tree complex wavelet transform (DT-CWT) to the above-mentioned images. The same experiments were repeated using the image obtained by applying the local binary pattern to the chest X-ray images. Within the scope of the study, four new result generation pipeline algorithms having been put forward additionally, it was ensured that the experimental results were combined and the success of the study was improved. The training sessions were carried out using a23-fold cross-validation method for both datasets.

The aim of the study by Apostolopoulos & Mpesiana (2020) was to evaluate the performance of state-of-the-art CNN architectures proposed for the classification of medical images. A dataset of X-ray images from patients with common bacterial pneumonia confirmed Covid-19 disease, and normal incidents were utilized for the automatic detection of Covid-19. Following two datasets have been utilized in this experiment:

- A collection of 1427 X-ray images including 224 images with confirmed Covid-19 disease, 700 images with confirmed common bacterial pneumonia, and 504 images of normal conditions.
- A dataset including 224 images with confirmed Covid-19 disease, 714 images with confirmed bacterial and viral pneumonia, and 504 images of normal conditions.

The data was collected from the available X-ray images on public medical repositories. The results suggest that Transfer Learning with X-ray imaging may extract significant biomarkers related to the Covid-19 disease, while the best accuracy, sensitivity, and specificity obtained are 96.78%, 98.66%, and 96.46% respectively.

In another related study by Ozturk *et al*. (2020), a new model for automatic COVID-19 detection using raw chest X-ray images was presented. The proposed study is developed to provide accurate diagnostics for binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. pneumonia). The model produced a classification accuracy of 98.08% for binary classes and 87.02% for multi-class cases. The DarkNet model was used in their study as a classifier for the You Only Look Once (YOLO) real-time object detection system. They implemented 17 convolutional layers and introduced different filtering on each layer. Their model can be employed to assist radiologists in validating their initial screening, and can also be employed via the cloud to immediately screen patients.

Cheema *et al.*(2020) designed an intelligent computing paradigm called the Levenberg-Marquardt artificial neural networks (LMANNs) as an attempt to solve the mathematical model SEIPAHRF (Susceptible, Exposed, Infected and symptomatic, super Propagation, infected but Asymptomatic, Hospitalized, Recovered, Fatality)of Coronavirus disease propagation via human to human interaction. The model is represented with systems of nonlinear ordinary differential equations represented with susceptible, exposed,

symptomatic and infectious, super spreaders, infection but asymptomatic, hospitalized, recovery, and fatality classes. The reference dataset of the COVID-19 model is then generated by exploiting the strength of the explicit Runge-Kutta numerical method for metropolitans of China and Pakistan. The created dataset is arbitrarily used for training, validation, and testing processes for each cyclic update in Levenberg-Marquardt's back-propagation for the numerical treatment of the dynamics of the COVID-19 model. The effectiveness and reliable performance of the design LMANNs are endorsed based on assessments of achieved accuracy in terms of mean squared error-based merit function, error histograms, and regression studies. In a separate study, Singh *et al*. (2020) designed and implemented another deep convolutional neural network approach for the data. They tuned the hyper-parameter of CNN using Multi-objective adaptive differential evolution (MADE). Considering the benchmark COVID-19 dataset, extensive experiments were performed. In comparison, it was found that the proposed technique outperforms the competitive machine learning models in terms of various performance metrics.

Li *et al*. (2020) proposed an automated measure of COVID-19 pulmonary disease severity on chest radiographs, for longitudinal disease tracking and outcome prediction based on a convolutional Siamese neural network-based algorithm, which was trained to output a measure of pulmonary disease severity on CXRs. Severity scores based on Siamese neural networks automatically measure radiographic COVID-19 pulmonary disease severity that can be used to track disease change and predict deaths. Before this paper, studies used manually annotated features from chest imaging to predict outcomes, such as mortality, need for intensive care, and other adverse events. However, barriers to the adoption of these systems include inter-rater reliability and the learning curve for users. In this study, raters assessing longitudinal change showed only moderate inter-rater agreement. The automated Siamese neural network-based approach addresses these challenges.

Ardakani *et al*. (2020) suggested a rapid and valid method for COVID-19 diagnosis using multiple AI techniques and then compared them. Ten well known convolutional neural networks were used to distinguish infection of COVID-19 from non-COVID-19 groups: AlexNet, VGG-16, VGG-19, SqueezeNet, GoogleNet, MobileNet-V2, ResNet-18, ResNet-50, ResNet-101 and Xception. Among all the networks, the best performance was achieved by ResNet-101 and Xception. Although the Xeption network gave the best performance, it did not have the best sensitivity. In contrast, ResNet-101 could diagnose COVID-19 infection with the highest sensitivity and implied lower specificity compare to the Xeption network. The authors suggest that a method with the highest sensitivity to diagnose all patients with COVID19 is desirable. In this regard, the ResNet-101 has an advantage over other networks, especially Xeption, due to its highest sensitivity and AUC. ResNet-101 is trained based on residual learning. This kind of learning can facilitate the training of networks by considering the layer inputs as a reference. In addition, residual networks are optimized easier, and the accuracy can be improved as the depth increases, as confirmed by He *et al*. (2016). This specific residual learning can lead to better training and provide a robust model. Hence, the best performance can be obtained using ResNet-101.

Cough analysis is another important area in clinical diagnosis. Bansal *et al*. (2020) proposed a CNN-based audio classifier using the open cough dataset. The dataset is labeled manually into cough categories with the final labeling of COVID and Non-COVID classes. Two approaches proposed in this paper are based on Mel-frequency Cepstral Coefficients (MFCC) features and spectrogram images as input to the CNN network. They found that the MFCC approach produced 70.58% test accuracy with 81% sensitivity and is better than the spectrogram-based approach. The authors opine that the methodology of their

paper is based on a novel CNN architecture with MFCC input and that being a contactless methodology, it serves as a first-hand assessment of the Covid infection for masses with a Smartphone along with other symptoms.Similar work has been seen in the paper by Hassan *et al*. (2020), where they highlight the importance of speech signal processing in the process of early screening and diagnosing the COVID-19 virus by utilizing the LSTM for analyzing the acoustic features of cough, breathing, and voice of the patients. Their result shows that the accuracy of the voice test compared to both coughing and breathing sound is lower.

Haque *et al*. (2020) proposed another CNN model to detect COVID-19 patients from chest X-ray images. This model is evaluated with a comparative analysis of two other CNN models. The architecture of this CNN is such that it has four convolutional layers, the first one being a 2D convolutional layer with 3×3 kernels and a Rectified Linear Unit (ReLU) activation function. The next three layers are 2D convolutional layers along with the ReLU activation function and Max pooling. Max pooling accumulates the features of the convolutional layer by convolving filters over it. In each of the three layers, a 2×2 Max pooling layer is added after the convolutional layer to avoid overfitting and to make the model computationally efficient. Then, the output of the convolutional layers is converted to a long 1D feature vector by a flattened layer. This output from the flattened layer is fed to the fully connected layer with dropout. In a fully connected layer, every input neuron is connected to every activation unit of the next layer. All the input features are passed through the ReLU activation function and this layer categorizes the images to the assigned labels. The Sigmoid activation function makes the classification decision depending on the classification label of the neurons. Finally, in the output layer, it is declared if the input X-ray image is COVID-19 positive or normal. The proposed model performs with an accuracy of 97.56% and a precision of 95.34%.

Govindarajan and Swaminathan (2020) aimed to differentiate COVID-19 conditions from healthy subjects in chest radiographs using simplified end-to-end CNN model and occlusion sensitivity maps. The images were considered from publicly available datasets. In terms of architecture, each convolutional layer was made of a varied number of filters as 8, 16, 32, 64, 128, 256, and 512 with varying filter sizes 3 × 3 and 5 × 5. The convolutions were all zero-padded to preserve the input resolution. Further, each convolutional 2D layer was followed by a batch normalization layer and ReLU activation layers. An epsilon value of 1e-5 was chosen for batch normalization. Further, a max-pooling layer was followed with a stride of 2 and a pool size of 2. Resultant feature maps were fed to a fully connected layer, followed by the Softmax layer followed by yet another classification layer. In this work, the output size of 2, corresponding to the two classes was fixed. The weights were initialized using the Glorot method on all layers where each weight was initialized from$N(0, k), k < \infty$. Finally, the classification scores are obtained to generate visualization-based occlusion sensitivity maps. Significant biomarkers representing critical image features were extracted from the CNN by experimentally investigating cross-validation methods and hyperparameter settings. Occlusion Sensitivity Maps work when the occluded images are passed through the CNN network and Euclidean distances are calculated for each iteration. The differences between the occluded distances and nonoccluded distances are computed. This difference increases once a patch occludes an area in the image relevant to the network, thus creating a heat map. The pathological regions correspond to higher probability and a drop in the value indicates that the pathological locations have been occluded. In this study, a mask size of 15 and a stride value of 10 pixels are heuristically chosen. Thus, COVID-19 specific CXR biomarkers can be localized providing an approximate visual diagnosis. The performance of this network was evaluated using standard metrics. Results describe that the simplified CNN model with

optimized parameters can extract significant features with a sensitivity of 97.35% and F-measure of 96.71% to detect COVID-19 images.

The primary innovation of the method proposed by Hartono (2020) for developing a transmission dynamics predictor that takes advantage of the time difference among many countries with respect to the transmission of this disease is that the proposed method only requires the transmission similarities between countries in the publicly available data for this current disease. For this purpose, two models have been used, the topical autoencoder (TA), a hierarchical neural network having a low-dimensional topological hidden layer that is a simplified version of the soft-supervised topological autoencoder (STA) proposed in Hartono (2020), in which the basic mathematical properties were proposed in Hartono *et al*. (2015), and a vanilla LSTM.

Hawas (2020) used RNNs to predict the daily infections in Brazil. His article introduces long-term time-series predictions for the virus's daily infections in Brazil by training forecasting models on limited raw data. The prediction data was generated by training 4200 recurrent neural networks (54 to 84 day's validation periods) on raw data. Militante *et al*. (2020) tried to find out if the patient has these diseases; experts conduct physical exams and diagnose their patients through Chest X-ray, ultrasound, or biopsy of lungs. The study employs a flexible and efficient approach of deep learning applying the model of CNN in predicting and detecting a patient unaffected and affected with the disease employing a chest X-ray image. The study utilized a collected dataset of 20,000 images using a 224x224 image resolution with a 32 batch size is applied to prove the performance of the CNN model being trained. Their trained model gives a 95% accuracy rate. Also, they can try to detect and predict COVID-19, bacterial, and viral-pneumonia diseases based on chest X-ray images.

The ANN-based models proposed in Niazkar and Niazkar (2020) were utilized to estimate the confirmed cases of COVID-19 in China, Japan, Singapore, Iran, Italy, South Africa, and the United States of America. These models exploit historical records of confirmed cases, while their main difference is the number of days that they assume to have an impact on the estimation process. Their results show that the ANN-based model that takes into account the previous 14 days outperforms the other ones. This comparison reveals the importance of considering the maximum incubation period in predicting the COVID-19 outbreak.

Sridhar *et al*. (2020) have analyzed the similarity in features between the novel coronavirus 2019 and various other lung diseases such as Pneumonia, Pneumothorax, Atelectasis, Pleural Thickening, etc. based on Chest X-ray scans in the posteroanterior view for various diseases. CNN using the Residual Network (ResNet) is built to identify similar regions in the chest X-rays of COVID-19 and various lung diseases. The regions of similarity are visualized using class activation maps. A total of eleven conditions affecting the lungs are studied and compared to COVID-19. Their results show that Atelectasis, Consolidation, Emphysema, and Pneumonia are most similar to COVID-19 of the eleven diseases considered.

Toraman *et al*. (2020) have used some deep learning models like CNN, DNN, etc. which are mostly used in radiology images to determine the positive cases. They studied a novel artificial neural network, Convolutional Capsule Neural Network (CapsNet) for the detection of COVID-19, and further proposed using chest X-ray images with capsule networks. The proposed approach is designed to provide fast and accurate diagnostics for

COVID-19 diseases with binary classification (COVID-19, and No-Findings), and multi-class classification (COVID-19, and No-Findings, and Pneumonia). Their proposed method achieved an accuracy of 97.24% and 84.22% for binary class and multiclass respectively.

2.2. Other important models and methodologies

Li *et al*. (2020) introduced a statistical disease transmission model using case symptom onset data to estimate the transmissibility of the early phase outbreak in china and provide sensitivity analyses with various assumptions of diseases. They fitted the transmission model to several publicly available sources of the outbreak data until 11, Feb 2020, and estimated lockdown intervention efficacy of Wuhan city. The estimated R_0 was between 2.7 and 4.2 from plausible distribution assumptions of the incubation period and relative infectivity. They finalized on the most agreed R_0 value of 2.2. Then they compared different modelling approaches. On comparison they found that, there was no negative correlation in the spatial distribution of Covid-19 incidence, and there were no high-value clusters.

In their article, Martin *et al*. (2020) have summarized some of the most important pitfalls, provided intuitive clinical examples (mostly related to other types of infections), and referred to methodological articles explaining the mathematical frameworks of some of the most common neural networks. Their article also describes the common errors which may occur in the presence of competing risks and time-dependent covariates, or when conditioning on future events. They have used Kaplan–Meier survival plots, Cox or logistic regression models as well for comparative purposes.

Wang *et al*. (2020) study publicly released data for 1212 COVID-19 patients in Henan, China using various statistical probabilistic models and network analysis. The incubation period was statistically estimated, and they proposed a state transition diagram that could explore the epidemic stages of emerging infectious diseases. They found that although the quarantine measures were gradually working, strong measures still might have been required for a while, since 7.45% of patients may have had very long incubation periods. They also found that migrant workers and/or college students were at high risk. A related paper by Sivapalan *et al*. (2020) describes the detailed statistical analysis plan for the evaluation of primary and secondary endpoints of the proPAC-COVID study. The purpose of this paper is to provide primary publication of study results to prevent selective reporting of outcomes, data-driven analysis, and to increase transparency. The proPAC-COVID is a multi-center, randomized, placebo-controlled, double-blinded clinical trial. Data is analyzed using intention-to-treat (ITT) principles, and main analyses will also be subject to modified ITT analysis and per-protocol analysis. The paper by Talmoudi *et al*. (2020) describes the transmission dynamics of the outbreak and the impact of intervention measures that are critical to planning responses to future outbreaks and providing timely information to guide policymakers' decisions. They estimated the serial interval and temporal reproduction number of SARS-CoV-2 in Tunisia. They used the maximum likelihood approach to estimate the dynamics of R_t. They collected data of investigations and contact tracing between March 1, 2020, and May 5, 2020, as well as illness onset data during the period February 29-May 5, 2020 from the National Observatory of New and Emerging Diseases of Tunisia.

The findings of Shruthi and Ramani (2020) signify that the statistical volatility transmission differs post food price crisis. Moreover, risk transmission materializes as an additional element of the statistical interrelations among agricultural and energy markets. The major contributions of the article are:

- The various qualities of volatility spillover impacts among each commodity market and crude oil after and before the COVID-19 are evaluated systematically, and the impact of crude oil fluctuations in the prices on different commodity markets is being compared.
- The statistical conditional correlation between each commodity market and the crude oil market is significantly high. The time differing qualities of correlations after and before the COVID-19 are also significant.

Phucharoen *et al*. (2020) have analyzed Thailand's COVID-19 data using the probit model to investigate the risk factors for transmission from confirmed COVID-19 cases to their high-risk contacts. The analysis further focused on the impact of quarantine measures in state-provided facilities on contacts' probability of infection and interpreted that the quarantine policy, which mandated individual isolation in the state provided facilities for all high-risk contacts, diminished the contact's chance of infection from the confirmed cases, especially in the epicenter districts. A comparative study paper by Nayak *et al*. (2020) has made some statistical analysis based on Gaussian modeling, ANOVA test, and probabilistic model approach. After ANOVA they conclude that the recovery rate for all the countries is significantly higher than the mortality rate except for the UK where the mortality rate is significantly higher than the recovery rate. Gaussian modeling applied here was able to predict the original peak values of confirmed cases of countries. Using the probabilistic model approach, the authors were able to predict that there is around a 5% probability that a person in India will be tested positive for COVID-19 on 100 tests.

Milano and Cannataro (2020) analyzed different Italian COVID-19 data at the regional level for the period 24th Feb to 29th Mar 2020. They built multiple similarity matrices using statistical tests and then they mapped those similarity matrices into networks where nodes represent Italian regions. Then the network-based analysis of Italian COVID-19 data can elegantly show how regions form communities, *i.e*., how they join and leave them according to time and how community consistency changes over time and with respect to the different available data.They found that the regions where the epidemic had a greater impact, such as the Lombardi, Veneto, Piemonte and Emilia, had a different behaviour with respect to other regions. This was evident in the community detection in which the regions most affected by the epidemic form individual communities or they are part of the same community. In addition, this study also led to identifying similar behaviours of regions that are geographically distant but that together form community. Another interesting finding is given by Hamzah *et al*. (2020), where the real-time data query is done and visualized; the queried data is used for Susceptible-Exposed-Infectious-Recovered (SEIR) predictive modeling. They utilized SEIR modeling to forecast COVID-19 outbreaks within and outside China based on daily observations. They also analyzed queried news, and classified the news into negative and positive sentiments, to understand the influence of the news on people's behavior both politically and economically.

Li *et al*. (2020) collected available data with respect to all confirmed Covid-19 cases in China and analyzed the spatial patterns of these cases by a Geographical Information Systems (GIS) approach applying ArcGIS software to create a spatial database, after which they applied the ArcGIS geographic statistical analysis module to investigate the spatial distribution trends. Also, Faes *et al*. (2020) investigated the time from symptom onset to diagnosis and hospitalization or the length of stay in the hospital and whether there are differences in the population. The distribution of different event times of different patients

groups is estimated accounting for interval censoring and right truncation of the time intervals.

An interval-censored likelihood framework was adopted in the paper by Zhao *et al*. (2020) to fit three different distributions including Gamma, Weibull, and Lognormal that govern the serial interval of COVID-19. They select the distribution according to the AIC (corrected) for a small sample size. They found that the lognormal distribution performed slightly better than the other two distributions in terms of AIC. Another model-based paper by Livadiotis (2020) focuses on the impact of environmental temperature on the exponential growth rate of cases infected by COVID-19 for US and Italian regions. Also, the researcher has derived growth rates for regions characterized by a readable exponential growth phase and has plotted them against the environmental temperatures averaged within the same regions. This is used to derive the relationship between temperature and growth rate and has evaluated its statistical confidence.

Using simulated epidemic data, O'Driscoll*et al*. (2020) have assessed the performance of seven commonly used statistical methods to estimate R_0 as they would be applied in real-time outbreak analysis scenario, fitting to an increasing number of data points over time and with varying levels of random noises in the data. Method comparison was also conducted on empirical outbreak data, using Zika surveillance data. On early analysis of the SARS-CoV-2 epidemic in China, they estimated a declining trend in the reproduction number from 7.93 (95% CI: 5.00-12.00) on the 29th December 2019 to 2.60 (95% CI: 0.57-5.17) on the 18th January 2020 using the Wallinga and Teunis method. Most methods considered here frequently over-estimate R_0 in the early stages of epidemic growth on simulated data.

3. Open Research Areas and the Future of the Pandemic

Over the last few years, deep neural networks have proven superior over most other models in terms of accuracy and forecasting ability. However, there are arguments against their use, including the fact that they do not play any role in identifying key relationships between variables, among other critiques mentioned in the work of Zador (2019).

One more critique in the case of neural networks may be in the way it is used by researchers worldwide. The fact that the architecture of the ANN is a defining feature of the model itself provides an opportunity for researchers to employ known combinations of varied architecture with considerable ease and, regrettably in some cases, even without proper reasoning, the primary defence of which extends to higher accuracy and architecture novelty. This may be the primary reason behind why so many CNN's have been discussed in this paper, most of which are concerned with the classification of patients into infected/non-infected or COVID-19/Pneumonia/Other Pulmonary Conditions, based on CT-Scan or X-ray Images.

From the works of Shams *et al*. (2020), we can see that they plan to use a Region of Interest feature extractor instead of the whole image to extract more details for medical images. Yasar and Ceylan (2020) planned to analyze the effects of using the obtained results through direct transfer learning in pipeline classification algorithms, on the study results. Ozturk *et al*. (2020), as well as Toraman *et al*. (2020), intend to validate their model by incorporating more images. The resultant model may be placed in a cloud to provide instant diagnosis and help in the rehabilitation of affected patients quickly. Haque *et al*. (2020) have postulated that in addition to improvements in the model, clinical testing shall prove to be helpful in diagnosis. It is also stated in the work by Militante *et al*. (2020) that further studies

may further improve the CNN architecture by hyper-parameter tuning and transfer learning combinations.

Aside from the suggestions put forward by the authors cited in this section, one may engage in trying to combine the results from CT-Image-based classification and X-Ray based classification. One may even think of developing an ensemble learning algorithm, combining the two models, thereby introducing more information as well as complexity. Keep in mind that the choice of architecture may be the key here. One may even think of developing a unanimous CNN with multiple input gates, to accommodate both types of images as input. Another alternative would be to accommodate bivariate data in the CNN, obviously, after pixel-correction of any one of the two images.

As far as transmission dynamics are concerned, we have already seen several RNNs and Time Series Models which have been developed during the pandemic, trying to estimate the same with great accuracy. We have found that a lot of focus is being spent on vanilla LSTMs, and on consolidating them with various other supervised and unsupervised learning pipelines. Seldom have we seen any set of researcher(s) publish advanced versions of LSTMs, although, it may have happened that advanced LSTMs might not have reaped the desired accuracy. A viable alternative approach to this problem might lie in developing Hybridized Time Series Models and seeing if it provides greater accuracy.

Aside from the Neural Network realm, Milano and Cannataro (2020) plan to extend their study considering the evolution of the communities at greater time intervals to demonstrate a new pattern of regions with respect to COVID-19 data. Shruthi and Ramani (2020) state that future investigations may expand their writing in any event in two different ways. Primarily, inquiries about whether profit by the adaptability of auxiliary vector auto regression models, force hypothetical limitations to recognize effects of vitality and money showcase on agrarian markets. Furthermore, multivariate unpredictability overflow examination which gives space to apply causality in fluctuation tests and drive reaction investigation dependent on multivariate models may give new experiences.

We can see that, not many researchers have focused on panel data modelling when addressing the pandemic itself; however, owing to the increasing availability of data, as time passes, the prospect might be even more attractive to the point that, a lot of statisticians may be working on it right now. Network or Random Graph Models have also been approached by a handful of researchers, but all of them have mentioned explicitly, that lack of data may be a contributing factor to somewhat poor results. Attempting to adopt such methodologies might prove fruitful now, although the reliability of most data sources that are used to generate similarity matrices may be an issue of concern. The standard SIR/SEIR model has been reproduced quite frequently during this pandemic to trace transmission dynamics. But it is also noteworthy, that most papers pertaining to this model have restricted themselves to simply adding intervening events as model enhancements. One might look into combining SEIR models with Latent Variable Models and other advanced models to gain better insight.

In overview, the statistician today, has the massive responsibility to, firstly, seamlessly combine conventional statistical techniques and machine learning and deep learning tools and techniques, and to act as the go-to-entity for validation of statistical as well as ML and DL oriented models/analyses. They must also ensure that modern ML and DL techniques are treated with a strong statistical acumen.

Acknowledgments

The authors thankfully acknowledge the support from the referee and the Chair Editor.

References

Apostolopoulos, I. D. and Mpesiana, T. A. (2020). Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 1-6.

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N. and Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*, 103795.

Bansal, V., Pahwa, G. and Kannan, N. (2020). Cough Classification for COVID-19 based on audio mfcc features using convolutional neural networks. *IEEE International Conference on Computing, Power and Communication Technologies* (*GUCON*), 604-608.

Chatterjee, A., Gerdes, M. W. and Martinez, S. G. (2020). Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death. *Sensors,* **20**, 11-39.

Cheema, T. N., Raja, M. A., Ahmad, I., Naz, S., Ilyas, H. and Shoaib, M. (2020). Intelligent computing with Levenberg-Marquardt artificial neural networks for a nonlinear system of COVID-19 epidemic model for future generation disease control. *The European Physical Journal Plus,* **11**, 1-35.

Faes, C., Abrams, S., Beckhoven, D. V., Meyfroidt, G., Vlieghe, E., Hens, N. and Surveillance, B. C. (2020). Time between symptom onset, hospitalisation and recovery or death: Statistical analysis of Belgian COVID-19 patients. *International Journal of Environmental Research and Public Health,* **17**, 7560.

Govindarajan, S. and Swaminathan, R. (2020). Differentiation of COVID-19 conditions in planar chest radiographs using optimized convolutional neural networks. *Applied Intelligence*, 1-12.

Hamzah, F. A., Lau, C. H., Nazri, H., Ligot, D. V., Lee, G., Tan, C. L., . . . Emmaanue, R. (2020). Corona Tracker: Worldwide COVID-19 outbreak Data Analysis and Prediction. *Bulletin of the World Health Organisation,* **1**, 32.

Haque, K. F., Haque, F. F. and Gandy, L. (2020). Automatic detection of COVID-19 from chest X-ray images with convolutional neural networks. *2020 International Conference on Computing, Electronics and Communications Engineering (iCCECE)*, 125-130.

Hartono, P. (2020). Mixing autoencoder with classifier: Conceptual data visualization. *IEEE Access,* **8**, 105301-105310.

Hartono, P. (2020). Similarity maps and pairwise predictions for transmission dynamics of COVID-19 with neural networks. *Informatics in Medicine Unlocked*, **20**, 100386.

Hartono, P., Hollensen, P. and Trappenberg, T. (2015). Learning-regulated context relevant topographical map.*IEEE Transactions on Neural Networks and Learning Systems,* **26**, 2323-2335.

Hassan, A., Shahin, I. and Alsabek, M. B. (2020). COVID-19 detection system using recurrent neural networks. 2020 *International Conference on Communications, Computing, Cybersecurity, and Informatics* (*CCCI*), 1-5.

Hawas, M. (2020). Generated time-series prediction data of COVID-19's daily infections in Brazil by using recurrent neural networks. *Data in Brief,* **32**, 106175.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 770-778.

Li, H. H. (2020). Statistical and network-based analysis of Italian COVID-19 data: Communities detection and temporal evolution. *International Journal of Environmental Research and Public Health*, 1-62.

Li, H., Li, H., Ding, Z., Hu, Z., Chen, F., Wang, K., . . . Shen, H. (2020). Spatial statistical analysis of Coronavirus Disease 2019 (Covid-19) in China. *Geospatial Health,* **15**, 867-875.

Li, H., Li, H., Ding, Z., Hu, Z., Chen, F., Wang, K., . . . Shen, H. (2020). Statistical and network-based analysis of Italian COVID-19 data: Communities detection and temporal evolution. *International Journal of Environmental Research and Public Health,* **15**, 1-62.

Li, M. D., Arun, N. T., Gidwani, M., Chang, K., Deng, F., Little, B. P., . . . Kalpathy-Cramer, J. (2020). Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence,* **2**, 200079.

Livadiotis, G. (2020). Statistical analysis of the impact of environmental temperature on the exponential growth rate of cases infected by COVID-19. *Plus One,* **15**, 5-26.

Martin, W., Jerome, L., Maja, v. C., Lars, B., Marlon, G., Derek, H., . . . Klaus, K. (2020). Statistical Analysis of Clinical COVID-19 Data: A concise overview of lessons learned, common errors and how to avoid them. *Clinical Epidemiology,* **12**, 925-928.

Milano, M. and Cannataro, M. (2020). Statistical and network-based analysis of italian COVID-19 data: Communities detection and temporal evolution. *International Journal of Environmental Research and Public Health,* **17**, 4182-4194.

Militante, S. V., Dionisio, N. V. and Sibbaluca, B. G. (2020). Pneumonia and COVID-19 detection using convolutional neural networks. 2020 *Third International Conference on Vocational Education and Electrical Engineering (ICVEE)*, 1-6.

Nayak, S. R., Arora, V., Sinha, U. and Poonia, R. C. (2020). A statistical analysis of COVID-19 using Gaussian and probabilistic model. *Journal of Interdisciplinary Mathematics*, 1-15.

Niazkar, H. R. and Niazkar, M. (2020). Application of artificial neural networks to predict the COVID-19 outbreak. *Global health research and policy,* **5**, 1-11.

O'Driscoll, M., Harry, C., Donnelly, C. A., Cori, A. and Dorigatti, I. (2020). A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics with implications for the current COVID-19 pandemic. *medRxiv*, 1-27.

Ozturk, T., Talo, M., Yildirim, E. A., Baloglud, U. B., Yildirim, O. and Acharya, U. R. (2020). Automated detection of Covid-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 103792.

Phucharoen, C., Sangkaew, N. and Stosic, K. (2020). The characteristics of COVID-19 transmission from case to high-risk contact, statistical analysis from contact tracing data. *EClinicalMedicine*, 1-11.

Pirouz, B., Nejad, H. J., Violini, G. and Pirouz, B. (2020). The role of artificial intelligence, MLR and statistical analysis in investigations about the correlation of swab tests and stress on health care systems by COVID-19. *Information, MDPI,* **11**, 454-474.

Shams, M. Y., Elzeki, O. M., Elfattah, M. A., Medhat, T. and Hassanien, A. E. (2020). Why are Generative Adversarial Networks Vital for Deep Neural Networks? A case study on COVID-19 chest X-Ray images. *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach, Springer, Cham*, 147-162.

Shruthi, M. and Ramani, D. (2020). Statistical analysis of the impact of COVID 19 on India commodity markets. *Materials Today: Proceedings*, 1-6.

Singh, D., Kumar, V., Yadav, V. and Kaur, M. (2020). Deep convolutional neural networks based classification model for COVID-19 infected patients using chest X-ray images. *International Journal of Pattern Recognition and Artificial Intelligence*, 1-19.

Sivapalan, P., Ulrik, C. S., Lappere, T. S., Eklöf, J. V., Shaker, S. B., Bødtger, U. C., . . . Biering-Sørensen, T. (2020). Proactive prophylaxis with azithromycin and

hydroxychloroquine in hospitalized patients with COVID-19(proPAC-COVID): a statistical analysis plan. *Trials,* **21**, 1-9.

Sridhar, S., Seetharaman, R., Mootha, S. and Rayan, A. X. (2020). A study on co-occurrence of various lung diseases and COVID-19 by observing chest X-ray Similarity using deep convolutional neural networks. 2020 *Fourth International Conference On Intelligent Computing in Data Sciences (ICDS). IEEE*, 1-8.

Talmoudi, K., Safer, M., Letaief, H., Hchaichi, A., Harizi, C., Dhaouadi, S., . . . Bougat, S. (2020). Estimating transmission dynamics and serial interval of the first wave of COVID-19 infections under different control measures: a statistical analysis in Tunisia from February 29 to May 5, 2020. *BMC Infectious Diseases,* **20,** 1-8.

Toraman, S., Alakus, T. B. and Turkoglu, I. (2020). Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks. *Chaos, Solitons and Fractals,* **140**, 110122.

Wang, P., Lu, J.-a., Jina, Y., Zhu, M., Wang, L. and Chen, S. (2020). Statistical and network analysis of 1212 COVID-19 patients in Henan, China. *International Journal of Infectious Diseases*, 391-398.

Yasar, H. and Ceylan, M. (2020). A new deep learning pipeline to detect Covid-19 on chest X-ray images using local binary pattern, dual-tree complex wavelet transform, and convolutional neural networks. *Applied Intelligence,* **51**, 2740-2763.

Zhao, S., Gao, D., Zhuang, Z., Chong, M. K., Cai, Y., Ran, J., . . . Wang, M. H. (2020). Estimating the serial Interval of the Novel Coronavirus Disease (COVID-19): A Statistical Analysis Using the Public Data in Hong Kong From January 16 to February 15, 2020. *Transboundary and Emerging Diseases*, 1-17.

Zador, A.M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications,* **10**(**1**), 1-7.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference of SSCA, 24-28 February 2021; pp 87-92

Augmented Reality: A Computational Framework Applied to Modeling the Dynamics of Air Pollution

Saumyadipta Pyne[1,2], Ryan Stauffer[3] and Benjamin Kedem[4]

[1]*Health Analytics Network, Pittsburgh, PA, USA.*
[2]*Public Health Dynamics Lab, and Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA.*
[3]*Atmospheric Chemistry and Dynamics Lab, NASA Goddard Space Flight Center, Greenbelt, MD, USA.*
[4]*Department of Mathematics and Institute for Systems Research, University of Maryland, College Park, MD, USA.*

Received: 10 June 2021; Revised: 18 June 2021; Accepted: 21 June 2021

Abstract

In recent years, we have developed a new Augmented Reality (AR) framework to combine real data with computer-generated synthetic samples to "look under the hood", as it were, for gaining insights into rare, dynamic phenomena. Using data fusion and density ratio model, AR allows us to estimate the tail probabilities of exceeding large thresholds that are far beyond the limited range of observations in moderately sized data. Such thresholds represent extreme events such as the drastic change in air pollution levels in Washington DC caused by lockdown due to the COVID-19 pandemic in 2020, as modeled in this study.

Key words: Data fusion; Tail probabilities; Density ratio model; Synthetic data; Air pollution.

1. Introduction

In its February 4, 2017, edition, *The Economist* claimed that "Replacing the real world with a virtual one is a neat trick. Combining the two could be more useful." Combining real data with synthetic data produces augmented reality (AR), which, we believe, opens up new perspectives regarding statistical inference. Indeed, augmentation of observations of the real world with virtual information is transforming engineering, healthcare and AI with emerging powerful technologies such as robotics, Internet of Things, and more recently, Digital Twins (Tao and Qi, 2019).

In a recent article, we advanced the notion of repeated AR in the estimation of very small tail probabilities even from moderately sized samples (Kedem and Pyne, 2021). Our approach, much like the bootstrap, is computationally intensive. However, unlike bootstrap, we look repeatedly outside the sample. Synthesis of a given sample of real world observations repeatedly with computer-generated data is based on repeated out of sample fusion (ROSF, Kedem *et al.* 2019; Zhang, Pyne and Kedem, 2020). This strategy proves to be useful for inference in various surveillance applications in which the available datasets usually have a limited range of observations and a moderate size due to limited storage capacity.

In particular, we are interested to estimate the tail probability p of observations exceeding a given high threshold T. Our repeated AR approach is based on numerous data fusions. We use an iterative method that can generate a large number of upper bounds B_i for

Corresponding Author: Saumyadipta Pyne
Email: spyne@pitt.edu

p. Say, such a method is fast and probabilistic, and the upper bounds exceed p with a 95% chance. Thus, many of these exceed p but many do not. Therefore, there are subsequences of ordered upper bounds which approach p from above and from below. We showed how upper bounds can be produced by repeated fusion of real data with computer-generated samples, where the number of fusions is arbitrarily large, and where the support of the generated data is large enough so that it ranges beyond T. Hence, using the connection between the real and generated data, we have a computational approach to "peek" into the realm above T.

Notably, the repeated AR approach allows us to model many phenomena of sudden yet significant change that are of great interest to researchers, *e.g.*, for predicting stock market crashes, disease outbreaks, and extreme climatic events. On January 12, 2021, it was reported in the *New York Times* that "America's greenhouse gas emissions from energy and industry plummeted more than 10 percent in 2020, reaching their lowest levels in at least three decades as the coronavirus pandemic slammed the brakes on the nation's economy". It pointed out that "transportation, the nation's largest source of greenhouse gases, saw a 14.7 percent decline in emissions in 2020 as millions of people stopped driving to work" due to lockdowns that were implemented in many states of the U.S. over the course of the COVID-19 pandemic.

Nitrogen dioxide (NO_2) is a gaseous pollutant emitted from the burning of fossil fuels at high temperatures primarily by vehicles, and thus, its level is a good indicator of traffic volume at a given area over a given interval of time. According to the American Lung Association, NO_2 causes a range of harmful effects on the lungs, including increased inflammation of the airways, worsened cough and wheezing, reduced lung function, increased asthma attacks, and a greater likelihood of emergency department and hospital admissions. The U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standard (NAAQS), therefore, measures NO_2 as an indicator for the NO_X family of air pollutants.

Given the sharp reduction in traffic after stay-at-home orders were enforced in many areas of the U.S., in this study, we are interested to model the resulting dynamics of air pollution at a given area. At the capital Washington DC, the stay-at-home order came into effect on April 1, 2020. To analyze the differences between the two periods, pre- and post-order, of 3 months on each side, we resort to two methods. First, we test for similarity in the levels of NO_2 in the morning air in the two periods by using their respective probability distributions. This is done by fusion of data from the two periods as described in a previous study (Kedem *et al.*, 2017). Second, we estimate the tail probability of NO_2 level exceeding $T = 100$ parts per billion (ppb) in each of the two periods. This is done by repeated fusion of the data with computer generated samples (Kedem *et al.*, 2019, Kedem and Pyne, 2020).

2. Data and Methods

2.1. Air Pollution Data

The NO_2 emissions data were collected at four monitoring stations of the U.S. Environmental Protection Agency (EPA) in Washington DC area, for the pre- and post-lockdown periods of January-March and April-June, 2020. In this study, we focused on the morning readings, *i.e.*, the hourly surface levels of NO_2 between 6 am and 9 am. For each period, a random sample of size 200 was selected from the data collected at the locations with the (latitude, longitude) coordinates of (38.895572, −76.958072), (38.921847,−77.013178), (38.970092, −77.016715), and (38.89477, −76.953426). Thus, we obtained a NO_2 sample of size 200 from the first period (January 1–March 31), and another sample of size 200 from the second period (April 1–June 30).

2.2. Testing for equidistribution

Let X_0 be a sample of NO_2 observations of size 200 from January-March, following an unknown probability density (pdf) $g(x)$, $x \in (0, \infty)$, and let $G(x)$ denote the corresponding unknown distribution function (CDF). Similarly, let $X_1 \sim g_1, G_1$ be a sample of size 200 from the second period of April-June, following unknown pdf $g_1(x)$ and CDF $G_1(x)$, $x \in (0, \infty)$. We assume the *density ratio model* (Qin and Zhang 1997, Lu 2007)

$$\frac{g_1(x)}{g(x)} = \exp(\alpha_1 + \beta_1' h(x)) \tag{1}$$

where α_1 is a scalar parameter, β_1 is an 2×1 vector parameter, and $h(x) = (x \log x)$. We now combine or fuse the two samples and estimate the parameters in (1) from the combined sample of size of 400. Kernel density estimates of g, g_1 are shown in Figure 1. From the fits in Figure 1 (bottom panel), we see that the estimated g, g_1 are very close to the corresponding histograms, indicating that the choice of "gamma tilt" $h(x) = (x \log x)$ is sensible.

2.3. Estimation of tail probabilities

The basic idea here is to fuse each NO_2 sample with numerous computer-generated "synthetic" samples. This strategy is referred to as *repeated out of sample fusion* (ROSF in Kedem *et al.*, 2019) or *repeated augmented reality* (repeated AR in Kedem and Pyne, 2020).

If $p = P(X > T)$ is a tail probability to be estimated, we generate numerous upper bounds B's for p where most are above p but many are below p. If $B_{(j)}$ are the corresponding order statistics, then there are $B_{(j)}$ which bound p from above and there are $B_{(j)}$ which bound p from below, yet some $B_{(j)}$ fall in a small neighborhood of p. In this paper, we generated 10,000 such B's. The problem is to find $B_{(j)}$ in a small neighborhood of p. This is addressed by an iterative algorithm which produces subsequences of the $B_{(j)}$ sequence which converge to a small neighborhood of p from above and from below as follows:

$$B_{(j^1)} < B_{(j^2)} < B_{(j^3)} < \cdots < B_{(j^n)} < p < B_{(j_m)} < \cdots < B_{(j_3)} < B_{(j_2)} < B_{(j_1)}$$

where $B_{(j^n)}$ and $B_{(j_m)}$ are very close to p.

We now have two relationships. For a sufficiently large number of fusions, say 10,000, there are $B_{(j)}$ which approach p from above and from below, so that there is a $B_{(j)}$ closest to p. This establishes a relationship between $B_{(j)}$ and p.

With $N = 1000$ (see the remark below), another relationship between $B_{(j)}$ and p is obtained from the well known distribution of order statistics,

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p)]^k [1 - F_B(p)]^{N-k} \tag{2}$$

where F_B is the distribution of B_i (not $B_{(j)}$), which can be computed since F_B practically coincides with the empirical distribution of $B_1, ..., B_{10,000}$.

The iterative algorithm consists of the following steps, starting from some j.
Step 1:
From (2) we can get the smallest p_j such that

$$P(B_{(j)} > p_j) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p_j)]^k [1 - F_B(p_j)]^{N-k} \leq 0.95, \tag{3}$$

Step 2:
From Step 1 we get a j corresponding to the smallest p_j. Use this j and go back to Step 1.

Convergence is reached when for some k. we get the same probability values p_{j_k}. For further details of the algorithm, see Kedem and Pyne (2020).

Remark: We get 10,000 upper bounds $B_1, ..., B_{10,000}$ from which F_B is obtained. However, due to computational limitations, in (2) we use $N = 1000$.

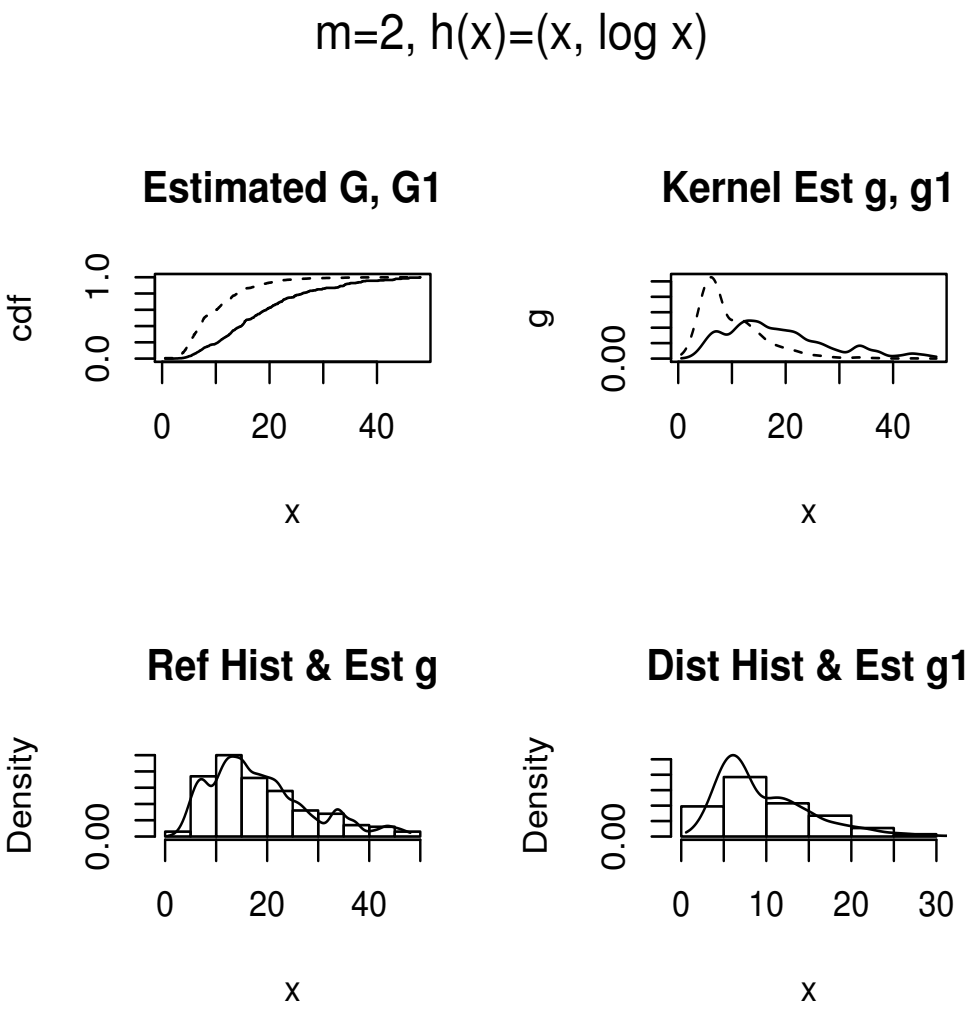


Figure 1: January-March (solid) vs. April-March (dashed) NO_2 distributions

3. Results

The likelihood ratio test of equidistribution $H_0 : \beta_1 = 0$ gives a p-value of 0, indicating that the behavior in the two periods is completely different. This is also seen graphically from the plots of the two CDF's in in Figure 1 (top panel). We see that $\hat{G}$ (solid line) is shifted much to the right relative to $\hat{G}_1$ (dashed line), indicating a great reduction in NO_2 levels in the second period of April-June relative to the first period of January-March.

Indeed a 95% confidence interval for the NO_2 mean in January-March is approximately (14.14,17.24) whereas the same for April-June is approximately (9.11,10.69), again indicating a great reduction in NO_2 levels in latter period.

By setting the threshold $T = 100$, we estimated $p = P(X > 100)$, the probability that NO_2 exceeds a level of $T = 100$, for the two different (pre- and post-shutdown) periods in 2020. Recall that X_0 is a sample of size 200 from the period of January-March. Fusing X_0 10,000 times with generated Uniform(0,180) samples, the algorithm after one iteration gave:

$$B_{(810)} \to 0.00016 \to B_{(808)} \to 0.00016 = \hat{p} = 0.00016 \leftarrow B_{(809)} \leftarrow 0.00016 \leftarrow B_{(813)}$$

Thus, for the period of January-March 2020, we get $\hat{p} = 0.00016$.

Recall that X_1 is a sample of size 200 from the period of April-June. Again, fusing X_1 10,000 times with generated Uniform(0,180) samples, the algorithm gave after eight iterations going down, and a single iteration going up:

$$B_{(500)} \to 6.2e-07 \to B_{(680)} \to 6.2e-07 = \hat{p} = 6.2e-07 \leftarrow B_{(690)} \cdots \leftarrow 2.5e-05 \leftarrow B_{(900)}$$

Thus for the period April-June 2020, we get $\hat{p} = 0.00000062$, which is much smaller than $\hat{p} = 0.00016$ from January-March 2020, echoing the previous results that the NO_2 levels had, in comparison, decreased significantly during April-June 2020 in Washington DC.

4. Discussion

The COVID-19 pandemic has highlighted the need for systematic monitoring and rigorous modeling of dynamic phenomena that can exact a high toll in the form of human suffering and rapid losses in various sectors such as breakdown of supply chains and reduced mobility. Similar lessons are learnt from other areas including extreme climatic events and sudden crashes in the markets. In public health, surveillance is conducted routinely to guard against disease outbreaks and environmental exposures. In this study, we demonstrated how the repeated AR approach could provide a computational framework for modeling the dynamics of air pollution due to traffic emissions during a period of sudden, sharp change.

While estimation of small tail probabilities has long been a topic of research, many of the commonly used methods rely on large number of observations, which makes them less practical for modeling of dynamic phenomena. Methods such as peaks-over-threshold (POT) require observations beyond a threshold, whereas block maxima (BM) require sufficient data such that maxima from each block can be used for estimation. In comparison, both the availability as well as the reliability of computer-generated samples that are representative of real data are increasing with the development of new, powerful computational platforms, e.g., generative adversarial networks (GANs), thus allowing for easier data augmentation.

In this study, we built our AR approach on a density ratio model that starts with a common reference distribution for all sources of information, and then models the individual distributions as *distortions* (*e.g.*, gamma tilt) of that "baseline". Further, we estimate very small tail probabilities even from moderately sized samples. Fusing these repeatedly with computer-generated synthetic data is particularly insightful when the data at hand falls short of the high threshold of interest. Our iterative algorithm constructs a "B-sequence" of bounds that contains a point whose ordinate is very close to the target tail probability, as ensured by the Glivenko-Cantelli theorem. For further details as well as the strengths and the limitations of the repeated AR approach, the reader is referred to Kedem and Pyne (2020) and Kedem *et al.* (2021).

A limitation of the present study is that it does not explicitly account for the fact that the level of NO_2 typically shows a decrease with the advent of spring and summer as it

dissociates in sunlight, and tends to collect less near the surface during that period. While the shift in NO_2 was much larger in 2020 compared to previous years, a multi-year extension of our model would be more insightful. Since the primary aim of this study is to introduce our AR computational framework, we plan to address this in our future work.

References

Kedem, B., De Oliveira, V. and Sverchkov, M. (2017). *Statistical Data Fusion.* World Scientific, Singapore.

Kedem, B., Pan, L., Smith, P. and Wang, C. (2019). Estimation of small tail probabilities by repeated fusion. *Mathematics and Statistics*, **7**, 172–181.

Kedem, B. and Pyne, S. (2020). Estimation of tail probabilities by repeated augmented reality. *Journal of Statistical Theory and Practice*, **15**, 25.

Kedem, B., Stauffer, R., Zhang, X. and Pyne, S. (2021). On the probabilities of environmental extremes. *International Journal of Statistics in Medical Research*, **10**, 72-84.

Lu, G. (2007). *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting.* Ph.D. Dissertation, University of Maryland, College Park.

Qin, J. and Zhang, B. (1997). A Goodness of fit test for logistic regression models based on case-control data. *Biometrika*, **84**, 609–618.

Tao, F. and Qi, Q. (2019). Make more digital twins. *Nature*, **573**, 490–491.

Special Proceedings: ISBN#: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 93-99

Benini Distribution: A Less Known Income-Size Distribution

V.S. Vaidyanathan
Department of Statistics,
Pondicherry University,
Puducherry

Received: 11 June 2021; Revised: 20 June 2021; Accepted: 24 June 2021

Abstract

Size distributions have potential applications in economics and actuarial science. Pareto and log-normal family of distributions are quite popular income-size distributions studied broadly in the literature. However, some less known size distributions are available but are not studied extensively. This paper discusses the properties and applications of one such probability distribution namely, Benini distribution.

Key words: Benini distribution; Maximum likelihood estimation; Lehmann-Scheffe theorem; Sufficient statistic; Claim amount.

AMS Subject Classifications: 60E05, 62F10

1. Introduction

Size distributions have potential applications in modeling income sizes, claim sizes and particle sizes, *etc.* These distributions usually are long- tailed and are skewed. For example, in modeling claim sizes, it is often the case that claims made by insurers are unusually high in very few cases, and therefore a long-tailed skewed distribution having support in the positive real line can be used to model the claim sizes. Pareto, log-normal and Gamma family of distributions are some well known choices in the family of size distributions to model size amounts. The cumulative distribution function (cdf) of a two-parameter Pareto distribution is given by

$$F(x) = 1 - (\frac{x}{x_0})^{-\alpha}, x \geq x_0 > 0, \alpha > 0, \tag{1}$$

where x_0 and α respectively denote the location and shape parameters. The corresponding probability density function (pdf) is

$$f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, x \geq x_0 > 0, \alpha > 0. \tag{2}$$

This distribution was developed by the Italian economist Vilfredo Pareto in the year 1895. Pareto observed a decreasing (downward) linear relationship between the logarithm of income

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

(x) and the logarithm of (T_x), the number of persons receiving income greater than x, $x \geq x_0$. He formulated the relationship as

$$log(T_x) = A - \alpha log(x), x \geq x_0 > 0.$$

Upon normalizing by the number of income receivers, the distribution function of X is obtained as given in (1). Various modified versions of Pareto distributions are available in the literature and are broadly categorized as Type I, Type II, ..., Type V Pareto distributions. A review of Pareto distributions along with its applications, properties and characterizations can be found in Arnold (2015). Soon after the publication by Pareto, economists, both applied and theoretical, started working on this new income-size distribution and found that it reasonably provided a good fit to model many economic phenomena. However, Rodolfo Benini, an Italian statistician and demographer, while modeling the distribution of wealth left as will (legacy), found that assuming a quadratic relationship provided a better fit for modeling legacies instead of a linear relationship as assumed in Pareto distribution. He obtained a new size distribution involving the quadratic term which came to be known as Benini distribution (Benini, 1905). The cdf and pdf of Benini distribution are respectively given by

$$F(x) = 1 - exp\{-\alpha log(\frac{x}{x_0}) - \beta(log(\frac{x}{x_0}))^2\}, x \geq x_0 > 0, \alpha, \beta \geq 0; (\alpha, \beta) \neq (0, 0) \quad (3)$$

and

$$f(x) = exp\{-\alpha log(\frac{x}{x_0}) - \beta(log(\frac{x}{x_0}))^2\} \left[\frac{\alpha}{x} + \frac{2\beta log(\frac{x}{x_0})}{x}\right], x \geq x_0 > 0, \alpha, \beta \geq 0; (\alpha, \beta) \neq (0, 0). \quad (4)$$

A distribution function having the form given in (3) with corresponding density function as given in (4) is known as Benini distribution and denoted as Benini(x_0, α, β). Here, x_0 denote the scale parameter and α, β denote the shape parameters of the distribution. Sphilberg (1977) has independently obtained this distribution in the context of modeling loss amounts due to fire accidents. Though Benini distribution has been established long ago, review of literature reveals that very little research has been carried in studying its properties and characterization. The present article is an attempt to fill this gap by deriving some properties and results of this distribution. In Section 2, some properties of two-parameter Benini distribution are discussed and new results are derived. Parametric Estimation is carried out in Section 3 and an illustration of the same in modeling insurance claim amount data is provided in Section 4. Conclusion of the paper is given in Section 5.

2. Two-Parameter Benini Distribution: Properties and New Resuts

Taking $\alpha = 0$ in (3) and (4), we obtain two-parameter Benini distribution having cdf and pdf respectively as

$$\begin{aligned} F(x) &= 1 - exp\{-\beta(log(\frac{x}{x_0}))^2\} \\ &= 1 - \left(\frac{x}{x_0}\right)^{-\beta(log(x) - log(x_0))}, x \geq x_0 > 0, \beta > 0 \end{aligned} \quad (5)$$

and

$$f(x) = \left[\frac{2\beta log(\frac{x}{x_0})}{x}\right] exp\{-\beta(log(\frac{x}{x_0}))^2\}, x \geq x_0 > 0, \beta > 0. \quad (6)$$

The plot of the density and cumulative distribution functions of the two-parameter Benini distribution for different values of the shape parameter β and $x_0 = 1$ is depicted in Figure 1. From the plots, it can be observed that the density curve becomes more peaked with thin-longer tails for large values of the shape parameter.

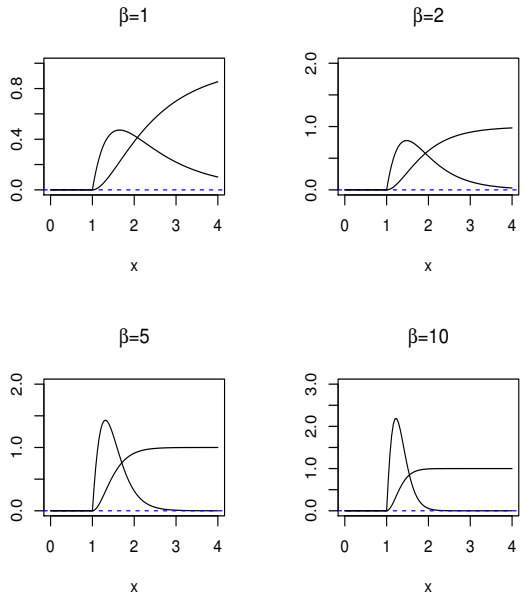


Figure 1: Plot of density and cumulative distribution functions of two-parameter Benini distribution

2.1. Properties

- The quantile function corresponding to (5) is given by

$$F^{-1}(u) = x_0 exp\left\{\sqrt{[\frac{-1}{\beta}]log(1-u)}\right\}, 0 < u < 1. \quad (7)$$

  Given a u, x_0 and β, random samples can be generated from two-parameter Benini distribution by inverse transformation method using (7).

- Equating the cdf given in (5) to 0.5, the median m of two-parameter Benini distribution is obtained as $m = x_0 exp\left\{\sqrt{[\frac{1}{\beta}]log(2)}\right\}$.

- Klieber (2013) has shown that Benini distribution suffers from the problem of moment indeterminacy *i.e.,* moments of all order exists but the distribution is not determined by the moment generating function. This property is true for log-normal distribution also.

- From (6), it can be seen that for fixed x_0, $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ is sufficient statistic for β based on n independent and identically distributed (iid) samples.

- Also, for fixed x_0, Benini distribution belong to one-parameter exponential family of distributions and therefore $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ is a complete sufficient statistic for β.

2.2. New Results

Let the random variable X has two-parameter Benini distribution with cdf and pdf as given in (5) and (6) respectively.

Result 1: Let $Y = log(\frac{X}{x_0})$. Then the distribution of Y is Rayleigh.
Proof: Consider the cdf of Y namely, $F(y)$, which is given by

$$\begin{aligned} F(y) &= P[Y \leq y] \\ &= P[log(\frac{X}{x_0}) \leq y] \\ &= P[X \leq e^y x_0] \\ &= 1 - \left(\frac{e^y x_0}{x_0}\right)^{-\beta(log(e^y x_0) - log(x_0))} \\ &= 1 - e^{-\beta y^2}, y \geq 0. \end{aligned}$$

This shows the distribution of Y is Rayleigh.

Result 2: Let $Y = aX$. Then Y is distributed as two-parameter Benini with parameters ax_0 and β.
Proof: The cdf of Y is

$$\begin{aligned} F(y) &= P[Y \leq y] = P[aX \leq y] \\ &= P[X \leq \frac{y}{a}] \\ &= 1 - \left(\frac{y}{ax_0}\right)^{-\beta(log(\frac{y}{a}) - log(x_0))} \\ &= 1 - \left(\frac{y}{ax_0}\right)^{-\beta(log(y) - log(ax_0))}, y \geq ax_0 > 0. \end{aligned}$$

Comparing with the cdf of Benini distribution given in (5), it is clear that Y is Benini with parameters ax_0 and β.

Result 3: Distribution of the first order statistic.
Proof: The cdf of the first order statistic $X_{(1)} = min\{X_1, X_2, ..., X_n\}$ based on n iid observations from two-parameter Benini distribution is given by

$$\begin{aligned} F(x_{(1)}) &= 1 - [1 - F(x)]^n \\ &= 1 - \left[1 - \left(1 - \left(\frac{x}{x_0}\right)^{-\beta(log(x) - log(x_0))}\right)\right]^n \\ &= 1 - \left(\frac{x}{x_0}\right)^{-n\beta(log(x) - log(x_0))}. \end{aligned}$$

Thus $X_{(1)}$ is distributed as Benini with parameters x_0 and $n\beta$ *i.e.,* the first order statistic is also distributed as Benini.

Result 4: Distribution of the sufficient statistic of the shape parameter β.
Proof: As mentioned in Section 2.1, for fixed x_0, $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ is sufficient for the shape parameter β. From Result 1, the distribution of $Y = log(\frac{X}{x_0})$ is found to be Rayleigh with parameter β. Since the sum of squares of n iid Rayleigh random variables has Gamma distribution, $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ has Gamma distribution with parameters n and β. In other words, the distribution of the sufficient statistic for the shape parameter β is Gamma(n, β).

3. Parametric Estimation

In this section, estimation of the parameters of two-parameter Benini distribution by maximum likelihood (ML) method is discussed. The likelihood function based on n iid samples from two-parameter Benini distribution is given by

$$L(\beta, x_0|x_1, x_2, ..., x_n) = \prod_{i=1}^{n}\left[\frac{2\beta log(\frac{x_i}{x_0})}{x_i}\right] exp\{-\beta(log(\frac{x_i}{x_0}))^2\}, x_i \geq x_0, \beta > 0.$$

Using the likelihood function, it is easy to verify that the ML estimator of x_0 is $\hat{X}_{(0)} = min(X_1, X_2, ..., X_n)$ and that of β is $\hat{\beta} = \frac{n}{\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2}$. The distribution of $\hat{X}_{(0)}$ is Benini as established already in Result 3. To derive the distribution of $\hat{\beta}$, we proceed as follows. Note that for fixed x_0, the distribution of $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ is Gamma with parameters n and β as established in Result 4. Let $\hat{\beta}$ be denoted as $\hat{\beta} = \frac{n}{Y}$, where $Y = \sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$, and Y follow Gamma(n, β). The distribution of $\frac{1}{Y}$ is inverse Gamma with pdf

$$f\left(\frac{1}{y}\right) = \frac{\beta^n}{\Gamma(n)} e^{-\beta(\frac{1}{y})} \left(\frac{1}{y}\right)^{n+1}, y > 0.$$

Hence, the pdf of $Z = \frac{n}{Y}$ is

$$f(z) = \frac{\beta^n}{\Gamma(n+1)} e^{-\beta(\frac{z}{n})} z^{n+1}, z > 0. \tag{8}$$

In other words, (8) is the probability density function of the ML estimator of β. Using (8), the expected value and variance of $\hat{\beta}$ is obtained as

$$E(\hat{\beta}) = \frac{n\beta}{(n-1)} \tag{9}$$

and

$$V(\hat{\beta}) = \frac{n^2\beta^2}{(n-1)}\left[\frac{1}{(n-1)(n-2)}\right]. \tag{10}$$

From (9), it is clear that the $E(\hat{\beta}) \neq \beta$ *i.e.,* the ML estimator $\hat{\beta}$ is not unbiased for β. However, as $n \longrightarrow \infty$ in (9) and (10), $E(\hat{\beta}) \longrightarrow \beta$ and $V(\hat{\beta}) \longrightarrow 0$, implying $\hat{\beta}$ is a

consistent estimator of β. Using (9), an unbiased estimator of β is obtained as

$$T = \frac{(n-1)}{n}\hat{\beta}$$
$$= \frac{(n-1)}{n}\frac{n}{\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2}.$$

Remark 1: For fixed x_0, as mentioned in 2.1, Benini distribution belong to the one-parameter exponential family of densities with complete sufficient statistic $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$. Since T is a function of complete sufficient statistic, by the application of Lehmann-Scheffe theorem, T is a minimum variance unbiased estimator of β. The variance of T is given by $V(T) = \frac{\beta^2}{(n-2)}$.

Remark 2: Since $\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2$ is sufficient for β, uniformly most powerful (UMP) test on β can be derived using the sufficient statistic. Suppose we wish to test the hypothesis $H : \beta \leq \beta_0$ against the alternate $K : \beta > \beta_0$, then the UMP test will reject the hypothesis H for large value of the sufficient statistic. Since for large values of β, the distribution has thin-long tail, the above test can be used to check whether the given data is from a Benini distribution with thin-long tail. The p-value of the test can be computed using Gamma distribution.

4. Illustration

To illustrate the computation of the estimate of the parameters, a real-life data on insurance claim amount due to fire loss is considered in this section. The data set corresponds to Danish reinsurance claim data and consists of 2167 insurance claim amounts (in millions of Danish Krone) during the period from 3rd January 1980 until 31st December 1990. This data

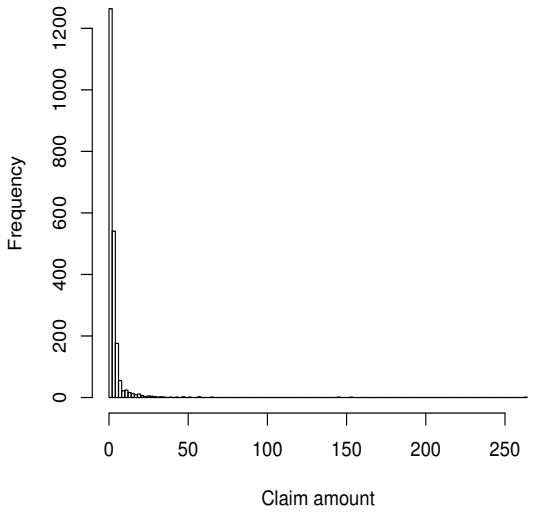


Figure 2: Histogram of claim amount

set has been used by Nadarajah and Bakar (2014) to fit composite models. The summary

statistics of the data are: Minimum value is 1.000, first quartile 1.321, median 1.778, mean 3.385, third quartile 2.967 amd maximum value 263.250. The moment based skewness of claim amount is found to be 18.749 which indicates the data is positively skewed. The histogram of the claim amount is depicted in Figure 2. The histogram reveals that most of the claim amounts are small barring a few which are very large. Thus, it is reasonable to assume a positively skewed probability distribution having long-tail to model the claim amounts. Since Benini distribution is positively skewed and has long-tail, it is chosen to model the insurance claim amounts. As mentioned in the previous section, the ML estimator of the scale parameter x_0 is $\hat{X}_{(0)} = min(X_1, X_2, ..., X_n)$. Since the minimum value of claim amounts is one, $\hat{X}_{(0)} = 1$. The ML estimate of β is obtained using $\hat{\beta} = \frac{n}{\sum_{i=1}^{n}[log(\frac{X_i}{x_0})]^2}$ and is found to be 0.8828. The estimated standard error of $\hat{\beta}$ is 0.0189.

5. Conclusion

In this paper, a less known probability distribution namely, Benini distribution is reviewed and some new properties and distributional results are derived. The results obtained are used to find the distribution of the maximum likelihood estimator of the parameters of the distribution. In addition, minimum variance unbiased estimation of the shape parameter is carried out.

References

Arnold, Barry C. (2015). *Pareto Distributions*. Second Edition. CRC Press, Boca Raton, Florida.

Benini, R. (1905). I diagrammi a scala logaritmica (a proposito della graduazione per valore delle successioni ereditarie in Italia, Francia e Inghilterra). *Giornale degli Economisti*, **Series II, 16**, 222–231.

Klieber, C. (2013). On the moment indeterminacy of the Benini distribution. *Statistical Papers*, **54(4)**, 1121-1130.

Nadarajah, S. and Bakar, S. A. A. (2014). New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal*, **2**, 180-187.

Shpilberg, D. (1977). The probability distribution of fire loss amount. *The Journal of Risk and Insurance*, **44(1)**, 103-115.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 101-111

Innovations in Genomic Selection: Statistical Perspective

Dwijesh Chandra Mishra, Neeraj Budhlakoti, Sayanti Guha Majumdar and Anil Rai
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received: 08 June 2021; Revised: 25 June 2021; Accepted: 29 June 2021

Abstract

Genomic selection is a modified form of Marker Assisted Selection in which the markers from the whole genome is used to estimate Genomic Estimated Breeding Value (GEBV). The population of individuals with both phenotypic and genotypic data is used to estimate model parameters that are subsequently be used to calculate GEBVs of selection candidates having only genotypic data. These GEBVs are then further be used to select the individuals for the purpose of advancement in the breeding cycle. Several estimators are available to estimate GEBV. However, various issues such as high dimensionality of the data, multicollinearity among the markers, a smaller number of individuals and a greater number of markers (large *p* and small *n* problem) are the major challenges in estimation of GEBVs. This paper discusses most commonly used methods for estimation of GEBVs, *viz*., Ridge Regression, Genomic Best Linear Unbiased Prediction (GBLUP), Bayesian Alphabets and Least Absolute Shrinkage and Selection Operator (LASSO) with the aim to meet the challenges associated with estimation of GEBVs. Apart from this, some semi and non-parametric methods of genomic selection have been discussed as well. Moreover, another problem like presence of outliers in the data of genomic selection has also been conversed. Furthermore, a case study deals with non-linearity of the data has also been presented and illustrated using multi traits data. At the end, some future directives of research in this area are highlighted.

Key words: Genomic estimated breeding values; Statistical models; Genomic best linear unbiased prediction (GBLUP); Bayesian methods; Least absolute shrinkage and selection operator (LASSO).

1. Introduction

Right from the beginning of agriculture, farmers used to select the best plants on the basis of their phenotypic characters such as higher yields, larger seeds, or sweeter fruits for the purpose of growing them for next season. In this way, they tried to alter the genetic makeup of plants. Afterward, farmers came to know about artificial mating of the plants by cross pollination, from which breeding approach emerged out. Breeding approach is basically a process by which humans use animals and plants to selectively develop particular phenotypic trait (characteristics) by choosing or selecting best males and females of animal or plant which can sexually reproduce and have offspring together. Breeding approach was more advanced than traditional approach. However, genetic gain through this technique was found to be very low, time-consuming and inefficient, especially when; the traits under consideration have low genetic variance (low heritability), traits are limited to particular sex (sex-limited traits) and when generation interval is large or traits appear late in the life. Other

Corresponding Author: Dwijesh Mishra
Email: dwij.mishra@gmail.com

limitation of this approach was that one does not know about the genetic basis of the transmission of traits from parents to their offspring consequently which causes the problem of Linkage drag (*i.e.*, Transfer of genes governing undesirable trait along with the gene of interest).

Later on demerits of breeding approach was overcome by Marker Assisted Selection (MAS) based breeding. Where molecular markers associated with the trait of interest are used to select the superior plant for breeding purpose. It is a simpler method as compared to phenotypic screening used in traditional breeding, especially for the traits which require laborious screening. Through this approach, time and resources may be saved as selection can be done even at seedling stage for the important trait like grain quality which appears late in their life cycle. It also enhances reliability as there is no environmental effects play a role in the selection as in the case of traditional breeding. According to one study, 8-38% extra genetic gain can be observed when marker information's are included in Best Linear Unbiased Prediction (BLUP) methodology for the prediction of breeding value (Meuwissen *et al*., 1996). However, there are certain limitations of this approach such as Marker should be tightly linked to the trait of interest, it cannot be used in polygenic trait or Quantitative trait, where multiple minor genes play a role in governing the particular trait. But it is well known fact that most traits of economic interest in agriculture are quantitative in nature. Another problem related to this approach is that every marker needs to be statistically tested with respect to their association with the trait of interest which causes Multiple Hypothesis Testing Problem.

In order to overcome the limitations of MAS, Meuwissen et al. (2001) proposed a variant of MAS that is known as Genomic Selection (GS). It is a form of marker-assisted selection in which genetic markers covering the whole genome are used to identify quantitative trait loci (QTL) which are in linkage disequilibrium (LD) with at least one marker. Genomic Selection has been successful and the main reason behind the success is that it incorporates all markers information in the prediction model. In this approach, a prediction equation on training population containing phenotypic as well as genotypic data is generated and subsequent prediction of the breeding values of the individuals (testing population) having only genotypic data is carried out. Breeding value calculated by prediction equation is termed as Genomic Estimated Breeding Value (GEBV) and depending on the outcomes of GEBV, the selection decision is made on the breeding population.

This approach is better than the above-mentioned approaches as it offers more accurate prediction of Genomic Wide Estimated Breeding Value (GW-EBV) than MAS, consequently high accuracy in breeding values with respect to desired trait. Using this approach, there is a drastic reduction of breeding interval than traditional breeding, faster genetic gains (more than 30% reported in animals) and long-term low cost of breeding which ultimately enhances the production and productivity which ensure the food and nutritional security. However, there are certain factors such as training population size, trait heritability, influence of Genotype-Environment ($G \times E$) interaction, marker density, effective population size of breeding population, (Genetic diversity of breeding population), genetic relationship between training population and selection candidates influences the accuracy of the prediction of GEBVs. Apart from this there are certain statistical issues or challenges such as large number of markers or predictors (p) as compared to small number of observations or samples (n), multi-collinearity where markers are related with each other, presence of outliers/missing data, exist which should be taken care of.

In this paper, a brief review of the models used in the prediction of Genomic Estimated Breeding Values (GEBVs) has been presented. Starting with the very simple linear model to advance nonlinear/nonparametric models with the aim to meet various statistical challenges and issues arises during prediction of GEBVs have been briefly discussed in this paper. Apart from this, a non-linear model called Multivariate Kernelized Least Absolute Shrinkage and Selection Operator (Multivariate Kernelized LASSO) has been suggested as a case study which takes care the problem of non-linearity as well as pleiotropy present in the data of genomic selection. As a concluding remark, some potential future research prospective in this area have been highlighted.

2. Methods of Genomic Selection

Predicting GEBVs on which selection of the suitable individuals is done in genomic selection starts with simple linear model.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = n \times 1$ vector of observations; $\boldsymbol{\beta} = p \times 1$ vector of marker effects; $\boldsymbol{\varepsilon} = n \times 1$ vector of random residual effects; $\boldsymbol{X} =$ design matrix of order $n \times p$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_e^2\mathbf{I})$.

One major problem in linear model is that number of markers exceed the number of observations (large p and small n problem ($p >> n$)and this creates a problem in parameter estimation. Subset of the significant markers can be an alternative for dealing with large $p$and small n problem.

Meuwissen *et al.* (2001) used a modification of least squares regression for GS. Performed least squares regression analysis on each maker separately with following model

$$\boldsymbol{Y} = \boldsymbol{X}j\beta j + \boldsymbol{\varepsilon}$$

where,

$\boldsymbol{X}j = j^{\text{th}}$ column of the design matrix
βj = genetic effect of j^{th} marker

Markers with significant effects are selected by plotting the log likelihood of this model against the position of the marker. The marker with significant effects (QTL) are further used for estimation of breeding value

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where,

$\boldsymbol{X}$ = the design matrix of order $n \times q$ $(q << p)$
$\boldsymbol{\beta}$ = vector of genetic effect of order $q \times 1$
q = number of significant markers

It is very likely that multicollinearity exist among markers (explanatory variables) and this can negatively affects the performance of variable selection methods. This problem is solved by using ridge regression (Meuwissen *et al*., 2001). Here, the goal is to derive an estimator of β with smaller variance than the least square estimator. Due to tradeoff between variance and bias of an estimator, there is a price to pay as ridge regression estimator of $\boldsymbol{\beta}$ is biased. In Ridge Regression (RR), penalty function is added to normal equation. So instead of minimizing sum of square of residuals, it minimizes:

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

where λ is the penalty parameter and can be chosen by variety of ways, one solution is given by (Hoerl *et al*., 1975)

$$\lambda = \frac{ps^2}{(\widehat{\boldsymbol{\beta}})'(\widehat{\boldsymbol{\beta}})}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

where, p = number of markers, s^2 = estimate of error variance ($i.e.\hat{\sigma}^2$).

Ruppert *et al*. (2003) showed that ridge regression is a special case of the Best Linear Unbiased Prediction (BLUP) where mixed linear model is implemented. Restricted Maximum Likelihood Estimation (REML) is a good choice for finding a realistic value for the penalty parameter and estimating the variance component. Here objective is to minimize the function-

$$(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{\beta}'\boldsymbol{G}^{-1}\boldsymbol{\beta}$$

where $E(\boldsymbol{u}) = \boldsymbol{0}$ and $E(\boldsymbol{e}) = \boldsymbol{0}$ and $var\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\sigma^2$, $\boldsymbol{G}$ and $\boldsymbol{R}$ are known positive definite matrices and σ^2 is a positive constant.

Similar to RR, Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani1996; Usai *et al*., 2009) is other variant of penalized regression which can be obtained by altering the penalty function (*i.e*. by giving linear penalty). Objective function of this variant is defined as-

$$(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda|\boldsymbol{\beta}'\boldsymbol{1}|$$

This constraint shrinks some of the marker effects and sets some of them to zero.

It may be possible that not all markers have equal variance. Therefore, variance of marker positions needs to be modeled. For this purpose, the Bayesian approach has been used. In this approach, it is assumed that there is prior distribution of marker effect. Where, inferences about model parameters are obtained on the basis of posterior distribution. Several variants of Bayes such as Bayes A, Bayes B, Bayes Cπ and Bayes Dπ were proposed for estimation of GEBVs (Meuwissen *et al*., 2001 and Habier *et al*., 2011).

The Bayes A approach applies the same prior distribution for all of the variances of the marker positions whereas Bayes B assumes that not all markers contribute to the genetic variation. In Bayes A approach, inverse chi-squared probability distribution $\chi^{-2}(\vartheta, S^2)$ can be used as the prior distribution. It is a conjugate prior as the posterior distribution is also an inverted chi-square distribution $\chi^{-2}(\vartheta + n_j, S^2 + \acute{\beta}_j\beta_j)$ where n_j is the number of haplotype effects at marker position.

The Bayes B approach has a prior density on the variance that is a mixture. It has a high probability mass at $\sigma_{\beta j}$ = 0, it can be summarized *as* $\sigma_{\beta j}$ = 0 with prob = π and $\sigma_{\beta j} \sim \chi^{-2}(\vartheta, S)$ with prob = (1 – π).The choice of degrees of freedom and the scale

parameters of the scaled inverse chi-square distribution can influence the outcome (Gianola *et al*., 2009).

Improved Bayesian methods were developed by Habier *et al*. (2011). Bayes Cπ and Dπ are the modification of Bayes A and Bayes B where the probability π of having a zero effect SNP is estimated.

The presence of outliers in genomic as well as phenotypic data is a common phenomenon. The presence of outliers may distort the distribution and adversely affect the accuracy of genomic prediction. Presence of outliers in genomic data increases the computational time and when the size of outliers increases, the sample size increases and consequently genetic variance decreases. Rajaratnam *et al.* (2019) recently proposed an approach for detection of influential observation based on LASSO technique. They proposed four different measures *i.e*., df-model- it measures the change in model selected; df-lambda: it measures the change in tuning parameter λ, df-regpath: it measures the changes observed in LASSO regularization path and df-cvpath: it observes changes in LASSO cross-validation path.

Df-Model: This measures the changes in model selection through LASSO when observation is discarded. To quantify this change, df-model for i^{th} observation can be defined as:

$$wdf - model(i) = \frac{\delta(i) - E\{\delta(i)\}}{\sqrt{var\{\delta(i)\}}}$$

where $\delta(i) = \sum_{j=1}^{p} \left| I\{\hat{\beta_j}^{lasso} = 0\} - I\{\hat{\beta_j}^{lasso}(i) = 0\} \right|$ is termed as model difference and it simply measures the difference in the no. of selected variable for full model vs. when i^{th} observation is dropped.

Df-lambda: It measure the changes observed in regularization parameter λ for full LASSO model *vs*. when i^{th} observation is dropped. Measuring this change is important because this parameter tells that at what extent selected LASSO model is shrinking the estimates. To quantify this change, df-lambda for i^{th} observation can be defined as

$$df - lambda(i) = \frac{\hat{\lambda} - \hat{\lambda}(i) - E\{\hat{\lambda} - \hat{\lambda}(i)\}}{\sqrt{var\{\hat{\lambda} - \hat{\lambda}(i)\}}}$$

This involves fitting LASSO, $n + 1$ times then computes difference of $\widehat{\lambda} - \hat{\lambda}(i)$. $E\{\hat{\lambda} - \hat{\lambda}(i)\}$ and $var\{\hat{\lambda} - \hat{\lambda}(i)\}$ can be simply estimated using sample mean and variance of n observed value of $\widehat{\lambda} - \hat{\lambda}(i)$, cut-off for df-lambda is justified at ± 2.

Df-Regpath: It measures the deviation in the LASSO regularization when an observation is dropped from LASSO path. If a significant deviation occurs from LASSO original path it means dropped observation could have huge impact on LASSO estimates which further may affect the conclusion and interpretation for LASSO solution. Df-Regpath for i^{th} observation could be defined as

$$df - regpath(i) = \frac{\Delta_1 \hat{\beta}^{lasso}(i) - E\{\Delta_1 \hat{\beta}^{lasso}(i)\}}{\sqrt{var\{\Delta_1 \hat{\beta}^{lasso}(i)\}}}$$

where $\Delta_1\hat{\beta}^{lasso}(i) = \int_l^u \left\|\hat{\beta}^{lasso}(s) - \hat{\beta}^{lasso}(s,i)\right\| ds$ and l, u are specified interval [l, u] defines possible λ values.$\hat{\beta}^{lasso}(s)$ represents vectors of parameter estimates obtained at λ = s using LASSO with full model, whereas$\hat{\beta}^{lasso}(s,i)$ represents vector of parameter estimates obtained at $\lambda = s$ by excluding i^{th} observation from the model.

Df-Cvpath: It measures the changes in predictive performance of LASSO when an observation is dropped from LASSO path. Quantifying this is crucial as if large change in predictive performance of LASSO, suggests that it has infrequent response hence observation has huge impact on LASSO solution. It generates a cross-validation error curve $\gamma(s)$ which gives estimate of prediction error on test data after LASSO is trained on data for range of values for regularization parameter λ. df-cvpath for i^{th} observation could be defined as

$$df - cvpath(i) = \frac{\Delta_\gamma(i) - E\{\Delta_\gamma(i)\}}{\sqrt{var\{\Delta_\gamma(i)\}}}$$

where $\Delta_\gamma(i) = \int_l^u \left|\gamma(s) - \gamma(s,\, i)\right| ds$ and $\gamma(s,\, i)$ represents the cross-validation error when i^{th} observation is dropped from path.

These measures detect outlier from high dimensional genomic data based on LASSO regression. It can be observed that all these measures *i.e.*, df-model, df-lambda, df-regpath and df-cvpath detect influential observations which affect model directly or indirectly, have difference in their results regarding detection of influential observations as they are used for optimizing different parameters (Budhlakoti *et al.*, 2020a). To overcome these limitations, Budhlakoti *et al.* (2020b) have proposed a more robust measure for detection of influential observation by integrating above discussed measures using *p*-values based meta-analysis approach (Figure 1). It has been observed that this method significantly improves the prediction accuracy of genomic selection in the presence of outliers in the data.

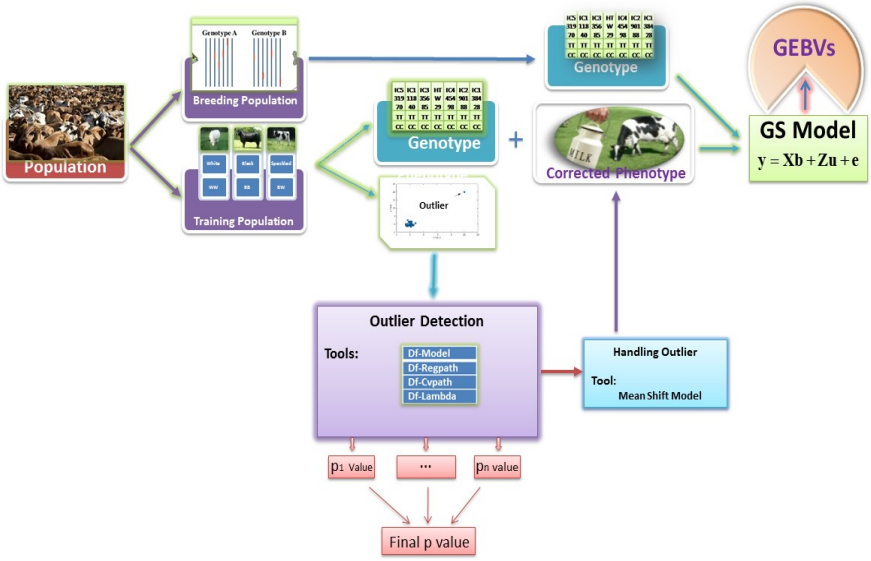


Figure 1: Workflow of the method developed by Budhlakoti *et al.* (2020) for detection of outlier in genomic selection

Genomic selection based on single trait (STGS) has been utilized successfully in recent years. But it is unable to perform well in the case of pleiotropy *i.e*., one gene links with multiple traits. A mutation in a pleiotropic gene may influence several traits simultaneously. It was found that traits with low heritability can borrow information from correlated traits and consequently achieve higher prediction accuracy. So Multi Trait Genomic Selection (MTGS) gave more accurate GEBVs than STGS for the trait with low heritability and for the trait having missing data. Jia *et al*. (2012) presented three multivariate linear models (*i.e*., GBLUP, Bayes A, and Bayes Cπ) and compared them to univariate models and a detailed comparison of various STGS and MTGS based methods also been deliberated by Budhlakoti *et al*. (2019). Moreover, the models, we generally use for GS are linear. But this assumption is generally violated. So nonlinear multi-trait-based approach may be more accurate for genomic selection.

Multivariate LASSO

This is an extension of simple LASSO model. Here the sharing involves which variables are selected, since when a variable is selected, a coefficient is fit for each response. Statistical formulation in this case is same as LASSO with some minor differences. It can be written in the form of simple statistical model as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

Here all notations are as such in LASSO. Only difference is that Y is a matrix of responses instead of vector earlier. It minimizes following objective function:

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda|\boldsymbol{\beta}'\mathbf{1}|$$

Kernelized Multivariate LASSO

To take advantage of higher dimensional feature spaces, we can introduce the data via nonlinear functions. For example, we can replace the inner product of the data by a kernel function. $k(\boldsymbol{X_i}, \boldsymbol{X_j}) = (\Phi(\boldsymbol{X_i})'\Phi(\boldsymbol{X_j}))$. Here we can apply so-called "kernel trick"; *i.e.* the fact that $\Phi(\boldsymbol{X_i})'\Phi(\boldsymbol{X_j}) = k(\boldsymbol{X_i}, \boldsymbol{X_j})$, we can see that $\Phi\Phi'$represents the $(n \times n)$ *Kernel Gram Matrix* $\boldsymbol{K}$ of the cross dot products between all mapped input data points $\{\Phi(\boldsymbol{X_i})\}_{i=1}^{n}$.

Some commonly used choice of kernel functions include: the Gaussian radial basis function $k(\mathbf{x}, \mathbf{z}) = exp(-\sigma\|\boldsymbol{x} - \boldsymbol{z}\|^2)$, where σ is the bandwidth parameter, the Laplace radial basis function $k(\mathbf{x}, \mathbf{z}) = exp(-\sigma\|\boldsymbol{x} - \boldsymbol{z}\|)$.

Here, we have suggested kernelized Multivariate LASSO for estimation of GEBVs. For illustration of MTGS based methods we have considered Brasica napus dataset (Kole *et al*., 2002). Dataset has 4 responses for 103 lines (individuals) genotyped for 300 markers. Lines are derived from two cultivars namely Stellar and Major. Marker genotypes are represented in 0/1, where 0 represent a Stellar allele and 1 represent Major allele. First, we applied LASSO (Multiresponse) technique for MTGS to improve the GEBVs. We have observed reasonable accuracy gain for predicted breeding value *i.e*., GEBVs for various traits under study. Then to capture nonlinearity component in populations we have also used Kernelized LASSO (Multiresponse). Very good level of accuracy has been observed for most of the traits under study.

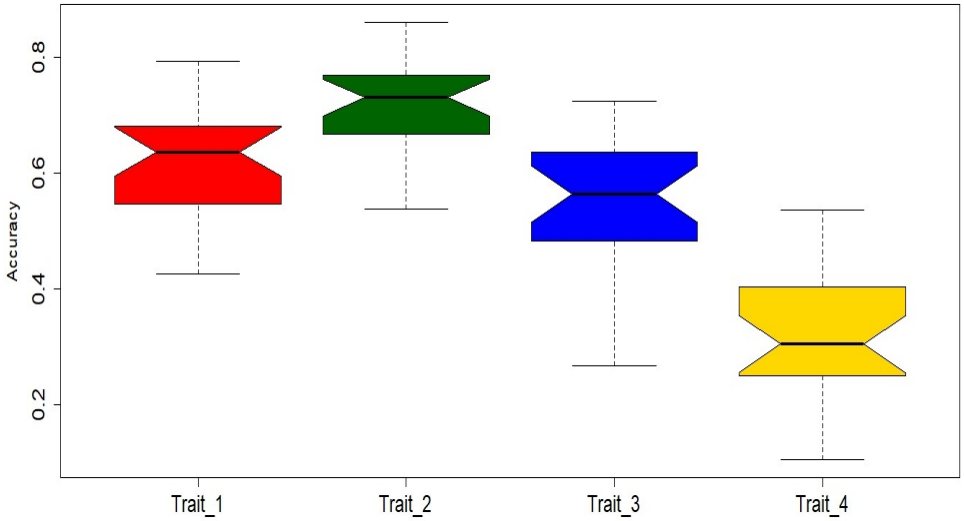


Figure 2: Performance Accuracy of the method of Multivariate Kernelized LASSO in four datasets of Brassica species

These all above discussed methods performs satisfactory well only in case of additive genetic architecture *i.e*., partitioning of genetic variance into additive, dominance, additive × additive, additive × dominance, etc. But it only holds under conditions of linkage equilibrium, random mating of male and female parents, no inbreeding, no assortative mating, no natural or artificial selection and no genotyping errors. In breeding programs, these conditions are all violated. Epistatic interaction may play a crucial role for explaining genetic variation for quantitative traits, as ignoring this kind of interaction in the model may end up with lower genomic prediction accuracy (Cooper *et al*. 2002). Gianola *et al*. (2006) first used non-parametric and semi-parametric methods for modeling complex genetic architecture, as they also include such type of higher order interaction in these models. Subsequently, several statistical methods were implemented to model both main and epistasis effects for genomic selection (Cai *et al*., 2011, Xu, 2007). Recently, some semi-parametric (Legarra *et al*., 2018) and other robust approaches (Tanaka 2020; Majumdar *et al*., 2019a; Majumdar *et al*., 2019b; Budhlakoti *et al*., 2020a; Budhlakoti *et al*., 2020b; Sehgal *et al*., 2020) have also been proposed and implemented in genomic selection.

Gianola *et al*. (2006) proposed non-parametric and semi-parametric methods to model the relationship between the phenotype and the markers that are available within the GS framework.

Nadaraya-Watson estimator

$$Y_i = g(\boldsymbol{X_i}) + e_i$$

where,
Y_i phenotypic measurement on individual i, $i = 1,2,\dots,n$,
$\boldsymbol{X_i}$ is a $p \times 1$ vector of dummy SNP covariates observed on individual i,
$g(.)$ is some unknown function relating genotypes to phenotypes, $g(\boldsymbol{X_i}) = E(Y_i|\boldsymbol{X_i})$
where, $E(Y_i|\boldsymbol{X_i})$ is a conditional expectation of Y_i relative to $\boldsymbol{X_i}$
e_i is a residual effect for i^{th} individual and $e_i \sim (0, \sigma^2)$.

The conditional expectation can be written as

$$g(X) = \frac{\int Yp(\boldsymbol{X}, \boldsymbol{Y})dY}{p(\boldsymbol{X})}$$

Reproducing Kernal Hilbert Space

In this semi-parametric kernel mixed model approach, features of a nonparametric model are combined with a mixed model framework by *Gianola et al.* (2006). The model can be written as:

$$Y_i = \boldsymbol{w'}_i\boldsymbol{\beta} + \boldsymbol{z'}_i\boldsymbol{u} + g(\boldsymbol{X_i}) + e_i$$

where $i = 1,2, \ldots, n$ and β is a vector of fixed unknown effects (*e.g.*, physical location of an individual), $\boldsymbol{u}$ is a $q \times 1$ vector which represents additive genetic effects, $\boldsymbol{w_i}$ and $\boldsymbol{z_i}$ are known incidence vectors, $g(\boldsymbol{X}_i)$ is an unknown function of the SNP data and the vector of residuals, an $\boldsymbol{e}$ is assumed to have a $N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ distribution.

Overall summary of the methods used in genomic selection is provided in the **Figure 3:**

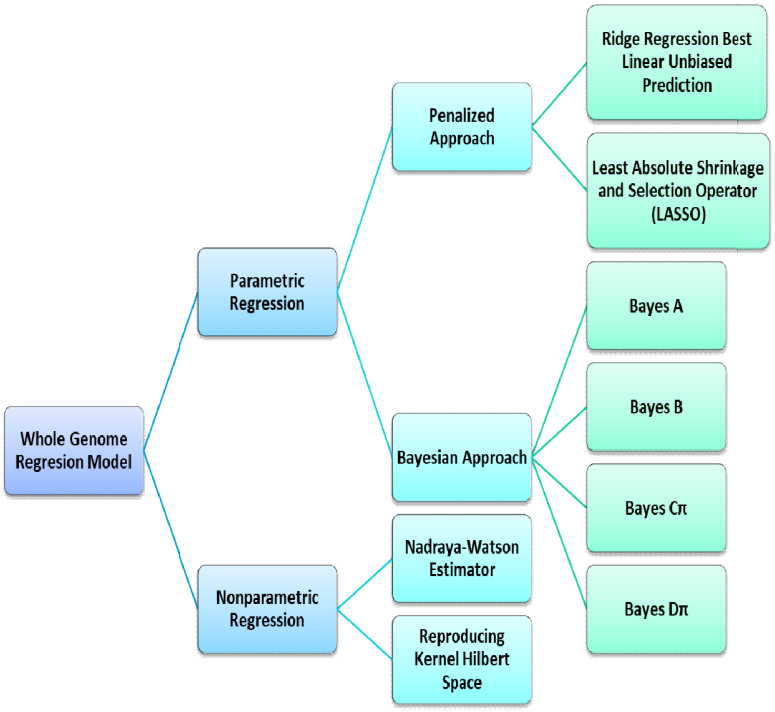


Figure 3:Overall summary of the methods used in Genomic Selection

A detailed comparison of various non-parametric methods for genomic selection at different combination of population size and heritability using simulated data was presented by Budhlakoti *et al.* (2020c). For detailed comparison and review of various genomic selection methods one can also refer to Howard *et al.* (2014).

3. Future Research Direction in This Area

Above mentioned methods of genomic selection mainly deals with genetic architecture containing additive effects, dominance effects and somewhat epistatic effects. But in real

situation, more degree of epistatic effects are present. We need to develop some more advanced statistical models or methods which could efficiently deal with epistatic effects present in the data. Moreover, since genomic selection models are based on genotypic as well as phenotypic data, therefore there is also the possibility of environmental effects and their interaction with the genetic component. Therefore, it is imperative to develop a model for genomic selection by incorporating the environmental effects and their interaction with the genotypic effects. Furthermore, genotypic data generated for genomic selection are having lot of missing data. Therefore, it is also required to develop a method which could take care of incomplete data situation. Apart from this, role of epigenetics in genomic selection can also be possible. New research initiative should be taken where epigenetic effects for genomic selection can be modelled.

References

Budhlakoti, N., Mishra, D. C., Rai, A., Lal, S. B., Chaturvedi, K. K. and Kumar, R. R. (2019). A comparative study of single-trait and multi-trait genomic selection. *Journal of Computational Biology*, **26**(**10**), 1100-1112.

Budhlakoti, N., Rai, A. and Mishra, D. C. (2020a). Effect of influential observation in genomic prediction using LASSO diagnostic. *Indian Journal of Agricultural Sciences,***90**(**6**), 1155-1159.

Budhlakoti, N., Rai, A. and Mishra, D. C. (2020b). Statistical approach for improving genomic prediction accuracy through efficient diagnostic measure of influential observation. *Scientific Reports*, **10**(**1**), 1-11.

Budhlakoti, N., Rai, A., Mishra, D. C., Jaggi, S., Kumar, M. and Rao, A. R. (2020c). Comparative study of different non-parametric genomic selection methods under diverse genetic architecture. *Indian Journal of Genetics*, **80**(**4**), 395-401.

Cai, X., Huang, A. and Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC Bioinformatics*, **12**(**1**), 1-13.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E. and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**(**1**), 347-363.

Gianola, D., Fernando, R. L. and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**(**3**), 1761-1776.

Habier, D., Fernando, R. L., Kizilkaya, K. and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**(**1**), 1-12.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: some simulation. *Communications in Statistics*, **4**, 105–123.

Howard, R., Carriquiry, A. L. and Beavis, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics*, **4**(**6**), 1027-1046.

Jia, Y. and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, **192**(**4**), 1513-1522.

Kole, C., Thormann, C. E., Karlsson, B. H., Palta, J. P., Gaffney, P., Yandell, B. and Osborn, T. C. (2002). Comparative mapping of loci controlling winter survival and related traits in oilseed Brassica rapa and B. napus. *Molecular Breeding*, **9**(**3**), 201-210.

Legarra, A. and Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, **50**(**1**), 1-18.

Majumdar, S. G., Mishra, D. C. and Rai, A. (2020a). Effect of genotype imputation on integrated model for genomic selection. *Journal of Crop and Weed*, **16**(**1**), 133-137.

Majumdar, S. G., Rai, A. and Mishra, D. C. (2020b). Integrated framework for selection of additive and nonadditive genetic markers for genomic selection. *Journal of Computational Biology*, **27**(**6**), 845-855.

Meuwissen T. H. and Goddard M. E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution*, **28**, 161-176.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**(**4**), 1819-1829.

Rajaratnam, B., Roberts, S., Sparks, D. and Yu, H. (2019). Influence diagnostics for high-dimensional lasso regression. *Journal of Computational and Graphical Statistics*, **28**(**4**), 877-890.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. 12, Cambridge University Press.

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J. and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Frontiers in Plant Science*, **11**, 197.

Tanaka, E. (2020). Simple outlier detection for a multi-environmental field trial. *Biometrics*, **76(4)**, 1374-1382.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(**1**), 267-288.

Usai, M. G., Goddard, M. E. and Hayes, B. J. (2009). LASSO with cross-validation for genomic selection. *Genetics Research*, **91**(**6**), 427-436.

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, **63**(**2**), 513-521.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 113–123

Inferential Procedures to Compare Parallel, Superior and Crossover Multivariate ROC Curves

Vishnu Vardhan Rudravaram
Department of Statistics, Ramanujan School of Mathematical Sciences
Pondicherry University, Puducherry

Received: 13 June 2021; Revised: 22 July 2021; Accepted: 24 July 2021

Abstract

Receiver Operating Characteristic (ROC) curve is widely used and accepted tool to assess the performance of a classifier or procedure. Along with this, comparing the diagnostic test procedures or ROC curves is also of major concern and interest. A multivariate extension of ROC (MROC) curve considers a linear combination of several markers for classification. In this work, some inferential procedures are given to compare MROC curves that are parallel, superior and crossover using the scores of MROC curve and also using mean vectors and dispersion matrices. Further, a modified version of AUC (mAUC) under MROC setup is proposed to address the case of crossover MROC curves. It is also shown that mAUC performs better than AUC. The performance of mAUC in the aspect of crossover curves is supported by a real dataset and simulation studies at different sample sizes.

Key words: MROC Curve; Modified AUC and Crossover curves

1. Introduction

In theory of Statistics, there are several tools and techniques available and being developed to address various practical issues in diversified areas. One such prominent area is *Classification Scenario*. The term *"Classification"* indicates the method of allocating or assigning a group of objects/individuals into one of the predefined classes or populations. The prominent fields of research where the logical thinking and analytical processing of classification techniques can flourish are *Diagnostic Medicine, Life Sciences, Experimental Psychology* etcetera.

In general, there are two major objectives in classification problems: the first one is to define a *classifier rule* and the second is to determine an *optimal cutoff*. These two objectives are to be met in such a way that it should minimize the rate of misclassification. The classifier rule will help in generating a decision matrix (usually referred as *confusion matrix*) with probabilities of correct and incorrect classifications. The techniques available to handle such classification problems are Logistic Regression, Discriminant Analysis and Receiver Operating Characteristic (ROC) curve analysis. All these techniques are branched from the hub of *Statistical Decision Theory* (SDT). The first two techniques meet the above mentioned criteria of obtaining a classifier rule and optimal cutoff. In addition to the mentioned

Corresponding Author: Vishnu Vardhan Rudravaram Email: vrstatsguru@gmail.com

objectives, ROC curve has yet another feature of providing the accuracy of a classifier.

ROC curve took its origin during World War II and was first used in Signal Detection Theory for analyzing radar images. This technique has its applications in wide variety of fields such as Medicine, Experimental Psychology, Banking, Finance and many more. However, the promising area for the theoretical development of ROC curve is *Diagnostic Medicine*. Apart from providing a classifier rule and optimal threshold another important advantage of ROC curve is in assessing the performance of diagnostic test and in choosing a better one when there are two tests for a particular scenario. Over the years, application of ROC curve analysis has been observed in many fields and a few to mention are Experimental Psychology, Diagnostic Medicine and Radiology (Krzanowski and Hand, 2009), Machine learning (Provost *et al.*, 1998).

Let there exist two populations denoted by '0' (normal or healthy) and '1' (abnormal or diseased) where 'c' be the cutoff. The individual/object is said to belong to population '1' if the score 'S' is greater than c otherwise belongs to population '0'. Four probabilities and their associated classification rates can be defined as

- The probability that an individual/object from '1' is correctly classified as '1'

$$\text{i.e, True Positive Rate } TPR = P(S > c|1)$$

- The probability that an individual/object from '0' is misclassified as '1'

$$\text{i.e., False Positive Rate } FPR = P(S > c|0)$$

- The probability that an individual/object from '0' is correctly classified as '0'

$$\text{i.e., True Negative Rate } TNR = P(S \leq c|0)$$

- The probability that an individual/object from '1' is misclassified as '0'

$$\text{i.e., False Negative Rate } FNR = P(S \leq c|1)$$

The ROC curve underpins an unknown monotonic transformation and it can be defined as the tradeoff between two intrinsic measures namely *1- specificity (FPR)* and *sensitivity (TPR)*. It is a unit square plot ranging from (0, 0) to (1, 1) and a line connecting these points is called the chance diagonal. Sensitivity is the probability that the test result is positive when the condition is present and Specificity is the probability that the test result is negative when the condition is absent.

$$\text{Sensitivity}(S_n) = P(S > c|1); \;\; \text{Specificity}(S_p) = P(S \leq c|0) \tag{1}$$

The typical forms of ROC curve are depicted in Figure 1. The figure constitutes of three cases of ROC curves: best, moderate and worst case. Each case indicates the extent of classification that can be performed using the marker. The curve for best case reaches the top left corner of the graph, the moderate case lies between the top left corner and chance diagonal and the worst case runs parallel to the chance diagonal.

A widely used summary measure of the ROC curve is the Area under the Curve (AUC) which is the probability that a randomly chosen individual/object from population '1' has a higher score than a randomly chosen individual/object from population '0'. It depicts the amount of correct classification that can be achieved using the cutoff of a marker under study. Probabilistically,

$$AUC = P(S_1 > S_0) \tag{2}$$

where S_0 and $S_1 \in S$ are the test scores of populations '0' and '1' respectively. A practical lower bound for AUC is 0.5 and any test with AUC = 0.5 is said to have random classification. As the value of AUC gets closer to 1, better the performance of a test. The mathematical formulation of *Binormal ROC model* is given in detail in the next section.

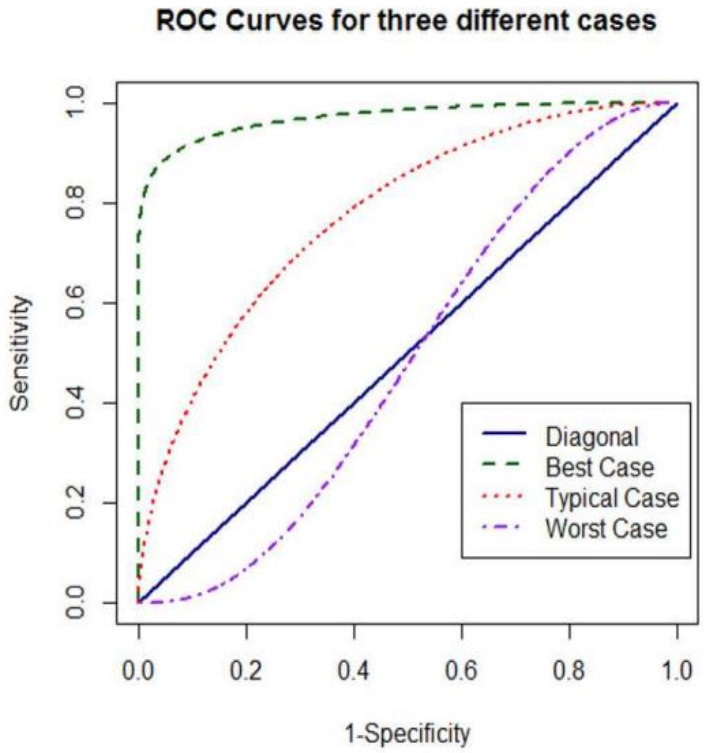


Figure 1: Typical forms of ROC curve

2. Binormal ROC Curve and its Ramifications

Let S_0 and S_1 be two random variables which denote scores from populations '0' and '1' respectively. These are assumed to follow normal distribution with means μ_0 and μ_1 and standard deviations σ_0 and σ_1 respectively (Green and Swets, 1966). i.e., $S_0 \sim N(\mu_0, \sigma_0^2)$ and $S_1 \sim N(\mu_1, \sigma_1^2)$. It is assumed that the mean of population '1' is greater than population '0' ($\mu_1 > \mu_0$) but no restrictions are posed on standard deviations. The intrinsic measures of ROC curve, TPR and FPR at a cutoff 'c' are defined as

$$FPR = x(c) = \Phi\left(\frac{\mu_0 - c}{\sigma_0}\right); \;\; TPR = y(c) = \Phi\left(\frac{\mu_1 - c}{\sigma_1}\right) \tag{3}$$

using FPR, we obtain $c = \mu_0 - \sigma_0 \Phi^{-1}[x(c)]$, then TPR can be written as

$$TPR = y(c) = \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}[x(c)]\right) \tag{4}$$

The expression for Binormal ROC curve becomes

$$y(c) = \Phi(a + b\ \Phi^{-1}[x(c)]) \tag{5}$$

where $a = \frac{(\mu_1 - \mu_0)}{\sigma_1}$, $b = \frac{\sigma_0}{\sigma_1}$ and $\Phi(.)$ is the standard normal deviate.

The AUC of the ROC curve can also be defined as the average true positive rate over all possible false positive rates in the range (0, 1).

$$AUC = \int_0^1 y(c)\ dx(c) \tag{6}$$

The expression for the AUC of Binormal ROC model is given as

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right) \tag{7}$$

3. ROC Models with Multiple Markers

One of the problems in Obstetrics and Gynecology is to identify a better procedure which helps in studying the blood flow from womb of the mother to baby for identifying the baby's growth. The study has multiple markers that need to be considered to identify whether there is a sufficient blood flow which in turn helps in classifying the subjects into one of the two groups: with adequate and inadequate blood flow. Another situation pertaining to Ecology is also observed where there is a need to identify the species type of a bug as well as to distinguish the gender of a particular species based on the features/characteristics of that particular bug. Hence, one cannot always depend on a single marker to judge the individual's/object's status. This scenario creates a necessity to develop a model that considers more than one marker for classification.

Su and Liu (1993) proposed best linear combinations where both healthy and diseased populations follow multivariate normal distribution by considering two cases, one with proportional covariance matrices and the other with no restriction on covariance matrices. In the first case with proportional covariance matrices the linear combination, is said to maximize sensitivity over a range of specificities. In the case of populations with unequal covariance matrices, the linear combination is the one that maximizes AUC among all possible combinations. Further confidence intervals were developed for AUC of the Su and Liu model by Reiser and Faraggi (1997) and named it as *Generalized ROC model*. Schisterman *et.al.* (2004) discussed covariates effects on the generalized ROC model and provided approximate confidence intervals for the measure AUC.

Liu *et al.* (2005) proposed methods to estimate the best linear combinations by maximizing sensitivity at a fixed specificity for Su and Liu model. They proved that the linear combinations proposed outperform Su and Liu model when there exists heterogeneity among covariance structures.

Countable articles were found in literature for the above multivariate classification in the context of ROC curves. The model proposed by Su and Liu (1993) was used as a base model for a considerable number of these articles. However, Su and Liu model has mathematical complexity when dealing with non-proportional covariance structures. Several authors provided improvisations on the model but none of them suggested a single linear combination that can accommodate both equal and unequal covariance structures. This motivated to the development of a new ROC model that can accommodate equal and unequal covariance structures by linearly combining multiple markers at hand (Sameera *et al.*, 2016).

4. The Multivariate Receiver Operating Characteristic (MROC) Curve

Let $X = (x_1, x_2, \ldots, x_k)$ be the 'k' markers involved in the study. Let π_0 and π_1 be two independent populations (groups) assumed to follow multivariate normal distribution with mean vectors μ_0, μ_1; covariance matrices Σ_0, Σ_1 and sample sizes n_0, n_1 respectively and $n = n_0 + n_1$. Then the probability density function for $\pi_i, i = 0, 1$ is given by

$$f(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)\right\}$$

where Σ is positive definite.

Let $x(c)$ denote the false positive rate (FPR) and $y(c)$ denote the true positive rate (TPR) where 'c' is the cutoff. The expressions for FPR, c and TPR are defined as

$$FPR = x(c) = P(U > c|\pi_0) = 1 - \Phi\left(\frac{c - b^T\mu_0}{\sqrt{b^T\Sigma_0 b}}\right) \tag{8}$$

where $b(\neq 0)$ be a $kX1$ vector and 'U' is the test score. Using (8), the expression for 'c' is given as

$$c = b^T\mu_0 + \sqrt{b^T\Sigma_0 b}\ \Phi^{-1}(1 - x(c)) \tag{9}$$

where $\phi^{-1}(.)$ is the inverse function of $\Phi(.)$

$$TPR = y(c) = P(U > c|\pi_1) = \Phi\left(\frac{b^T\mu_1 - c}{\sqrt{b^T\Sigma_1 b}}\right) \tag{10}$$

substituting (9) in (10) we get

$$ROC(c) = \Phi\left[\frac{b^T(\mu_1 - \mu_0) - \sqrt{b^T\Sigma_0 b}\ \Phi^{-1}(1 - FPR)}{\sqrt{b^T\Sigma_1 b}}\right] \tag{11}$$

The expression in (11) is the form of MROC curve.

The linear combination is defined as $U = b^T X = b_1x_1 + b_2x_2 + \ldots + b_kx_k$, where the vector 'b' is obtained using Minimax procedure (Anderson and Bahadur ,1962) as

$$b = [t\Sigma_1 + (1-t)\Sigma_0]^{-1}(\mu_1 - \mu_0) \tag{12}$$

here 't' is a constant which lies in the interval (0, 1) and its value is determined by trial and error method. The cutoff 'c' at each 't' can be obtained through Minimax procedure by

equating TPR and FPR. On solving, we obtain

$$c = \frac{b^T\mu_1\sqrt{b^T\Sigma_0 b} + b^T\mu_0\sqrt{b^T\Sigma_1 b}}{\sqrt{b^T\Sigma_1 b} + \sqrt{b^T\Sigma_0 b}} \tag{13}$$

The AUC of MROC curve can be obtained by integrating (11) over [0, 1]. However, the expression of AUC can also be derived using probabilistic notations. Let U_0 and U_1 be the test scores randomly taken from π_0 and π_1 populations respectively, $(U_1 > U_0)$.

$$\therefore AUC = P(U_1 > U_0)$$

$$AUC = P(U_1 - U_0 > 0)$$

The test scores U_0 and U_1 are independent and follow normal distribution

$$i.e., U_0 \sim N(b^T\mu_0, b^T\Sigma_0 b) \text{ and } U_1 \sim N(b^T\mu_1, b^T\Sigma_1 b), \text{ then}$$

$$U_1 - U_0 \sim N(b^T\mu_1 - b^T\mu_0, b^T\Sigma_0 b + b^T\Sigma_1 b).$$

Hence, if 'z' denotes standard normal variable then,

$$AUC = \Phi\left(\frac{b^T(\mu_1 - \mu_0)}{\sqrt{b^T(\Sigma_1 + \Sigma_0)b}}\right) \tag{14}$$

Higher the AUC lower the overlapping area of two populations and vice versa.

The two intrinsic measures of ROC curve that are used in plotting the curve, Sensitivity (S_n) and Specificity (S_p) defined as abilities of correct identification of the two groups '1' and '0' respectively and are given as follows

$$S_n = P(U > c|\pi_1) = \Phi\left(\frac{b^T\mu_1 - c}{\sqrt{b^T\Sigma_1 b}}\right); \;\; S_p = P(U < c|\pi_0) = \Phi\left(\frac{c - b^T\mu_0}{\sqrt{b^T\Sigma_0 b}}\right) \tag{15}$$

5. Introduction to Crossover Curves

In classification, attention is required for those reference values (cutoff) of markers which provide at least a moderate amount of correct classification with a greater susceptibility. In usual context of assessing the performance of a test, scores which are nearer to reference value are given same amount of weightage as that of the scores farther from reference value. The AUC so computed will be contaminated and the true accuracy or the actual performance will be masked. This misleads the interpretation of the measures of ROC as well as the optimal cutoff and leads to high amount of misclassification. Let us consider two tests A and B for better identification of a particular abnormality in individuals. Suppose that the curves of tests A and B cross each other and have at most similar accuracies. Under these circumstances, it is very difficult to notify a better test which has more ability to distinguish individuals.

To fix this issue, a solution to compare two crossover curves by means of a modified version of AUC (mAUC) of MROC curve by eliciting the importance of mAUC over AUC. Numerical illustrations are given using real data sets.

5.1. Modified Area under the Curve (mAUC)

AUC is the probability that an individual/object from group '1' has a score greater than individual/object from group '0'. One small drawback with this definition is that, it does not take into account the amount by which the scores of group '1' and group '0' differ. To overcome this, a small weight is assigned to those scores where the difference between scores is comparatively small (Figure 2).

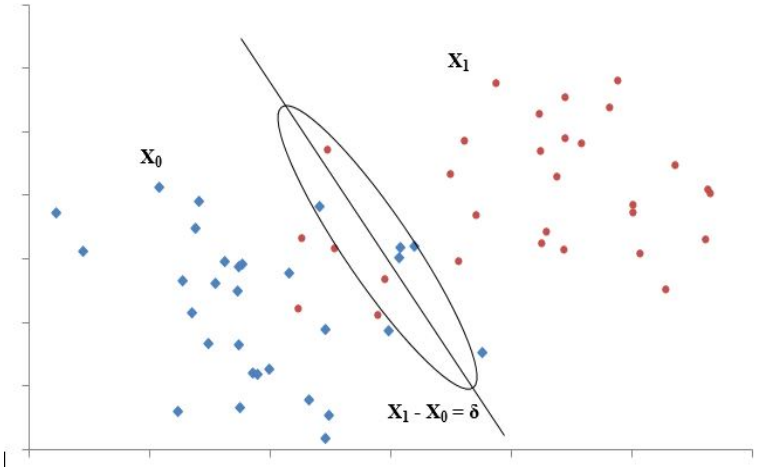


Figure 2: Hypothetical distribution of scores of two populations

mAUC was defined probabilistically by Yu *et al.* (2014) under univariate setup as weighted sum of two AUC's.

$$i.e.,\ mAUC = P(X_1 - X_0 > \delta) + (1-\lambda)P(0 < X_1 - X_0 \leq \delta)$$

$$mAUC = (1-\lambda)P(X_1 > X_0) + \lambda P(X_1 > X_0 + \delta) \tag{16}$$

The first part of this mAUC represents the conventional AUC with $(1-\lambda)$ *as its weight and the second part constitutes an additional parameter '*δ*' with '*λ*' as its weight.* The main role of 'δ' is to magnify the true status of scores of the individuals that are nearer to the reference value. Once the 'δ' value is imposed, a clear identification can be made about those scores that can be treated as true positives, which is the criterion of interest. This supports in giving out an accuracy which can be considered to be better than the conventional AUC. Using the above probabilistic notations, mAUC is derived for MROC model and is given as

$$mAUC = (1-\lambda)\ \Phi\left(\frac{b^T(\mu_1-\mu_0)}{[b^T(\Sigma_0+\Sigma_1)b]^{\frac{1}{2}}}\right) + \lambda\ \Phi\left(\frac{(b^T(\mu_1-\mu_0)-\delta)}{[b^T(\Sigma_0+\Sigma_1)b]^{\frac{1}{2}}}\right) \tag{17}$$

In equation (17), the values of parameters 'λ' and 'δ' are to be chosen in such a way that the true accuracy of a test can be extracted by minimizing the effect of nearby points of the threshold. If 'λ' value is taken to be 0, mAUC reduces to AUC and if it is taken as 1, the probability $P(X_1 > X_0+\delta)$ is only taken into account. Any value of 'λ' greater than 1 would result in making the probability $P(0 < X_1 - X_0 \leq \delta)$ value a penalty. Hence, a reasonable choice for 'λ' lies in the range (0, 1), larger the 'λ' value lower the importance on AUC. In order to choose a 'δ' that is meaningful and reasonable the following result is used.

Result: The confidence interval for mean vector of multivariate normal distribution is

$$\left(b^T \bar{X} - \sqrt{\frac{k(n-1)}{n(n-k)}\, F_{k,(n-k)}(\alpha)\, b^T S b},\;\; b^T \bar{X} + \sqrt{\frac{k(n-1)}{n(n-k)}\, F_{k,(n-k)}(\alpha)\, b^T S b} \right)$$

where n is the number of samples; k is the number of markers and $b^T Sb$ is the quadratic form. (Result 5.3, Johnson and Wichern (2007), p.225)

Using the above result, the upper bound of population '0' can be written as

$$b^T \bar{X}_0 - \sqrt{\frac{k(n_0-1)}{n_0(n_0-k)}\, F_{k,(n_0-k)}(\alpha)\, b^T S_0 b}$$

and parameter 'δ' can be chosen as

$$\sqrt{\frac{k(n_0-1)}{n_0(n_0-k)}\, F_{k,(n_0-k)}(\alpha)\, b^T S_0 b}$$

. The main reason for this choice of 'δ' is that it is part of the upper bound for the mean vector of population '0'. If an observed score is larger than this upper bound, then individual's status can be affirmatively called as true positive. The variance of mAUC expression cannot be derived explicitly and hence the concept of bootstrapping is used. If 'B' bootstraps are generated from the dataset, then the estimate and variance of mAUC is given as

$$\hat{mAUC}_{bs} = \frac{1}{B}\sum_{b=1}^{B} mAUC_b \;;\;\; Var(\hat{mAUC}_{bs}) = \frac{1}{B-1}\sum_{b=1}^{B}(mAUC_b - \hat{mAUC}_{bs})^2 \qquad (18)$$

5.2. Comparing two Crossover MROC curves

Two crossover curves can be compared using their mAUC values. The testing procedure proposed to test the hypothesis

$$H_0 : mAUC_{(1)} = mAUC_{(2)} \sim H_1 : mAUC_{(1)} \neq mAUC_{(2)}$$

for identifying the difference between two cross over MROC curves is defined as

$$Z = \frac{\hat{mAUC}_{bs(1)} - \hat{mAUC}_{bs(2)}}{var(\hat{mAUC}_{bs(1)}) - var(\hat{mAUC}_{bs(2)})} \qquad (19)$$

where $\hat{mAUC}_{bs(i)}$ and $var(\hat{mAUC}_{bs(i)});\; i = 1, 2$ can be estimated using equation (18). The Z statistic follows standard normal distribution asymptotically. The bootstrapped confidence interval for mAUC can be obtained using

$$\hat{mAUC}_{bs} \pm Z_{\left(1-\frac{\alpha}{2}\right)}\, var(\hat{mAUC}_{bs})$$

6. Results and Discussions

The proposed methodology is supported using real datasets namely, IUGRFDS and ILP datasets. For illustration purposes, all computations of mAUC and its confidence intervals under real datasets are given at $\lambda = 0.3, 0.5, 0.8$.

IUGRFDS dataset

The dataset IUGRFDS contains data collected from two independent diagnostic procedures CPR and MCA which exhibit a moderate amount of classification. Here, comparison is to be made between CPR and MCA procedures in order to find out which procedure is better in identifying the sufficient blood flow from the mother to baby. The AUC's and mAUC's of CPR and MCA along with their corresponding Z statistic value are computed and reported in Table 1. The crossover MROC curves for CPR and MCA procedures are shown in Figure 3.

Table 1: Comparison between CPR and MCA using mAUC

Measure	CPR (LL, UL)	MCA (LL, UL)	Z value (p-value)
$mAUC_{0.3}$	0.6551 (0.5196, 0.7815)	0.5902 (0.5034, 0.7254)	0.8008 (0.212^{NS})
$mAUC_{0.5}$	0.6369 (0.5106, 0.7631)	0.5774 (0.4453, 0.7090)	0.7070 (0.239^{NS})
$mAUC_{0.8}$	0.6095 (0.4817, 0.7577)	0.5581 (0.4558, 0.7032)	0.5777 (0.282^{NS})
AUC	0.6824 (0.5702, 0.7906)	0.6095 (0.5329, 0.7139)	0.9536 (0.170^{NS})

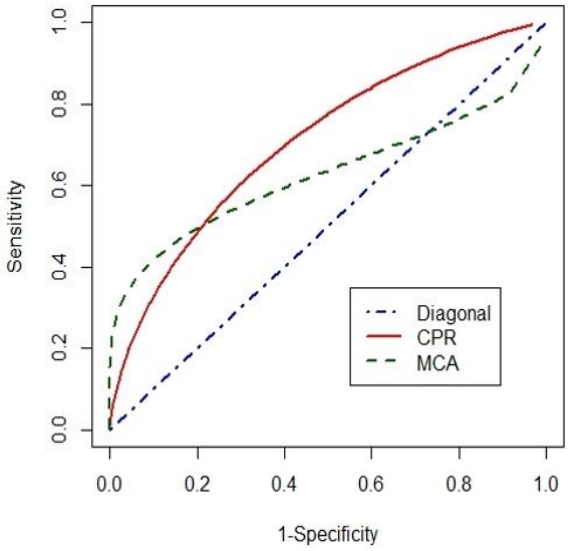


Figure 3: Crossover MROC curves for IUGRFDS dataset

Here NS = Not Significant, LL = Lower Limit, UL = Upper Limit.

For three values of λ, mAUC values are lower than that of AUC values. This is due to the fact that mAUC expression takes value of λ into account which results in assigning an appropriate weight to those scores that are closer to the threshold for extracting the true accuracy of a diagnostic procedure. However, the choice of λ should be in such a way that the accuracy is not too low. The results portrayed in Table 1 depict that both the procedures; CPR and MCA are equally effective in identifying the blood flow from mother to baby.

ILP dataset

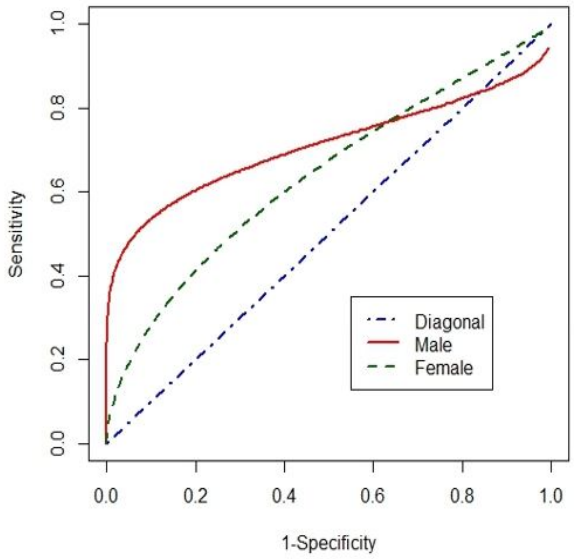
The ILP dataset is divided into two sets based on gender as males and females to check whether disease identification is identical in both the genders. MROC curves of males and females are then compared to check if classification is better in one gender compared to the other. The mAUC and AUC values are calculated for both datasets and placed in Table 2 along with their Z values and significance. The MROC curves obtained for males and females can be seen in Figure 4.

Table 2: Comparison between Males and Females using mAUC

Measure	Males (LL, UL)	Females (LL, UL)	Z value (p-value)
$mAUC_{0.3}$	0.6989 (0.6721, 0.7241)	0.6116 (0.5252, 0.7025)	1.8327 (0.033^{NS})
$mAUC_{0.5}$	0.6908 (0.6597, 0.7230)	0.5912 (0.5118, 0.6697)	2.0605 (0.019*)
$mAUC_{0.8}$	0.6788 (0.6571, 0.7124)	0.5606 (0.4512, 0.6598)	2.3824 (0.009*)
AUC	0.7109 (0.6759, 0.7297)	0.6422 (0.5591, 0.7203)	1.4726 (0.070^{NS})

Here NS = Not significant, * = significant, LL = Lower Limit, UL = Upper Limit.

Figure 4: Crossover MROC curves for ILP dataset



A better classification is seen in males than females when mAUC's obtained at $\lambda = 0.5$ and $\lambda = 0.8$. However, the Z value obtained for AUC's shows no difference between the curves indicating that the influence of scores close to the threshold is high and masking the true information. The result obtained at $\lambda = 0.3$ depicts an insignificant value stating that the weight assigned is not sufficient to extract true information from the scores of markers that are nearer to cutoff.

Observations

In this work, detailed discussion is made on a new summary measure for the MROC curve namely modified AUC is proposed. Inferential procedure is developed for this modified AUC in order to identify the true difference between MROC curves that cross each other.

References

Anderson, T. W. and Bahadur, R. R. (1962). Classification into Two Multivariate Normal Distributions with Different Covariance Matrices. *Annals of Mathematical Statistics*, **33**, 420–431.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley. New York.

Johnson, R. A. and Wichern, D. W. (2012). *Applied Multivariate Statistical Analysis*. 6th Edition, Prentice Hall India Learning Private Limited.

Krzanowski, W. J., and Hand, D. J. (2009). *ROC Curves for Continuous Data*. CRC Press.

Liu, A., Schisterman, E. F. and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, **24(1)**, 37-47.

Provost, F. J., Fawcett, T. and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *In Proceedings of International Conference of Machine Learning*, **98**, 445-453.

Reiser, B. and Faraggi, D. (1997). Confidence intervals for the generalized ROC criterion. *Biometrics*, **53(2)**, 644-652.

Schisterman, E. F., Faraggi, D. and Reiser, B. (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, **23(21)**, 3319-3331.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, **88(424)**, 1350-1355.

Yu, W., Chang, Y. I. and Park, E. (2014). A modified area under the ROC curve and its application to marker selection and classification. *Journal of the Korean Statistical Society*, **43**, 161 – 175

Special Proceeding: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 125-141

Evolutionary Computing for Optimum Crop Planning

Rajni Jain[1], **Shbana Begam**[2] **and Prem Chand**[1]
[1]*ICAR-National Institute of Agricultural Economics and Policy Research*
[2]*ICAR-National Institute for Plant Biotechnology*

Received: 03 August 2021; Revised: 15 August 2021; Accepted: 17 August 2021

Abstract

Evolutionary computing (EC) is a soft computing technique inspired by the biological concept of natural selection or Darwin's theory in genetics. An EC algorithm starts with creating a population consisting of individuals that represent solutions to the problem. The first population of solution is created randomly and then refined by an EC algorithm. Thus, an EC algorithm does the job of environmental pressure which leads to the survival of the fittest and in turn the increase of the average fitness of the population. The final converged population is the optimum solution to a problem.

Crop planning is essential for agricultural systems management for increasing productivity and sustainability of the resources. The central objective of optimal crop planning is to search for an optimal combination of different resources and crops to maximize the overall contributions by concurrently satisfying a set of constraints *e.g.* availability of land, water, capital *etc.*

While optimizing crop planning with more than one objective function, there exist several Pareto optimal solutions. To solve such a problem is not as straightforward as it is for a conventional single-objective optimization problem. Evolutionary algorithms have helped in solving complex problems to provide an optimum solution. In this paper, Particle Swarm Optimization algorithms are adapted for optimal crop plans using a case study. We conclude that evolutionary computing is a viable and convenient alternative for crop planning at a regional level and should be explored further for crop planning at a farm and country level.

Key words: Crop planning; PSO; Bundelkhand; Multi-objective optimization.

1. Introduction

Crop planning is essential for agricultural production systems management and it can resolve how much resources are allocated to different cropped areas in obtaining certain goals such as the maximization of return from cultivated land under the limitation of resources (Jain *et al*., 2015). The central objective of crop planning is to search for an optimal combination of crops amongst those considered to maximize the overall contributions while concurrently satisfying a set of constraints such as land availability and capital. An important issue in those problems is the optimization objective(s). Optimization problems may be single or multi-objective based on the number of goals to be achieved. Maximization of net returns,

Corresponding Author: Rajni Jain
Email: rajnijain67@gmail.com

maximization of gross margin, and minimization of water use are some of the goals that are desired.

Traditionally, linear programming based approaches have been used for crop planning similar to other domains like route optimization, resources optimization in the manufacturing industry *etc*. (Jain *et al.*, 2017). With the advent of high speed and high memory-based computers, EC methods need to be explored for optimum crop planning because of the four reasons namely: (i) easier implementation, (ii) exploration of non-linear solutions, (iii) avoidance of local solutions, and (iv) easier implementation of multi-objective optimization.

2. Review of Literature

Optimization techniques for crop planning have been in use for a long time (Jain *et al.*, 2018b). Different optimization models based on the concept of linear programming for utilizing the resources efficiently and providing optimal crop plan like the Regional Crop Planning model (Jain *et al.*, 2015; Jain *et al.*, 2017) and allocating water resources optimally (Sethi *et al.*, 2006) have been developed. Evolutionary algorithms have been successfully studied and applied extensively in the past few decades in agriculture, engineering and various other fields, and helped in solving complex problems to provide near to optimum solutions. Evolutionary algorithms like GA (Kumar *et al.*, 2006) have been used for allocating optimal crop water from an irrigation reservoir to maximize the sum of the relative yields from all crops in the irrigated area. The authors recommended optimum crop planning for maximizing irrigation benefits for a typical irrigation system. Other crop planning models based on GA, swarm intelligence, and differential evolution algorithms have been used to maximize total net benefit and production from farming (Sharma and Jana 2009; Nath *et al.*, 2020). Adeyemo and Otieno (2010) formulated a multi-objective optimal crop-mix problem and solved using the generalized differential evolution. Crop planning is a multi-dimensional problem, therefore it is desirable to have more than one objective function to solve the problem and to get more optimal results.

Pareto rank candidate solutions keep an archive of all non-dominated solutions in multi-objective optimization using EC (Coello *et al*., 2004). Some studies which already used multi-objective optimization using PSO are parameter estimation in hydrology (Gill *et al.* 2006), estimation of soil mechanical resistance parameter (Hosseini *et al.*, 2016) and optimization of feeding composition for methane yield maximization (Wang *et al.*, 2012). Jain *et al*. (2018b) presents a review of available crop planning techniques and concluded the need for exploring EC techniques for crop planning (Table 1). More on crop planning methodology, data preparation steps, case studies of optimum crop plans under different constraints are available in the literature (Jain *et al.*, 2019; Jain *et al.*, 2018a).

2.1. Evolutionary computing framework

The review reflects that evolutionary computing techniques have not been explored for crop planning. Evolutionary algorithms fall into the category of “generate and test” algorithms. They are stochastic, population-based algorithms. In genetic algorithms and evolutionary computation, crossover, also called recombination, is a genetic operator used to combine the genetic information of two parents to generate new offspring. The mutation is a genetic operator used to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next. The mutation alters one or more gene values in a chromosome from its initial state. In mutation, the solution may change entirely from the

previous solution. Variation operators (recombination and mutation) create the necessary diversity to facilitate novelty. Selection reduces diversity and acts as a force pushing quality (Figure 1).

Table 1: **Summary of past work related to optimization of crop plans**

Studies	Scope	Approach	Objective(s)	Constraints
Maleka 1993	Farm (Gwembe Valley, Zambia)	Target MOTAD	Maximize net revenue	Land, labour, credit, moisture, cost of risk taking
Tajuddin *et al*. 1994	Farm (Bangladesh)	LP	Max net returns	Land, labour, cap min cereal requireme
Sarkar and Lingard 2002	Country(Bangladesh)	LP (QSOM)	Max over all contribution of agriculture sector	Land, capital, area import bound, minim food requirement
Sethi *et al*. 2002	Farm(Coastal river	DLP& CCLP (QSB)	Maximize net revenue	Land, water
Sharma *et al*. 2007	District(Ghaziabad, India)	FGP (Lingo)	Max crop production, net profit, labour, min water, machine	Land, capital, food
Mohaddes and Mohayidin 2008	Farm (Atrak Watershed, Iran)	FGP	Max profit, Max employment, min erosion	Land, water
Sharma *et al*. 2009	State (Himachal	DNLP (GAMS)	Max profit	Land, labour, capital
Sethi and Panda 2011	Farm (Costal River	LP based DSS (QSB+)	Max net returns	Land, water
Soltani *et al*. 2011	Local(Kerman province, Iran)	FGP & LP (QSB)	Max crop production & net returns, min labour employment, water &	Land, labour, wa machine
Rani *et al*. 2012	Farm(Mahabubnagar ,	LP	Max profit, input cost min & water usage min	Land, water, min-r yield requirement
Karunakaran *et al*. 2012	Regional (Bhavani	LP (GAMS)	Max net income	Land, water, soil
Mirkarimi *et al*. 2013	Farm (Amol,	FGP (Lingo)	Max profit,	Self sufficiency
Mortazavi *et al*. 2014	Country (Iran)	LP, GP,FP, FGP (MCDA/MCDM)	Max gross revenue, max employment, min water consumption,	Land, labour, water
Kaur *et al*. 2010; Kaur *et al*. 2015	State (Punjab)	LP	Max profit	Land, labour, wa capital , crop maxim minima
Martin *et al*. 2015	Country (Spanish	LP (GAMS)	Max net return	greening constra based on CAP of EU

Source: Jain et al., 2018*b*

3. Material and Methods

3.1. Study area and the dataset

The Bundelkhand is a semi-arid region of India that comprises seven districts (Jhansi, Jalaun, Lalitpur, Mahoba, Hamirpur, Banda and Chitrakoot) of Uttar Pradesh and six districts (Datia, Tikamgarh, Bhhatarpur, Panna, Damoh and Sagar) of Madhya Pradesh (https://bundelkhand.in/). Agriculture in Bundelkhand is rainfed, diverse, complex, under-invested, risky and vulnerable. In addition, droughts, short-term rain, intermittent dry spells, flooding in fields, and abrupt temperature change add to the uncertainties and seasonal migrations (Jain *et al.*, 2020). Major problems in Bundelkhand are water deficiency, infertility of the land, soil erosion, improper land distribution, depleting groundwater resources and unscientific cultivation in terms of non-use of modern methods in agriculture (GoI, 2015). The yields of cereals, pulses and oilseeds in kharif and rabi seasons are generally lower (Yadav *et al.*, 1996) and even below the parent state average for the majority of the

crops (GoUP, 2018; GoMP, 2018). Therefore, considering water, land, capital and other constraints, Bundelkhand is a challenging region for optimum crop planning. The dataset and variables used for experimentation and optimum plan development are summarised in Table 2 and Table 3.

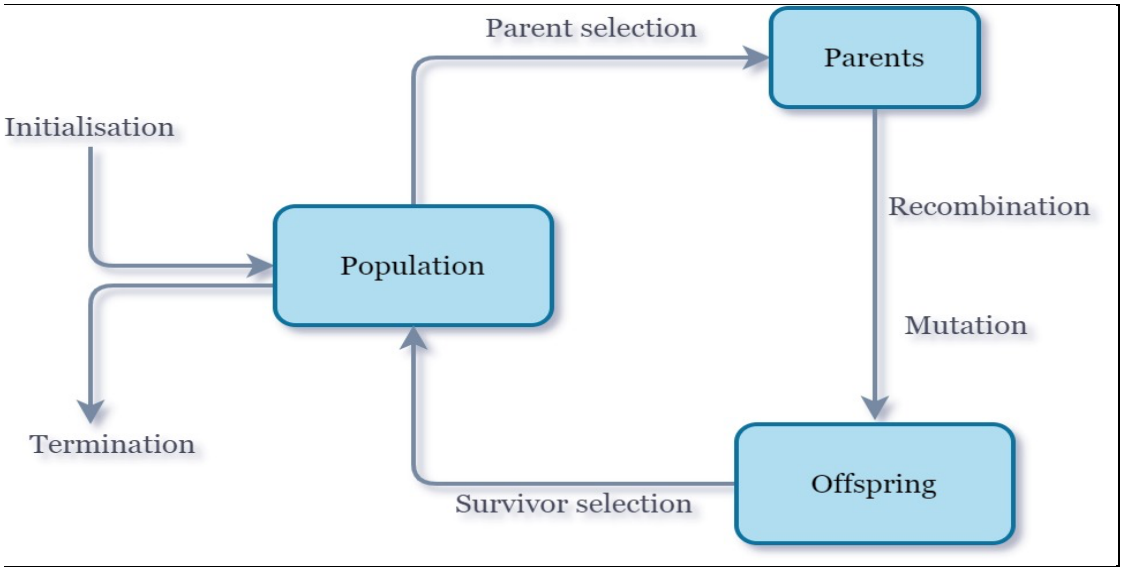


Figure 1: General scheme of evolutionary computing

The data values for the study are based on the unit level data available from the cost of cultivation scheme (https://eands.dacnet.nic.in/Cost_of_Cultivation.htm) for the study area for the year 2017-18 following the data extraction methodology developed by Jain *et al.* (2015). However, the data used for validation of model is interim and have not been generalized for a longer period. Methodology for estimating crop-wise water requirement is based on the literature (Chand *et al.*, 2020). The dataset was compiled as per Table 2 and Table 3. *MinArea* and *MaxArea* are estimated based on current areas, advice of the experts and food security considerations (Jain *et al.*, 2015). Table 4 illustrates a sample from such a dataset. The variables are self-explanatory.

Table 2: Description of the dataset

Item	Description
Location	Bundelkhand
Number of districts	13 *i.e.* Jhansi, Jalaun, Lalitpur, Mahoba, Hamirpur, Banda and Chitrakoot) of Uttar Pradesh and six districts (Datia, Tikamgarh, Bhhatarpur, Panna, Damoh and Sagar)
Number of variables	5 *i.e. MinArea*, *MaxArea*, *R* (Net Returns per ha), *M* (Working Capital per ha), *W* (Water Requirement per ha) for each crop
Number of crops	26 *i.e.* Arhar, Bajra, Barley, Berseem, Chillies, Chickpea, Groundnut, Guar, Jowar, Khesari, Lentil, Linseed, Maize, Mentha, Mesta, Moong, Mustard, Onion, Paddy, Pea, Sesamum, Soyabean, Sugarcane, Tomato, Urad, Wheat

Table 3: Description of the variables

Variable ID	Variable description	Datatype	Unit
Crop	Crops name	character	Identification variable
A	Area to be allocated for a crop	double	000' hectare (ha) per crop
NCA	Net cultivable area	double	000' ha for all crops in the region
MinArea	Minimum allocated area for the crop	double	000' ha per crop
MaxArea	Maximum allocated area for the crop	double	000' ha per crop
R	Net returns from a crop	double	rupees per ha
Z	Overall returns from the region	double	million rupees
M	Money (Working Capital) required for purchase of inputs and other basic things for a crop cultivation	integer	rupees per ha
C	Overall working capital available	double	rupees
W	Water requirement for a crop	double	m^3 per ha
V	Overall irrigated water available	double	billion cubic meters (BCM)

Table 4: Dataset illustration for selected sample crops

Crop	Minimum Area desired under a crop (*MinArea*) (ha)	Maximum Area desired under a crop (*MaxArea*) (ha)	Irrigated Water Requirement (*W*) (m^3/ha)	Net Returns (*R*) (`/ha)	Working Capital (*M*) (`/ha)
Chickpea	867074	1300611	1114	20734	10373
Groundnut	83205	124663	0	8933	12862
Guar	343	36000	0	18000	5938
Paddy	240000	360040	250	16721	16181
Wheat	1695819	2543729	3319	27565	9942

Table 5 presents the comprehensive list of crops, the existing cropping pattern and the crop calendar based on the environment and agricultural practices in the region. A value of '1' means the land may be cultivated by a crop in the specified calendar month.

3.2. Generalised mathematical model

Mathematical Programming can be used for developing and presenting the mathematical model of optimum crop plan (Jain *et al.*, 2019). It provides formulation to use limited resources and maximises the productivity with minimum working capital and/or minimum irrigation water use under many constraints like land availability *etc.* The

mathematical formulation for the optimum crop model is characterised by a set of equations in Table 6. The presented formulation is based on two objective functions defined in Equation 1 and Equation 2. In this model, *net returns from a crop*(R) = (*total value of the main product and by-products*) – (*variable costs*). Here variable costs contain Cost $A1$ and imputed value of family labour at market prices paid and received by the farmer or imputed in some cases. The methodology for estimation of R is adopted from Jain *et al*. (2015). Maximization of Z (the sum total of returns from all the crops over the region) is the prime objective of the optimum crop model (Equation 1). Bundelkhand has limited irrigation water availability, hence it is necessary to use minimization of irrigation water *i.e.* V as another objective function (Equation 2). In the case of single objective-based implementation, irrigation water availability needs to be known and it can be used as a constraint. In this paper, we have implemented both (single as well as multi-objective) kinds of mathematical models.

Table 5: Crops and respective calendar for crops used in the model

Crops	Area (ha)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Arhar	105616	0	0	0	0	0	1	1	1	1	1	1	1
Bajra	30587	0	0	0	0	0	1	1	1	1	0	0	0
Barley	67912	1	1	0	0	0	0	0	0	0	0	1	1
Berseem	2000	1	1	1	1	0	0	0	0	0	1	1	1
Chillies	4020	0	0	0	0	0	0	0	1	1	1	1	0
Chick	867074	1	1	0	0	0	0	0	0	0	1	1	1
Gnut	83204	0	0	0	0	0	1	1	1	1	1	0	0
Guar	3430	0	0	0	0	0	1	1	1	1	1	1	0
Jowar	85845	0	0	0	0	0	1	1	1	1	0	0	0
Khesari	5140	1	0	0	0	0	0	0	0	0	1	1	1
Lentil	268984	1	1	0	0	0	0	0	0	0	1	1	1
Linseed	22924	1	1	0	0	0	0	0	0	0	1	1	1
Maize	53948	0	0	0	0	0	1	1	1	1	0	0	0
Mentha	9500	0	1	1	1	1	1	1	0	0	0	0	0
Mesta	100	0	0	0	0	0	1	1	1	1	1	1	0
Moong	48426	0	0	0	0	0	1	1	1	1	0	0	0
R&m	111777	1	1	1	0	0	0	0	0	0	1	1	1
Onion	8833	1	1	1	0	0	0	0	0	0	1	1	1
Paddy	240000	0	0	0	0	0	1	1	1	1	1	0	0
Pea	334440	1	1	1	0	0	0	0	0	0	1	1	1
Sesame	373423	0	0	0	0	0	1	1	1	1	1	0	0
Soya	593011	0	0	0	0	0	1	1	1	1	0	0	0
Scane	16141	1	1	1	1	1	1	1	1	1	1	1	1
Tomato	515	0	0	1	1	1	1	0	0	0	0	0	0
Urad	520532	0	0	0	0	0	1	1	1	1	0	0	0

Wheat	1695819	1	1	1	0	0	0	0	0	0	0	1	1

Note: Self-explanatory short names of crops are used due to space constraint *e.g.* gnut: groundnut

Table 6: Mathematical formulation of Optimum Crop Plan

Eq.	Equation Functions	Variable Description
1.	$Max\ Z = \sum_{c=1}^{n} R_c A_c = \sum_{c=1}^{n} (Y_c P_c - C_c) A_c$	Z: overall net returns from all crops in the cultivated area R_c: Net returns from crop c per ha Y_c: yield of crop c P_c: market price of crop c C_c: cultivation cost of crop c A_c: area allocation for crop c n: number of crops available for cultivation *i.e.* 26
2.	$Min\ V = \sum_{c=1}^{n} W_c A_c$	V: Overall irrigated water W_c: water required for crop c per ha A_c: area allocation for crop c
3.	$\sum_{c=1}^{n} A_c M_c \ \leq\ C$	M_c: working capital for crop $c$in `/ha C: total working capital available for the region (regional constant constraint)
4.	$\sum_{c=1}^{n} (a_{tc}) A_c \ \leq\ NCA_t$ for all t from1 ..12	a_{tc}: crop calendar coefficient for month t and crop c. Coefficient a_{tc} = 0 means cultivation of the crop c in month t does not happen. NCA_t: net cultivable area in the region in a month(regional constant constraint). We have taken it the same for all months. Equation 4 represents 12 area constraints one for each month
5.	$A_c \geq MinArea_c$	$MinArea_c$: Minimum allowable area for crop c
6.	$A_c \leq MaxArea_c$	$MaxArea_c$: Maximum allowable area for crop c

3.3. Evolutionary computing based model

Particle Swarm Optimisation (PSO) was developed in 1995 by Kennedy and Eberhart, inspired by the behaviour of social organisms in groups, such as bird and fish schooling or ant colonies (Kennedy and Eberhart, 1995; Kennedy, 2010). The algorithm emulates the interaction between members to share information. PSO has been applied to numerous areas in optimisation and combination with other existing algorithms. This method performs the search of the optimal solution through agents, referred to as particles, whose trajectories are adjusted by a stochastic and a deterministic component. Each particle is influenced by its 'best' achieved position and the group 'best' position but tends to move randomly. One of the reasons for using the individual best is probably to increase the diversity in the quality solutions; however, this diversity can be simulated using some randomness. The individual best is useful when the optimization problem of interest is highly nonlinear and multimodal.

Consider a swarm (population) containing p particles in a k-dimensional continuous solution space. The position of the i^{th} particle is denoted as $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and each i^{th} particle has its position and velocity in k-dimensional vector. The best particle is denoted as *gbest* in the swarm. The best previous position of the i^{th} particle is recorded and represented as *pbest.* Finally, the velocity can be computed by using Equation 7 to Equation 8.

$$x_i^t = x_i^{t-1} + v_i^t - \quad (7)$$

$$v_i^t = inertia \times v_i^{t-1} + c_1 r_1 (pbest_i^{t-1} - x_i^{t-1}) + c_2 r_2 (gbest^{t-1} - x_i^{t-1}) \quad (8)$$

where *inertia* is the inertia weight, c_1, c_2 are acceleration coefficients, r_1, r_2 are random numbers between 0 and 1,v is particle velocity, x is particle position, i is a particle identifier in a swarm and t is an iteration number in the optimization process. Since each particle explores the possible solutions of space, each of them represents a candidate solution to the problem. For example, one candidate solution for optimum crop plan (based on 26 crops) for a given region is shown in Figure 2.

Figure 2: Example of an i^{th} candidate solution for PSO based optimum crop plan

In Figure 2, x_{ik} represents area allocation for a crop k where k varies from 1 to 26 in i^{th} candidate solution of our case study. Each candidate solution is called a crop plan (a particle in PSO *i.e.* x_i in Equation 7). Figure 3 presents a self-explanatory basic flow chart for implementing the PSO. We observe that it is an iterative algorithm that works until the solution converges.

3.4. Software

PSO based OCP has been implemented in this work using the MOPSOCD package in R (Naval, 2013; Table 7). Multi-objective optimization involves maximizing or minimizing multiple interacting/conflicting objective functions subject to a set of constraints. MOPSOCD is a multi-objective optimization solver based on particle swarm optimization that uses crowding distance computation to ensure an even spread of non-dominated solutions. Crowding distance is calculated by first sorting the set of solutions in ascending order of objective function values. The crowding distance value of a particular solution is the average distance of its two neighbouring solutions. The crowding distance mechanism together with a mutation operator maintains the diversity of non-dominated solutions in the external archive. The approach is highly competitive in converging towards the Pareto front and generates a well-distributed set of non-dominated solutions. The details of the package are available at https://cran.r-project.org/web/packages/mopsocd/ mopsocd.pdf.

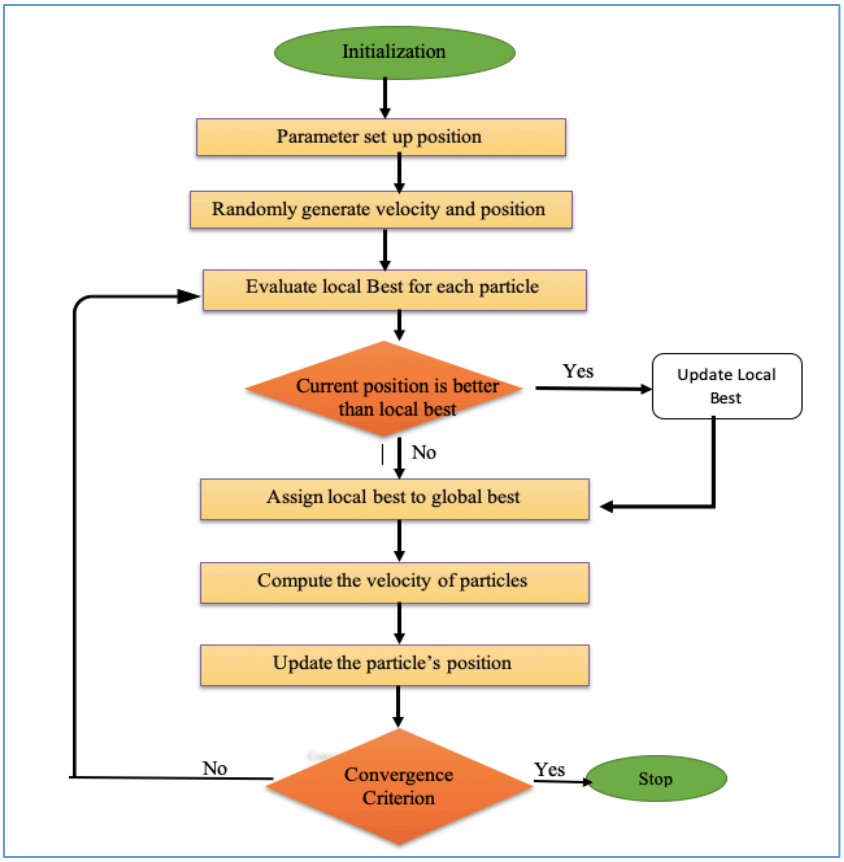


Figure 3:Basic flowchart of Particle Swarm Optimization algorithm

Table 7: Implementation details of using PSO for optimum crop planning in Bundelkhand

Algorithm and parameters	Implementing software	Package	Source
PSO $N = 100$ $c_1 = 1.041$ $c_2 = 0.948$ *inertia* = 0.629	R software	mopsocd	https://CRAN.R-project.org/package=mopsocd
		pso	https://CRAN.R-project.org/package=pso
		ggplot2	https://CRAN.R-project.org/package=ggplot2
		plot3D	https://CRAN.R-project.org/package=plot3D

Note: N is user-defined for the number of iterations;c_1, c_2 and inertia are defined in Equation 8 in Section 3.3.

4. Results and Discussion

Evolutionary computing approaches can be used for linear as well as non-linear optimization. Besides, they are convenient to use for single and multi-objective functions. We explored PSO for optimum crop plan development in Bundelkhand region for increasing farm income while utilizing the scarce resources efficiently. Presently, Bundelkhand has 5.3 million ha gross cropped area (GCA) with overall returns (Z) of `105.46 billion by using 18.6 BCM of water with an inefficiency of 52 per cent. Thus, the actual water requirement of crops that can be fulfilled is 9.12 BCM.

4.1. Single objective function

Firstly, we explain the results obtained from our experiments for single objective function optimization under various scenarios (Table 8-9). Later, we explain various scenarios under multi-objective function optimizations and the corresponding results. Table 8 presents the scenarios formulated depending on the availability of irrigation water in the region. Bundelkhand is a water-scarce region where irrigation facilities are available only in 50 per cent of the cultivable area. In the remaining area, rainfed crops are cultivated. Hence, there is a need to explore how the profitability changes under the variable availability of irrigation water.We explore a single objective function along with variable availability of irrigation water (Table 8) and variable availability of working capital (Table 9) and present the results in Table 10 and Table 11 respectively.

As water availability increases, net returns also increase. The maximum optimum profit under the current use of water is `159.41 billion which increases to `160.9 billion with an increase of water availability by 20 per cent. Under optimum plans, there is an increase in allocated area under crops that need less water. Allocation under berseem, guar, khesari, mesta and tomato increase substantially. Besides, allocation under chillies, mentha, rapeseed and mustard, onion and sugarcane decrease in most of the optimum plans. Further, we observe that area under all other crops shows an increase (Table 10). Bundelkhand has a scarcity of working capital due to the lack of formal credit infrastructure, hence it is important to develop optimum crop plans under variable constraints of working capital. With an increase in working capital by 10 per cent, we can increase net returns to `161.20 billion at the most. But, further increase in working capital will not increase the net returns due to limitations of other factors. In optimal plans, net returns increased by nearly 50-52 per cent (Table 10-11) and water use decreased up to 20 per cent. This has been made possible by the use of fallow land available in the state during rabi season. We observe that PSO using single-objective functions can achieve the optimum plans under various constraints. However, it requires estimation of the supply of resources which is not always feasible to estimate. For such scenarios, multi-objective optimization will be useful and the corresponding optimization results are explained further.

Table 8: List of scenarios and corresponding description for variable water constraints

SNo.	Short Name	Description	Water (*V*) Constraint (BCM)	Overall water used by crops (BCM)
1	W_NO	No water constraint in the model	-	12.16
2	W_Curr	Current use of water	18.65 **	9.10
3	W_Avail	Available irrigated water(after taking account of water efficiency in Bundelkhand) for model development	17.36	8.50
4	W_0.7	70% of the current water requirement	13.05	6.30
5	W_0.8	80% of the current water requirement	14.92	7.20
6	W_0.9	90% of the current water requirement	16.78	8.19
7	W_1.1	110% of the current water requirement	20.51	10.01
8	W_1.2Curr	120% of the water requirement	22.38	10.92

Notes: '*': Land and capital constraints were not changed; Model parameters land=4.2 million ha; working capital (C) =`53.15 billion

'**': Water use efficiency in Bundelkhand is 52.6%

Table 9: List of scenarios and description for variable working capital constraints

SNo.	Short name	Description	Working Capital (C) Constraint* (`billions)
1	C_NO	No working capital constraint in the model	NA
2	C_Curr	Current use of working capital is constrained	53.15
3	C_0.9Curr	90% of working capital is available	47.83
4	C_0.8Curr	80% of working capital is available	42.52
5	C_0.7Curr	70% of working capital is available	37.20
6	C_1.1Curr	110% of working capital is available	58.46
7	C_1.2Curr	120% of working capital is available	63.78

Note'*': Land and water constraints were not changed.
Model parameters are: Irrigated Water use= 18.65 BCM; land=4.2 million ha;

Table 10: Percent change in area allocations to different crops under different scenarios for the single objective function and variable irrigation water constraints

Crop	W_No	W_Curr	W_Avail	W_0.7	W_0.8	W_0.9	W_1.1	W_1.2
Arhar	10	32	34	33	23	28	45	44
Bajra	10	32	19	29	36	41	37	2
Barley	46	5	24	18	41	3	34	4
Berseem	184	630	393	1454	568	709	68	193
Chillies	–33	–5	–54	–12	17	19	1	–59
Chickpea	50	10	50	50	50	50	50	50
Gnut	47	34	21	24	29	50	34	38
Guar	399	764	865	949	735	–15	637	79
Jowar	35	48	25	29	20	48	13	38
Khesari	937	585	1032	590	635	986	299	626
Lentil	35	30	35	50	49	50	37	20
Linseed	22	23	24	44	29	18	22	9
Maize	13	17	26	10	23	5	24	14
Mentha	–47	–46	–28	–44	–39	–65	–52	–46
Mesta	2876	2931	3826	4366	820	5618	5098	4723
Moong	7	46	36	38	13	36	30	33
R&m	–34	–68	13	30	–31	50	16	–82
Onion	–17	–78	–63	–2	–60	15	–76	–53
Paddy	42	30	50	41	45	35	40	49
Pea	50	13	31	23	50	50	50	41
Sesame	47	47	47	36	41	46	44	47
Soybean	56	56	55	56	56	56	51	56
Scane	–1	–19	10	–28	28	53	–2	–17
Tomato	783	1362	1192	559	1289	1060	1135	392
Urad	49	49	48	49	49	48	49	49
Wheat	50	32	50	50	50	50	50	50
V in BCM	12.16	9.10	8.50	6.30	7.20	8.19	10.01	10.92
GCAin Mha	8.11	8.10	8.14	8.12	8.15	8.25	8.13	7.90
Z in ` 10^9	159.0	159.41	159.70	159.03	159.93	158.17	159.11	160.90
Change *Z* %	50.78	51.16	51.43	50.8	51.65	49.98	50.87	52.57

Note: Current GCA: 5.3Mha; current *Z*: `105.46 billion; Current *V* = 9.12 BCM

Table 11: Percentage change in area allocations to different crops under different scenarios for single objective function and variable working capital (C) constraints

Crop	Variable working capital (C) constraints						
	C_No	C_curr	C_0.7	C_0.8	C_0.9	C_1.1	C_1.2
Arhar	32	32	46	6	37	50	32
Bajra	32	32	12	11	32	16	32
Barley	5	5	47	20	21	17	5
Berseem	630	630	754	1630	958	1278	630
Chillies	–5	–5	–64	–35	–58	–37	–5
Chickpea	10	10	50	50	50	50	10
Gnut	34	34	38	50	43	50	34
Guar	764	764	586	945	675	851	764
Jowar	48	48	48	37	40	38	48
Khesari	585	585	366	388	1024	625	585
Lentil	30	30	43	35	50	50	30
Linseed	23	23	34	35	7	25	23
Maize	17	17	10	39	40	44	17
Mentha	–46	–46	–49	–26	–40	–64	–46
Mesta	2931	2931	1813	2479	3174	2962	2931
Moong	46	46	28	42	18	15	46
R&m	–68	–68	50	50	–66	50	–68
Onion	–78	–78	–69	–37	–13	–10	–78
Paddy	30	30	40	50	45	49	30
Pea	13	13	50	20	50	33	13
Sesame	47	47	46	29	47	40	47
Soybean	56	56	56	56	56	52	56
Scane	–19	–19	53	53	–45	–73	–19
Tomato	1362	1362	1867	1355	1001	1637	1362
Urad	49	49	49	31	19	49	49
Wheat	32	32	50	50	50	50	32
C in ` 10^9	75.43	53.15	37.20	42.52	47.83	58.46	63.78
GCA in Mha	7.96	8.10	8.01	8.00	8.02	8.20	8.22
*Z*in ` 10^9	156.70	159.41	157.60	158.60	158.40	161.20	160.80
% change *Z*	48.59	51.16	49.44	50.39	50.2	52.85	52.47

4.2. Multi-objective function

Evolutionary computing is also useful in developing optimal plans for multi-objective optimization (Table 12-13). Table 12 presents the various scenarios that were experimented with PSO. The list is only indicative to show the potential of evolutionary computing. The first two scenarios seek to optimize profit under minimum water and minimum working capital use respectively. The third scenario is just to explore the maximum capital requirement. Scenario 4 and scenario 5 make use of three objective optimizations. Table 13 shows only one such case for each scenario. However, the results of each of these optimizations are a Pareto-optimal curve showing multiple solutions for each scenario

(Figure 4-6). Thus, the evolutionary techniques can deliver the Pareto-optimal solutions as expected from multi-objective optimizations. We observe that net returns increase 15-30 per cent in different scenarios (Table 13).

The relationship of net returns (Z) with variable availability of water (V) or variable availability of working capital (C) is presented in Figure 4. These relationships are obtained based on the optimization of two objectives using evolutionary computing and the resulting Pareto optimal solutions are shown in Figure 5. We observe multiple solutions for three different models namely MaxP_MaxC, MaxP_MinC and MaxP_MinW (Table 12 for the description of scenarios). Profit maximization is the prime objective in each of these two objective-based optimizations. We observe the maximum profit in the case of MaxP_MaxC and the number of solutions is also limited with lesser variability. On the other hand, in the case of MaxP_MinC model, we get very high variability in solutions with Z ranging from 110-145 billion rupees with an average profit of 125 billion rupees in Bundelkhand. For the third MaxP_MinW model, we observe the intermediate variability of solutions with Z ranging from 122-140 billion rupees with an average profit of 132 billion rupees for the region. We observe that an increase in availability of both resources namely working capital (C) and water (V) increase net returns to the farmers. This emphasises the need to explore three objective optimizations. The number of solutions achieved for such scenarios is shown in Figure 6. Thus, a diverse set of solutions are available and any one can be chosen depending on the available resources. These models, if used at the farm level have the potential to recommend optimum plans to farmers based on the status of available inputs at farm level.

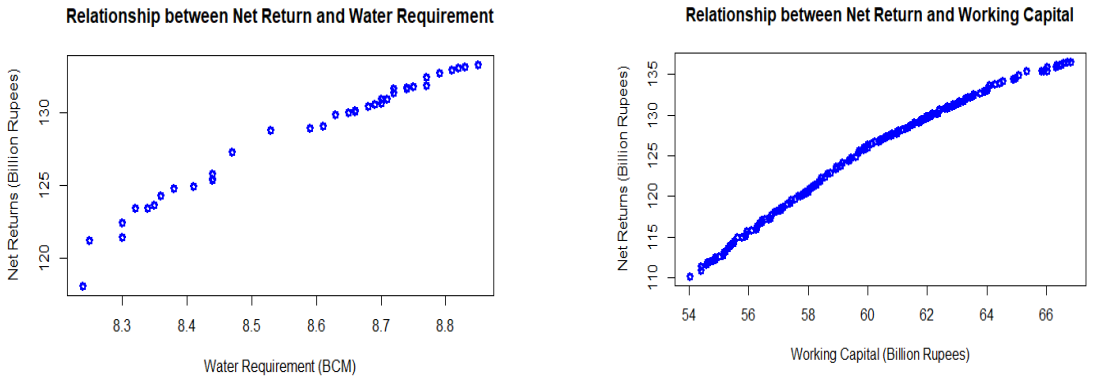
Table 12: List of multi-objective scenarios and the corresponding description

S.No*.	Short name	Description	Other parameters
1	MaxP_MinW	Maximise profit and Minimise water	C = `53.15 billion
2	MaxP_MinC	Maximise profit and Minimise Working capital	Water (V) = 9.10 BCM
3	MaxP_MaxC	Maximise profit and maximise working capital	Water(V) = 9.10 BCM
4	MaxP_MinW_MinC	Maximise profit, Min water, Min work capital	-
5	MaxP_MinW_MaxC	Maximise profit, Min water, Max work capital	-

Note '*': S.No.refers to scenario number

Table 13: Change in area allocations (%) to the crops under multi-objective scenarios

Crops	MaxP_MinW	MaxP_MinC	MaxP_MaxC	MaxP_MinW_MinC	MaxP_MinW_MaxC
Arhar	47	22	24	49	35
Bajra	9	13	6	21	31
Barley	15	31	22	5	29
Berseem	66	1221	65	511	155
Chillies	-51	-54	1	-35	-35
Chickpea	43	2	6	12	26
Gnut	32	26	47	28	28
Guar	682	284	632	579	372
Jowar	32	9	43	21	25
Khesari	761	377	267	567	856
Lentil	37	43	3	24	12
Linseed	22	17	16	33	34
Maize	10	16	42	34	32
Mentha	-61	-56	-35	-77	-59
Mesta	810	262	1900	3310	4859
Moong	44	28	22	39	13
R&m	-60	-57	-73	-64	-90
Onion	0	-65	26	-81	-87
Paddy	33	1	42	10	42
Pea	2	1	17	5	6
Sesame	38	17	46	13	40
Soybean	56	0	56	25	56
Scane	-6	-28	-11	-6	-94
Tomato	475	1781	1192	1043	1181
Urad	48	2	49	30	40
Wheat	0	21	29	13	0
C in ` 10^9	65.84	58.35	67.92	60.89	63.53
GCA in Mha	6.90	6.19	7.12	6.46	6.71
Z in ` 10^9	131.78	121.78	137.82	124.64	126.44
Z in % change	24.96	15.48	30.68	18.19	19.89



Max Net returns and Min Water **Max Net Returns and Min Working Capital**

Figure 4: **Relationship (i) Net returns (*Z*)*vs.* water used (*V*) (ii) Net returns (*Z*)*vs.* working capital (*C*) using the single objective function**

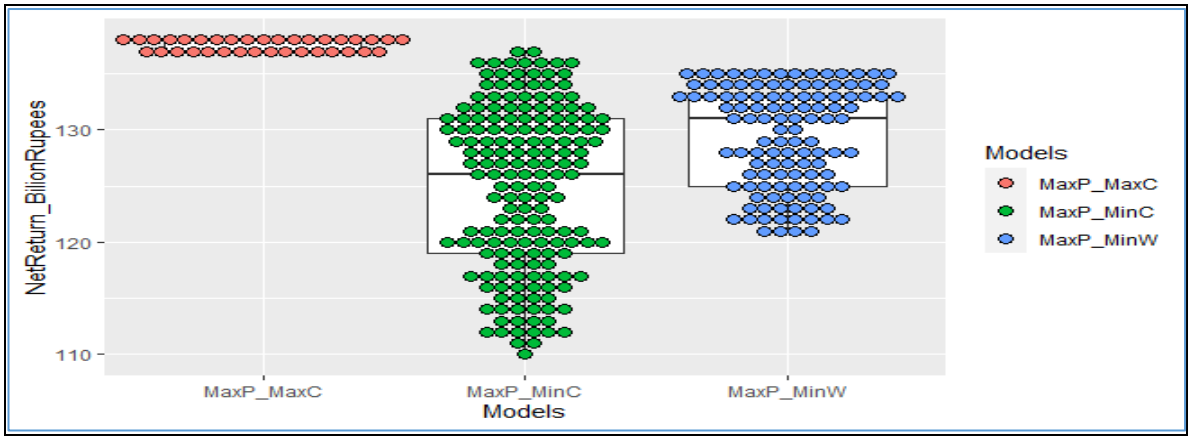


Figure 5: Number of solutions obtained from PSO based optimization for two objectives

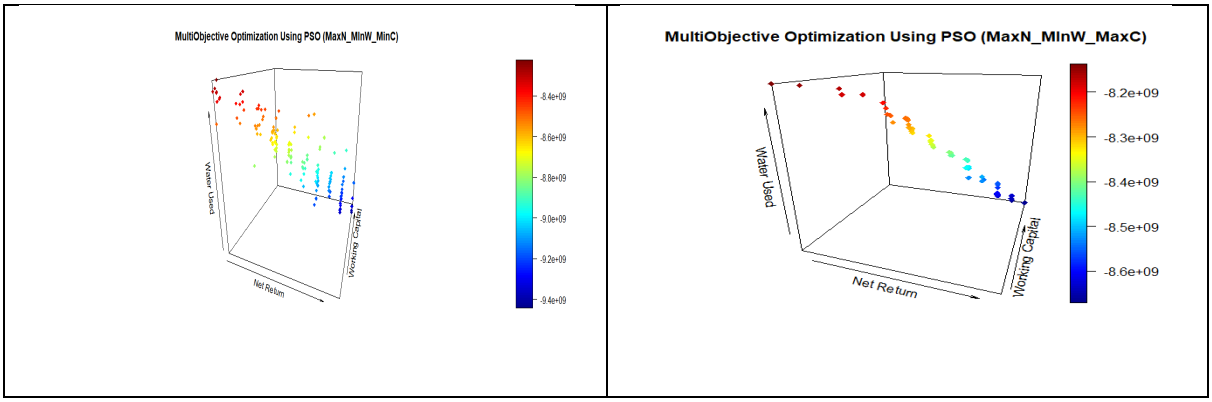


Figure 6: **Number of solutions obtained from PSO based optimization for three objectives**

5. Conclusions and Future Scope

Researchers have attempted various techniques for optimum cropping plans at various levels (farm, region, state, country) with various objectives and constraints depending on the resource availability and objectives of the farms or region. These models have been set up as a research tool or a teaching aid only as farmers and the end-users have directly used very few models. By using the versatile models presented in this paper, extension services can create awareness and give advisory services under expert guidance based on the available resources of individual farmers and the regions. This paper strengthens the opinion (Nath *et al.*, 2020) that the use of evolutionary computing for optimization in crop planning needs to be explored further and algorithms may be modified to suit the specific needs of crop planning.

References

Adeyemo, J. and Otieno, F. (2010). Differential evolution algorithm for solving multi-objective crop planning model. *Agricultural Water Management*, **97**(6), 848-856.

Chand, P., Jain, R., Chand, S., Kishore, P., Malangmeih, L. and Rao, S. (2020). Estimating water balance and identifying crops for sustainable use of water resources in the Bundelkhand region of India. *Transactions of the ASABE*, **63**(1), 117-124.

Coello, C. A. C., Pulido, G. T. and Lechuga, M. S. (2004). Handling multiple objectives with particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, **8**(3), 256-279.

Gill, M. K., Kaheil, Y. H., Khalil, A., McKee, M. and Bastidas, L. (2006). Multi-objective particle swarm optimization for parameter estimation in hydrology. *Water Resources Research*, **42**(7), at https://cran.r-project.org/web/packages/mopsocd/mopsocd.pdf.

GoI (2015). *Study on Bundelkhand - Concerns in Development and Issues for Action.* Planning Commission, Government of India, accessed from http://planningcommission.nic.in/ reports/sereport/ser/bndel/stdy_bndel.pdf.

GoMP (2018). *Statistical abstract of Madhya Pradesh*. Economic and Statistics Division, State Planning Institute, MP, available at mpdes.mp.nic.in.

GoUP (2018). *Statistical abstract of Uttar Pradesh*. Economic and Statistics Division, State Planning Institute, UP available at updes.up.nic.in.

Hosseini, M., Naeini, S. A. M., Dehghani, A. A. and Khaledian, Y. (2016). Estimation of soil mechanical resistance parameter by using particle swarm optimization, genetic algorithm and multiple regression methods. *Soil and Tillage Research*, **157,** 32-42.

Jain, R., Chand, P., Rao, S. C. and Agarwal, P. (2020). Crop and soil suitability analysis using multi-criteria decision making in drought-prone semi-arid tropics in India. *Journal of Soil and Water Conservation*, **19**(3), 271-283.

Jain, R., Kingsly, I., Chand, R., Raju, S. S., Srivastava, S. K., Kaur, A. P. and Singh, J. (2019). Methodology for region level optimum crop plan. *International Journal of Information Technology*, **11**(4): 619-624.https://doi.org/10.1007/s41870-019-00330-w.

Jain, R., Kingsly, I., Malangmeih, L., Deka, N., Raju, S. S., Srivastava, S. K., Kaur A. P. and Singh, J. (2018a). Linear programming based optimum crop mix for crop cultivation in Assam state of India. *In International Conference on Intelligent Systems Design and Applications* (*pp. 988-997*). Springer.DOI:10.1007/978-3-319-76348-4_95.

Jain, R., Malangmeih, L., Raju, S. S., Srivastava, S. K., Kingsley, I. and Kaur, A. P. (2018b). Optimization techniques for crop planning: A review. *Indian Journal of Agricultural Sciences*, **88**(12),1826-1835.

Jain, R., Kingsly, I., Chand, R., Kaur, A. P., Raju, S. S., Srivastava, S. K. and Singh, J. (2017). Farmers and social perspective on optimal crop planning for groundwater sustainability: a case of Punjab state in India. *Journal of the Indian Society of Agricultural Statistics*, **71**(1), 75-88.

Jain, R., Raju, S.S., Kingsly I., Srivastava, S.K., Kaur, A. P. and Singh, J. (2015). *Manual on Methodological Approach for Developing Regional Crop Plan.* ICAR-National Institute of Agricultural Economics and Policy Research (available at https://niap.icar.gov.in).

Kennedy, J. (2010). Particle swarm optimization. *Encyclopaedia of Machine Learning*, 760-766.

Kennedy, J., Eberhart, R. C. (1995). Particle swarm optimization. *Proceedings of the International Conference on Neural Networks; Institute of Electrical and Electronics Engineers*, **4**, 1942-1948. DOI: 10.1109/ICNN.1995.488968.

Kumar N., Raju D. K. S. and Ashok B. (2006). Optimal reservoir operation for irrigation of multiple crops using genetic algorithms. *Journal of irrigation and drainage engineering*, **132**(2),123-129.

Nath, K., Jain, R., Marwaha, S. and Arora, A. (2020). Identification of optimal crop plan using nature inspired metaheuristic algorithms. *Indian Journal of Agricultural Sciences*, **90**(8), 1587-92.

Naval, P. (2013). MOPSOCD: Multi-objective particle swarm optimization with crowding distance. *R Package Version* 0.5, 1.

Sethi, L. N., Panda, S. N. and Nayak, M. K. (2006). Optimal crop planning and water resources allocation in a coastal groundwater basin, Orissa, India. *Agricultural Water Management*, **83**(3), 209-220.

Sharma, D. K., and Jana, R. K. (2009). Fuzzy goal programming based genetic algorithm approach to nutrient management for rice crop planning. *International Journal of Production Economics*, **121**(1), 224-232.

Wang, X., Yang, G., Feng, Y., Ren, G., and Han, X. (2012). Optimizing feeding composition and carbon-nitrogen ratios for improved methane yield during anaerobic co-digestion of dairy, chicken manure and wheat straw. *Bioresource Technology*, **120**, 78-83.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 143-160

Rural Credit: Can We Make it More Inclusive?

K. J. S. Satyasai and Abhishek Tiwari
NABARD, Mumbai

Received: 16 July 2021; Revised: 14 August 2021; Accepted: 20 August 2021

Abstract

Exclusion from institutional rural credit network has been on at least three fronts: geography, gender and farm households. This paper discusses these three variants of exclusion, interventions and policies that were implemented over the decades, progress report so far and way forward. Credit has been oft reported constraint for farmers. Access to credit is not uniform across geographies, farm households, and gender. Small farmers need more outside funds for their farming and family expenditure due to lower savings and, ironically, they have restricted access to credit. It was argued during late 1980s, using data on distribution of credit and landholding area across farm size categories, that there was an institutional bias in favour of small and marginal farmers as the share of small farmers in credit exceeded their share in operational area. It goes to the credit of successive institutional efforts, following multi-agency approach, starting from cooperative movement, nationalization of banks in two batches, establishment of Regional Rural Banks (RRBs), Self-help Groups (SHGs), Joint Liability Groups (JLGs), new generation institutions such as Micro-Finance Institutes (MFIs) and Small Finance Banks (SFBs) topped by directed credit policy driven by Priority Sector Guidelines issued by Reserve Bank of India (RBI) from time to time. Lower credit penetration in geographies like North Eastern region, Eastern and Central regions is well known. Some of the initiatives take for inclusion such as SHG-Bank linkage programme has been more successful in Southern and Western regions who have higher share in total credit. Infrastructure bottlenecks such as poor road network, lack of irrigation facilities, poor communication network, etc. could have played major role. Even among the well-developed regions of the country, small and marginal farmers are disadvantaged in terms of credit access. Access to institutional credit remains poor with hardly 20% of the 12.56 crore small and marginal farmers. Tractor financing and the like are mostly through NBFCs and we hardly have any evidence to understand the inter-farm size equity in credit dispensation from these agencies. Tenants who dominant the agriculture space are hardly included in the institutional credit system. Though banks and MFIs have supported rural people through SHG and JLG-modes, the total credit disbursed is hardly 5% of the annual ground level credit purveyed. Thus, most rural people have availed loans from informal sources including private moneylenders which often comes at high rates of interest. Among the rural people, access of women to credit has been low and insignificant. Importantly, we hardly have any hard data on credit to women. While a lot of efforts went into linking women to banks through SHGs under the programmes of NABARD and Ministry of Rural Development (MoRD), the progress is yet to be impactful. The problem of exclusion is very difficult to address as the marginal cost of reaching out to new clients will be much higher for the financing institutions. And, these clientele groups need small ticket loans. Infrastructure creation, leveraging digital technology for internet/mobile banking, digital payment system, *etc*., innovative credit products to suit the low-end clientele are some of the ways around the problem.

Key words: Disparity in credit distribution; Herfindahl Index; Panel regression; Gender gap.

Corresponding Author: K.J.S. Satyasai
Email: satyasaik@outlook.com

1. Introduction

The relationship between agriculture credit and development in agriculture sector from a macro perspective is captured in various studies. Multiple studies have documented positive relationship between the credit and development of the sector (Demetriades and Hussein, 1996; Arestis and Demetriades, 1997; Khan, 2001; Rajan and Zingales, 2003). Over the period of time discussion on rural credit and development has evolved to incorporate various dimensions like accessibility, productivity, and disparity in credit distribution.

Credit is a critical facilitator that enables adoption of technology, and higher input use in agriculture as well as an effective means of rural development. Various agencies, including commercial bank, regional rural bank (RRBs), co-operatives, small finance banks (SFBs), NBFCs, micro-finance institutions (MFIs) and indigenous bankers together form the rural credit delivery system in India. Since independence government of India (GoI), RBI and other financial institutions are trying to minimise the share of non-institutional credit in agriculture. Over the years, multiple committees have been formed to recommend ways to increase the institutional credit in agriculture and rural areas. These include the R. V. Gupta Committee on Agricultural Credit through Commercial Banks, Vikhe Patil Committee on Co-operatives, V.S. Vyas Advisory Committee on Flow of Credit to Agriculture, and A. Vaidynathan Task Force on Revival of Co-operative Credit Institutions. The GoI has accepted most of the recommendations and brought reforms suggested by these committees. In this direction several initiatives have been taken time to time, *e.g*., accepting Rural Credit Survey Committee Report (1954), nationalisation of the commercial banks in 1969 and 1980, establishment of RRBs (1975), establishment of National Bank for Agriculture and Rural Development (NABARD) (1982), *etc.*

Apart from institutional changes, GoI, RBI and NABARD have brought multiple schemes to increase credit penetration in agriculture like establishment of Lead Bank Scheme, provision of priority sector lending, self-help group – bank linkage programme (SHG-BLP), lending to joint liability group, kisan credit card scheme, Rural Infrastructure Development Fund (RIDF),*etc*. to increase credit penetration in rural area. After the formation of NABARD, credit flow to agriculture sector has gone up significantly. Total credit flow in agriculture was just ₹4,352 crore in 1982-83 which has increased to ₹15,58,831 crore.

However, despite significant jump in credit flow to agriculture sector, disparity in credit distribution is observed on three front, regional, gender and farm landholding. Empirical data suggest that Southern region get almost half of the total agriculture credit flow and reach of institutional agencies have remained poor to the small and marginal farmers.

The significant growth and coverage of the MFIs in the country since new economic reforms plus the implementation of Interest Subvention Scheme or KCC scheme have also failed to cover many credit starved region, as evident from the regionally skewed distribution credit disbursement. For example, NER comprising of seven sister states and Sikkim, account for 8 percent of total national area, 3.7 percent of total population and 4.3 percent of net sown area, the share of ground level credit in agricultural has remained just 1.1 percent in 2020-21, despite overall CAGR of agriculture credit at the national level is in double digit during the period from 2013-14 to 2020-21 (NABARD).

Also, empirical evidence suggests that share of female farmers in the total agriculture credit flow is miniscule. For a scientific and empirical analysis of agriculture credit delivery,

it is important to examine disparity in credit distribution at micro level to distinguishing characteristics of the agricultural and rural households. Small farmers need more outside funds for their farming and family expenditure due to lower savings and, ironically, they have restricted access to credit. It was argued during late 1980s, using data on distribution of credit and landholding area across farm size categories, that there was an institutional bias in favour of small and marginal farmers as the share of small farmers in credit exceeded their share in operational area. Due to lack of access to formal credit, marginalised and weaker sections of society, women, small farmers *etc*. are forced to borrow from indigenous banker and exhaust their small savings. This has put many female household and small and marginal farmers in vicious circle of poverty, unemployment, and income inequality. This analysis would be useful in understanding the reasons for disparity in agriculture credit flow. This will also help in reorienting the agriculture credit policies and programmes for a better impact.

The observation that accessibility of finance is more acute underdeveloped region when compared to the developed region developed regions required a micro look at the underlying causes of credit distribution across states in developing economies like India. Against these backdrops, this study is undertaken to assess (i) Regional disparity in credit distribution, (ii) factor affecting the regional disparity in distribution of credit, (iii) gender-based disparity in agriculture credit distribution and (iv)disparity in credit distribution based on farm landholding.

The paper has been divided into six sections. The following section provides literature review. Section 3 gives brief description of data and the methods of analysis. Section 4 gives an overview of the performance of agriculture credit flow in last decade. Section 5, discuss about the findings of statistical analysis. Conclusion and policy implication are discussed in the last section.

2. Review of Literature

(Aportela, 1999) suggests that access of financial services increases savings, consumption and high investment (Dupas and Robinson, 2009). While lack of access to financial services may lead to a vicious circle of poverty and increase in the inequality (Banerjee and Newman, 1993; Galor and Seira, 1993; Aghion and Bolton, 1997; Beck *et al*, 2007). Many studies have shown that the level of financial inclusion inevitably rises in response to prosperity and declining inequalities (Sarma, 2008; Rajesh and Das, 2019). (Kumar *et al*., (2010), highlighted that the institutional credit has been conceived to play a pivotal role in the agricultural development of India. The empirical studies have highlighted that large number of institutional agencies are involved in the disbursement of credit to agriculture. However, the persistence of money lenders in the rural credit market is still a major concern.

Modern empirical studies confirm that economic development is not possible without credit(King and Levine, 1993; Levine *et al.,* 2000 and Beck and Levine 2004). Hulme and Mosely (1996) highlighted in their study that credit could be the important tool against urban and rural poverty as it helps to create long term asset to those who do not have capacity to finance large investment through their savings. However, many small and marginal farmers, people from marginalised section and small-scale entrepreneurs are deprived of access to credit in emerging countries like India, which is evident from a much lower CD ration per 1000 adults. Worldwide studies by various agencies have indicated that developed economies have the largest number of loans per 1000 people. Shetty (1997) highlighted that credit absorptive capacity of economy is one of the important instruments to achieve high ground

level credit flow. Credit absorptive capacity depends upon the infrastructure development, transportation facilities, banking infrastructure, public expenditure etc. World-wide as well as in India not many attempts have made to measure the disparity in the distribution of ground level credit in agriculture.

In this milieu, the present study has examined the performance of agricultural credit flow and has identified the determinants of increased use of institutional credit at the farm household level in India. The study based on the secondary data has revealed that the institutional credit to agriculture in real terms has increased tremendously during the past four decades. The structure of credit institutions has witnessed a substantial change and commercial banks have arisen as the major contributor in institutional credit in recent years. The quantum of institutional agri-credit availed by the farmers is affected by various socio-demographic factors such as education, landholding, family size, caste, gender, *etc*. The study has suggested that reduction in procedural complications may lead simplification of the procedure for a better access to agricultural credit of smallholders and less educated/illiterate farmers. Satyasai (2012) in his findings brought out that inequalities in the distribution of number of loans vis-à-vis operational holdings have increased over time. Chavan (2020) highlights that gender gap in credit access is significantly high.

3. Data and Methods

In this paper, the term "ground level credit" implies the total credit disbursement in agriculture at the end of the financial year. We employed the secondary data on state-wise distribution of number of credit accounts and amount of credit disbursement by various agencies during 2009-10 to 2020-21. This decade represents period of intensive policy-push through various measures to increase GLC in agriculture. Secondary data sourced from government/authentic reports published by the RBI, `Registrar General and Census Commissioner of India, the National Statistical Office and NABARD. All the 29 States (including united J and K as state)data have been included in the analysis. Union Territories are excluded due to their special features which are likely to affect the comparability of datasets.

Herfindahl Geographic Concentration Index

Ellison and Glaeser (1997), reformulated the original Herfindahl index to calculate geographical concentration index. The Herfindahl index is the one of the most used indices to measure geographic concentration (Spieza, 2002). This index takes into account differences in the gross cropped area. The formula appears as follows:

$$EG = \sum_{i=1}^{n} (Y_i - a_i)^2$$

where Y_i is the credit proportion of region/state, n stands for the number of regions/states being compared; and a_i is the gross cropped area of region i as a proportion of the total gross cropped area. If the credit share of each region equals its relative area, then there is no concentration and EG equals 0 indicating no regional disparity.

The EG index is the sum of all n squares for the entire nation. It shows the extent of credit disparity in agriculture among n states. Each of these squares would be a decimal number or a fraction, and the EG index is also most likely to be a decimal or fractional

number. Thus, to apply this formula to calculate the proportion of regional disparity that is contributed to the total by each state, the ratio of the square to EG is used. The formula would appear as follows:

$$(Y_i - a_i)^2/EG$$

Theil's Index

To get deeper understanding of regional disparity in ground level credit (GLC) distribution, weighted Theil's Index (*T*) has also been calculated. Theil's index measures inequality among 'm' regions using the formula given below (Conceição and Ferreira, 2000).

$$T = \sum_{i=1}^{m} w_i \, Log \frac{W_i}{n_i}$$

where,

w_i: Share of the *i*th state in agriculture credit to total agriculture flow in agriculture,

n_i: Share of the *i*th state in gross cropped area to total gross cropped area,

m: denotes region

The Index close to zero indicates a more equitable coverage.

Credit growth convergence

Before analysing the causes for regional disparity in ground level credit across states, we made attempt to check whether growth in GLC is converging across the states over the years. β-convergence method is used to empirically investigate the disparity in credit distribution in agriculture across Indian states. The fixed effect panel regression consisted of all 29 states (Jammu & Kashmir and Ladakh as considered one state) for 2011-12 to 2020-21 with eight-year average credit growth as a dependent variable and the base period credit growth as an independent variable. The coefficient of β found positive and statistically insignificant. The finding suggests no statistical evidence of convergence of credit growth of agri-credit across states over time.

Panel regressions

Panel regression is used to understand the factor affecting disparity in the agriculture credit distribution across the states. This will help in understanding the causes of disparity which can be helpful in formulating the policy to bridge the credit gap across the states. Multi-variables are used to explain divergence in agriculture credit across states and their economic rationale discussed as well. We examine whether the high agriculture growth of a particular state helps in attracting more credit to the sector in that particular state and vice versa. Second, impact of share of gross cropped area, irrigation facility, rural road, and electricity consumption for agriculture over ground level credit in agriculture. These variables used as a proxy of higher credit absorption capacity across the states.

Further, state-wise rural bank branches are taken to represent the availability of banking infrastructure and penetration of the banking services in the states. One of the key features of

this paper is the introduction of infrastructure in rural areas as explanatory variables for agriculture credit in state. The crucial role played by infrastructure in economic development has been well established in academic literature for a long time (Hirschman, 1958 and Rostow,1960). Better infrastructure facilities like transport, communication and power help in enhancing the productivity of investment in that region, which in turn propels competitiveness.

We employ a generic panel data model with ground level credit in agriculture as a dependent variable and function of agriculture gross value added, gross cropped area, bank branches in rural area, rural CD ratio, power consumption in agriculture, fertiliser consumption and irrigation coverage as independent variables. All the explanatory variables have been suitably standardized taking into account the size of the state. The equations are:

$$Y_i = a + b_1X_{i1} + b_2X_{i2} + b_3X_{i3} + b_4X_{i4} + b_5X_{i5} + b_6X_{i6} + b_7X_{i7} + e_i$$

where Y_i = Agriculture GLC, X_{i1} = Agriculture GVA, X_{i2} = Rural Bank Branch, X_{i3} = Gross Irrigated Area, X_{i4} = Rural Electricity Consumption, X_{i5} = Gross Cropped Area, X_{i6} = Rural CD ratio, X_{i7} = Fertiliser consumption, and e_i = error term.

4. Performance of Credit Flow in Last Decade

In last one decade agriculture credit has increased by more than 3 times. During 2020-21, the institutional credit flow to agriculture sector in India was to the tune of Rs.15.58 lakh crore, including Rs.8.85 lakh crore of short-term credit and 6.73 lakh crore of long-term credit. The share of long-term credit in total institutional credit flow to agriculture has been rising steadily and exceeds 40% mark in 2018-19. The share of LT credit which stood at 22.48% in 2011-12 has increased to 43.08% in 2020-21.

The declining trend seen in long term credit before 2014-15 was of concern to the policy makers as term credit purveyed to finance long term investments led to private capital formation in farm mechanisation, minor irrigation structures including pump sets, land development, orchards, farm ponds, micro-irrigation, *etc*., in the country. As we know, long term credit has been the major driver of the private sector capital formation in agriculture (PSCFA). There exists a high correlation between long term credit and private investment. This can also be seen in increasing share of private sector in Gross Capital formation in agriculture sector to 80% in 2017-18 from 56% in 1980-81. Driven by the understanding of this relationship and the fact that investment on the farm is indispensable for enhancing production as also building productive capacity on the farm, NABARD and other stakeholders made concerted efforts and the trend started reversing, reaching 43 per cent by 2020-21.

There was a year-on-year growth of 11.9 percent in the total credit flow in 2020-21, mainly on the strength of the 18.6 percent rise in the term loan whereas the y-o-y growth of short-term credit flow increased by 7.4 percent. There has been steady progress in the flow of institutional credit to agriculture sector over the years and the compound annual growth rate (CAGR) between 2010-11 and 2020-21, is 12.9 percent (Table 1), which means total credit flow has doubled in 6 years. During the period 2011-12 to 2020-21, the long-term credit, which adds to the capital formation in the agriculture and allied sector, increased at a faster rate (CAGR 20.7%) as compared to the short-term credit (CAGR 9.4%). Data is showing that term loan has doubled in less than 4 years in last decade while crop loan has doubled in almost 8 years. High growth in term loan is indication of higher capital investment

in the sector and greater mechanisation of the agriculture which may lead to higher production and productivity. Apart from increase in term loan, crop loan has also gone up significantly and helping farmers to get rid of moneylenders.

Kisan Credit Card is playing important role in increase in crop loan. Till 31st March 2019, total 1896 lakh kisan credit card issued which helped in increasing the cheap institutional credit to masses. GoI started Kisan Credit Card (KCC) saturation drive for publicity and awareness campaigns across districts to cover all beneficiaries under the Pradhan Mantri Kisan Samman Scheme to increase credit penetration throughout the country.

Table 1: Ground level credit flow to agriculture – ST / LT disbursement

Year	Short Term			LT/MT			Total GLC	
	Amount (₹ crore)	Annual growth (%)	% to total	Amount (₹ crore)	Annual growth (%)	% to total	Amount(₹ crore)	Annual growth (%)
2011-12	3,96,158	15.3	77.52	1,14,871	–15.6	22.48	5,11,029	9.1
2012-13	4,73,500	16.3	77.96	1,33,875	14.2	22.04	6,07,376	18.9
2013-14	5,48,435	13.7	75.12	1,81,687	26.3	24.88	7,30,123	20.2
2014-15	6,35,412	13.7	75.17	2,09,916	13.5	24.83	8,45,328	15.8
2015-16	6,65,313	4.5	72.67	2,50,197	16.1	27.33	9,15,510	8.3
2016-17	6,89,457	3.5	64.69	3,76,298	33.5	35.31	10,65,756	16.4
2017-18	7,53,214	9.5	64.61	4,09,402	9.9	35.39	11,62,616	9.6
2018-19	*7,52,209*	-0.1	59.85	*5,04,620*	23.3	40.15	*12,56,829*	9.1
2019-20	8,25,151	9.7	59.25	5,67,579	12.5	40.75	13,92,730	10.8
2020-21P	8,85,811	7.4	56.83	6,73,020	18.6	43.17	15,58,831	11.9
CAGR	**9.4**			**20.7**			**12.9**	
P = Provisional Data								

Source: *Handbook of state statistics and NABARD Annual Report*

With the concerted efforts of financing banks, operationalisation of Small Finance Banks, refinance support from NABARD under Long Term Rural Credit Fund (LTRCF) to RRBs and RCBs etc. the investment credit in agriculture sector has been overachieving the targets for past four consecutive years. It is heartening to note that Banks have been overachieving the long-term credit targets since 2016-17. Banks could hardly achieve 58 % of the LT target in 2012-13, however, since then the same has increased to 131% of the target in 2018-19, 117 percent in 2019-20 and 118 percent in 2020-21 of target has been achieved.

Table 2: Target and Achievement under Investment Credit (Amount in ₹ Crore)

Year	Target	Achievement	Target Achievement (%)
2011-12	1,95,000	1,14,871	59
2012-13	2,30,000	1,33,875	58
2013-14	2,00,000	1,81,687	91
2014-15	2,25,000	2,09,916	93
2015-16	2,55,000	2,50,197	98
2016-17	2,85,000	3,76,298	132
2017-18	3,20,000	4,13,530	129
2018-19	3,85,000	5,04,620	131
20-2019	4,86,675	5,67,579	117
2020-21	5,70,000	6,73,020	118

Source: *NABARD Annual Report*

4.1. Market shares of rural financial agencies in GLC

Commercial Banks have maintained their three-fourth share in the total agriculture credit and the rest about one-fourth was shared between the RRBs and RCBs. The RCBs have been steadily losing their share in GLC over time from 40 percent in 1999-2000 to 12 percent in 2020-21. Commercial Banks (CBs) enhanced their Agri-credit growth at a faster pace since 2004-05 (GoI announced doubling of credit flow) and increased their share from 53.7 percent in 1999-2000 to 75.8 per cent in 2020-21. RRBs accounted for the remaining share of 12.1 per cent in 2020-21, which is an improvement from their share of 6.9 per cent in 1999-2000. The agency wise share indicates that agriculture credit dispensation in the country is heavily dependent on commercial banks and points towards the poor credit delivery capability of RCBs and RRBs.

Table 3: Share of various agencies in total agricultural GLC(Amount in ₹ crore)

Year	Commercial Banks		RRBs		Cooperative Banks		Total GLC
	Amt.	% to total	Amt.	% to total	Amt.	% to total	Amount
1999-2000	24,836	53.7	3,172	6.9	18,260	39.5	46,268
2004-2005	81,674	65.2	12,404	9.9	31,231	24.9	1,25,309
2009-2010	2,85,800	74.3	35,217	9.2	63,497	16.5	3,84,514
2013-2014	5,27,506	72	82,653	11	1,19,963	16	7,30,122
2014-2015	6,04,376	71.5	1,02,483	12.1	1,38,469	16.4	8,45,328
2015-2016	6,42,954	70.2	1,19,261	13	1,53,295	16.8	9,15,510
2016-2017	7,99,781	75	1,23,216	11.5	1,42,758	13.5	10,65,755
2017-2018	8,77,155	75	1,40,959	12.1	1,50,389	12.9	11,68,503
2018-2019	9,54,822	76.0	1,49,666	12.1	1,52,340	12.3	12,56,829
2019-20	10,70,036	76.8	1,65,326	11.9	1,57,367	11.3	13,92,729
2020-21	11,81,558	75.8	1,89,505	12.2	1,87,769	12	15,58,831

Source: *NABARD*

The Share of term loan to total agriculture credit disbursed by RRBs during the last six years has increased from 12.8% to 16.4%. Strive for achieving the term loan targets to give boost to long term investments in agriculture, GoI has set up Long Term Rural Credit Fund (LTRCF) with NABARD for providing refinance support to RRBs and Cooperatives. The current interest rate on refinance under LTRCF is lower as compared to rate under normal refinance and this benefit may be passed on to ultimate borrowers.

Apart from SCB, RRBs and RCBs, NBFCs, micro-finance institutions and small finance banks have emerged as new force in agriculture and rural financing. Recently robust growth has been witnessed in credit disbursed by NBFCs to the agriculture sector. Credit by NBFCs have increased by CAGR of 22 % between 2015-16 to 2018-19 while the growth rate witnessed in total ground level credit to agriculture was 11% during the same period. NBFCs credit to agriculture has increased by 79 % between the years 2015-16 and 2018-19. This points towards development of deeper reach into the rural areas by the NBFCs. NBFCs also may score over banks due to their ability to provide a more hassle-free processes and better technology adoption. Small Finance Banks (SFBs) were set up in 2016 to offer basic banking services such as accepting deposits and lending to the unserved and the under-served

sections, including small businesses, marginal farmers, micro and small industries, and the unorganised sector. Small finance banks are spreading well in semi-urban areas while spread in rural areas are still slow.

5. Analysis

5.1. Regional imbalance in credit dispensation

Regional imbalance in the distribution of agriculture credit has persisted over the years. In 2020-21, Southern Region had the largest share (45.9%) followed by the Northern Region with almost half of the former (17.1%). Incidentally, the share of Southern Region in the total agriculture credit flow has increased, whereas the share of northern region has decreased from 2014-15 to 2020-21. This indicates a persistence and deeper regional imbalance of the credit flow across regions. In year 2020-21, north-eastern region has witnessed highest YoY growth of 39 percent, followed by western region (16.8%). Northern region has witnessed decline in ground level credit in agriculture by 8 percent compared to 2019-20. This will lead to further disparity in credit distribution.

The southern region has higher credit absorption capacity may be because of better infrastructure facilities, better outreach and credit availability leading to improvement in its share. Normally, low density of credit delivery outlets and weak financial health of Rural Financial Institutions could be the constraints for increasing credit flow in credit starved regions like Eastern and Central states. However, central and eastern regions account for 21 % and 19% share in rural and semi urban bank branches respectively pointing towards demand side bottlenecks in the regions.

Data further indicates that there is a growing disconnect between the real sector parameters and regional distribution of agriculture credit**.** For example, the Eastern Region has 11.7 % share in Net Sown Area and 11 per cent in NIA, but hardly accounted for 9.5 per cent of agriculture credit disbursed during the year 2019-20. In contrast, the southern region accounted for 19.6 per cent of Net Sown Area and 16.6 percent of NIA, but availed the highest share (45.9 %) of GLC disbursed during 2020-21.

Table 4: Regional distribution of agriculture credit and real sector indicators (%)

Regions	Share in total agricultural GLC (%)					Real Sector Indicators (%)		
	2018-19	2019-20	Share	YoY Growth (2020-21)	Share in Net Sown Area	Share in NIA	Share in Food grain production#	Share in rural/semi urban branches
Northern	21.7	20.4	17.1	–8.1	18.9	22.6	24.9	16.9
N.E.R.	0.9	0.8	1.1	39.3	3.27	1.0	3.0	3.7
Eastern	8.7	9.1	9.5	10.2	11.7	11.0	16.5	18.6
Central	13.7	14.1	14.4	14.1	26.6	37.7	32.5	20.9
Western	12.3	11.2	12.0	16.8	19.8	40.6	7.4	12.4
Southern	42.8	44.4	45.9	14.1	19.6	16.7	15.7	27.5
Note: (a) # denotes share for the period 2017-18 (b) Net Sown Area as per 2014-15, NIA- Net irrigated Area as per 2014-15								

Source: *Calculated based on data from MOA, RBI, SLBC and NABARD*

Table 5 shows that Kerala received highest loan per hectare (6.63 lakh) followed by Tamil Nadu (3.19 lakh) and West Bengal (2.07 lakh). In all 29 states (including united J and K), only 9 states have per hectare loan amount more than ₹ 1 lakh.

Table 5: Average disbursement per account (₹ Lakh)

States	All Farmers Accounts	Loan per hectare (NSA)
Haryana	1.9	1.76
Himachal Pradesh	1	0.73
Jammu and Kashmir	1.4	1.56
Punjab	2.2	1.96
Rajasthan	1	0.4
Arunachal Pradesh	1	0.01
Assam	0.7	0.24
Manipur	1.1	0.14
Meghalaya	0.6	0.07
Mizoram	3.3	0.32
Nagaland	0.5	0.02
Sikkim	1.2	0.18
Tripura	0.7	0.95
Bihar	0.7	0.55
Jharkhand	0.5	0.13
Odisha	0.5	0.59
West Bengal	0.8	2.07
Chhattisgarh	0.6	0.21
Madhya Pradesh	0.8	0.39
Uttarakhand	1.4	1.38
Uttar Pradesh	0.9	0.51
Goa	2.1	1.61
Gujarat	1.7	0.66
Maharashtra	1.3	0.41
Andhra Pradesh	1	1.53
Telangana	1.2	0.96
Karnataka	0.8	0.62
Kerala	1.1	6.63
Tamil Nadu	0.8	3.19

Source: *Calculated by authors from NABARD data*

In the above section, disparity in credit distribution is indicated in absolute form. To get the better understanding of the relative disparity, weighted Theil's and modified Herfindahl index is calculated. In the calculation, ground level distribution of credit is weighted with gross cropped area.

5.2. Result of Theil and Herfindahl Index

As illustrated in Table 6, the southern states make the highest contribution to the overall regional disparity as per Theil's and EG index. This means southern region is contributing about 58 % of the regional disparity in credit distribution for the entire country in agriculture. Even though North-eastern region is showing low disparity in EG index, however as per Theil's index, disparity in credit distribution in the region is relatively more.

Table 6: Theil's and Modified Herfindahl Index

S. No.	Region	Theil Index	EG Index
1	Southern	0.21	0.58
2	Northeast	0.19	0.003
3	Northern	0.11	0.02
4	Central	0.05	0.13
5	Western	0.03	0.02
6	Eastern	0.01	0.003
7	India	0.33	0.17
Note: Theil's Coefficient and EG index is estimated based on GLC data from Ensure (GLC data)			

5.3. Factor affecting disparity in agriculture distribution

Panel regression result and correlation among the variables are suggesting the plausible causes of regional disparity in the credit distribution. The list of variables is chosen based upon availability of data and their plausible impact on ground level credit in agriculture. Rural bank branches and CD ratio explains the banking services in the rural area while fertiliser consumption and electricity consumption are taken as input and rural infrastructure helping the spread of banking facility. Table 7 suggests that GLC is highly correlated to rural branches, CD ratio and electricity consumption among all the variables.

Table 7: Correlation Coefficient

Correlation	GLC	Agri-GVA	GCA	Fertiliser	Bank Branches	GIA	Electric
GLC	1.0000						
Agri-GVA	0.6651	1.0000					
GCA	0.5694	0.9262	1.0000				
Fertiliser	0.5655	0.4666	0.3122	1.0000			
Bank Branches	0.7864	0.7564	0.6720	0.5496	1.0000		
GIA	0.5559	0.8744	0.8582	0.4446	0.7186	1.0000	
Electric	0.7713	0.7658	0.8004	0.4765	0.6555	0.5914	1.0000
Rural Account	0.5511	0.7743	0.6716	0.4422	0.8623	0.8052	0.4326
CD Ratio	0.7081	0.3815	0.3747	0.5446	0.4316	0.2879	0.6791

Panel regression of last one decade suggests that rural bank branches, electricity consumption in rural area and CD ratio has significant positive impact on the agriculture credit (Table 8).As result suggests that agri-GVA, gross cropped area and fertiliser consumption do not have significant impact over credit distribution.

Table 8: Panel Estimates Explaining Credit Disbursement

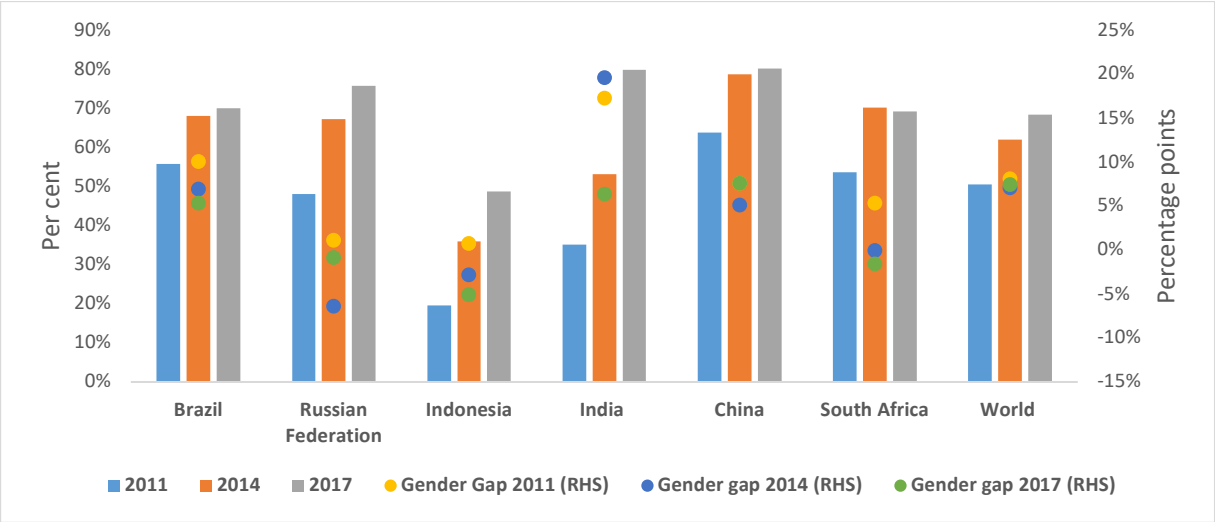
Log Variable	Estimate
Log Agri-GVA	0.0240 (0.26)
Log Fertilisers Consumption	–0.146* (–2.02)
Log Rural Bank Branches	1.088*** (10.35)
Log Electricity Consumption	0.149*** (4.57)
Log rural CD ratio	0.548*** (4.80)
***N* = 173, *t* statistics in parentheses,** *–p<0.05, **–p<0.01, *** –p<0.001	

5.4. Gender disparity in credit distribution

India's achievement in providing access of basic banking services is not very encouraging in compare to the other emerging countries in the world. The term gender gap in credit distribution indicates to the gap of access to women vis-a-vis men.

5.4.1. Gender gap in deposits

After the launch of Pradhan Mantri Jan Dhan Yojana, women share in total bank account has increased significantly, which has led to decrease in gender gap in availing the financial services. However, the gap is still very significant and it is highest among the BRICS nations (Figure 1).

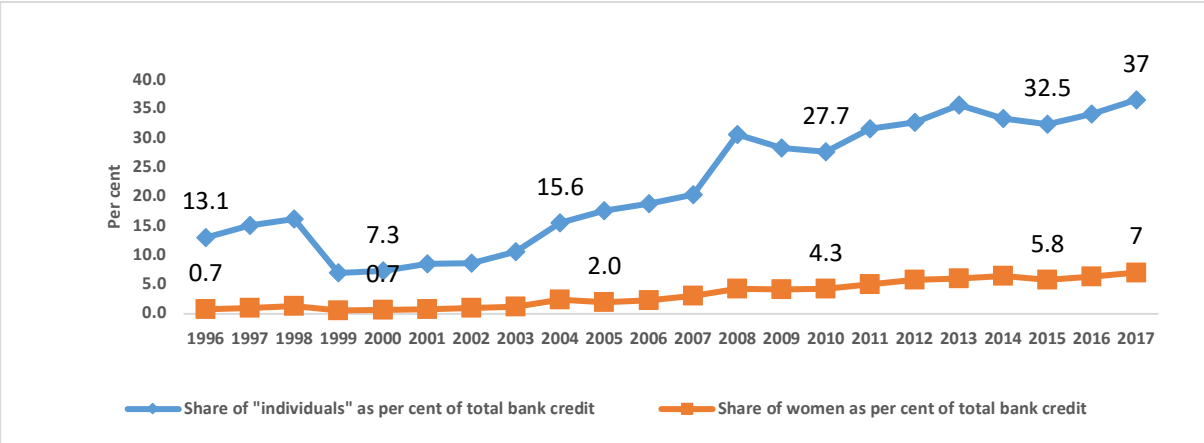


Source: *Trends in women banking- Pallavi Chavan*

Figure 1: Adult population with deposit accounts in financial institutions(in per cent)

5.4.2. Gender gap in availing credit

As data suggests globally, India is still lagging behind by distance with respect to the access to bank credit for its women. As per World Bank findex data, only 5 per cent of Indian women accessed bank credit. This shows that while women deposits have gone up over the period of time but they are still not getting credit from bank due to various reasons. As figure 2 is suggesting that women's share in bank credit is increasing over the period of time but at a very slow pace. Women share in 1996 was just 0.7 percent which has gone up to just 7 percent in 20 years as against 30 per cent for men. NABARD Shelf-Help Group- Bank Linkage Programme (SHG-BLP) helped rural women to avail credit but the pace is not very encouraging. Even after including the credit going to MFIs, and joint liability groups (JLGs)/trusts/non-governmental organisations (NGOs) as credit to women, there was only a marginal increase to 8 per cent in women's share.



Source: *BSR, RBI.*

Figure 2: Share of women in total bank credit in India (in per cent)

5.5. Disparity based on landholding

As per agri-census 2015-16, 86.58 percent farmers are small and marginal farmers holding less than 2 hectares of land. Apart from that in Indian around 14 crore cultivators are lessee, sharecroppers or tenant farmers.

Table 9 indicates that small and marginal farmers were getting just 44 percent of total credit disbursement in agriculture sector, which has slightly increased to 52 percent in 2019-20. While medium and large farmers who are just 14 percent of total farmers are getting 48 percent of the total agriculture credit. Main reason behind the low credit disbursement to small and marginal farmers is that decrease in share of RRBs and Cooperatives in total credit disbursement. As RRBs 66 percent and cooperatives 70 percent of total loan goes to small and marginal farmers while only 47 percent of total loan of commercial banks goes to small and marginal farmers which has largest share in total credit disbursement. Tenant and sharecropper are largely excluded in most of the government schemes and benefits.

6. Conclusion and Policy Recommendation

The preceding analysis exhibits that though government policies and initiatives have undoubtedly resulted in considerable improvement over the period of time especially in increase in ground level credit flow, but disparity in distribution one of the major concerns at present time. While much of the improvement in credit parameters is accounted for by a few high-income states, there exists severe disparities with the low-income states consistently maintaining their lower positions relative to the former, indicating absence of equity in distribution of credit. In Central and Eastern regions, the various real sector variables indicate relatively favourable situation for enhanced credit flow. For example, the share of Central region in net sown area is 26 per cent against its share in GLC at 13.7%, not commensurate with its share in irrigation area and net sown area. In Eastern region around 95 per cent of the operational holdings are of small and marginal farmers. In central and NER region the share is 86 per cent and 82 percent, respectively. There is a need to sensitize and develop credit products that are in sync with the need of the majority of the borrowers. The state-wise percentage of farm households with loan accounts is lower in eastern region as compared to southern states. In the eastern and central region, there is huge gap in the number of farm households to be covered under bank finance. States like Uttar Pradesh and Bihar need to be focused for more coverage of house hold accounts as well. Some of the States like Tamil

Nadu, Andhra Pradesh, Kerala, and Karnataka) have higher number of bank accounts than the number of farm house holds.

Table 9: Ground Level Credit Flow to Agriculture – Share of SF/MF

Yr.	Agency	No of accounts (lakh)			Loan disbursed (Rs. crore)			Average loan amt of SF/MF (Rs.)
		Total	SF/MF	Share of SF/MF (%)	Total	SF/MF	Share of SF/MF (%)	
2013-14	Com. Banks	385.3	232.5	60.4	5,27,506	2,01,296	38.2	86579
	Coop. Banks	321.4	206.5	64.1	1,19,964	69,352	57.8	33585
	RRBs	99.3	66.6	67.1	82,653	51,359	62.1	77116
	Total	805.9	505.2	62.7	7,30,123	3,22,007	44.1	63739
2014-15	Com. Banks	426.2	195.4	45.9	6,04,376	1,97,540	32.7	101095
	Coop. Banks	306.9	202.8	66.1	1,38,470	78,736	56.9	38824
	RRBs	120.5	87.8	72.9	1,02,483	70,390	68.7	80171
	Total	853.6	486.3	56.9	8,45,328	3,46,666	41.1	71286
2015-16	Com Banks	441.6	210.2	47.6	6,42,954	2,00,346	31.2	95312
	Coop. Banks	324.2	232.9	71.8	1,53,295	97,999	63.9	42078
	RRBs	133.2	97.0	72.8	1,19,261	81,653	68.5	84178
	Total	899.6	540.4	60.7	9,15,510	3,79,998	41.5	70318
2016-17	Com. Banks	664.2	482.5	72.6	7,99,781	3,62,675	45.4	75166
	Coop. Banks	269.5	190.1	70.5	1,42,758	89,178	62.5	46911
	RRBs	137.0	99.0	72.3	1,23,216	82,496	67.0	83329
	Total	1070.7	771.6	72.6	10,65,755	5,34,351	50.1	69252
2017-18	Com. Banks	732.7	556.9	76.0	8,71,080	3,89,866	44.8	70009
	Coop. Banks	254.6	183.7	72.2	1,50,321	98,109	65.3	53401
	RRBs	144.6	104.9	72.5	1,41,216	92,482	65.5	88191
	Total	1131.9	845.5	74.7	11,62,617	5,80,457	49.9	68655
2018-19	Com. Banks	850.1	631.8	74.3	9,54,823	4,28,063	44.8	67753
	Coop. Banks	255.5	192.9	75.5	1,52,340	1,06,849	70.1	55405
	RRBs	149.8	106.7	71.3	1,49,667	98,749	66.0	92539
	Total	1255.4	931.4	74.2	12,56,830	6,33,661	50.4	68036
2019-20	Com. Banks	934.3	714.4	76.5	10,61,215	5,02,172	47.3	70294
	Coop. Banks	270.6	198.6	73.4	1,49,694	1,04,883	70.1	52809
	RRBs	153.5	109.5	71.3	1,62,857	1,07,301	65.9	98010
	Total	1358.3	1022.5	75.3	13,73,766	7,14,356	52.0	69866

Source*: NABARD Annual Report*

Satyasai and Kumar (2020) have developed index (NAFINDEX) based on NABARD All India Rural Financial Inclusion Survey 2016-17 data. State wise performance of NAFINDEX is presented in Table 10. NAFINDEX is calculated based on 3 parameters *i.e.*, traditional banking product, modern banking services and payment mechanism. It is very encouraging to see that small state like Goa and North-eastern states have performed exceedingly well especially in the modern banking services. Modern banking services can be

one of the innovative ways to increase the penetration of credit in credit starved region. This will also help in promoting the regional equity in credit distribution.

Table 10: NAFINDEX

State	Traditional Banking products	Rank	Modern Banking Services	Rank	Payment mechanism	Rank	NAFINDEX	Rank
Goa	0.472	5	0.946	1	0.761	1	0.600	1
Punjab	0.617	1	0.473	12	0.383	19	0.486	2
Karnataka	0.533	3	0.430	14	0.438	13	0.483	3
Telangana	0.482	4	0.563	8	0.478	8	0.480	4
Andhra Pradesh	0.424	7	0.703	4	0.529	5	0.473	5
Kerala	0.609	2	0.446	13	0.362	21	0.470	6
Manipur	0.385	12	0.791	2	0.558	3	0.464	7
Tripura	0.366	14	0.523	10	0.558	3	0.452	8
Jammu & Kashmir	0.420	8	0.427	15	0.450	12	0.435	9
Odisha	0.379	13	0.381	24	0.477	9	0.425	10
Haryana	0.409	10	0.328	26	0.423	14	0.416	11
Mizoram	0.322	16	0.580	6	0.476	10	0.392	12
Assam	0.237	21	0.482	11	0.625	2	0.385	13
Himachal Pradesh	0.460	6	0.565	7	0.310	23	0.377	14
Meghalaya	0.318	17	0.240	29	0.403	17	0.358	15
Arunachal Pradesh	0.337	15	0.353	25	0.374	20	0.355	16
Sikkim	0.253	20	0.678	5	0.486	7	0.351	17
Nagaland	0.318	17	0.734	3	0.325	22	0.322	18
West Bengal	0.202	25	0.419	16	0.507	6	0.320	19
Maharashtra	0.224	22	0.416	18	0.416	16	0.305	20
Jharkhand	0.200	26	0.321	27	0.451	11	0.301	21
Gujarat	0.215	24	0.531	9	0.420	15	0.300	22
Uttar Pradesh	0.217	23	0.417	17	0.397	18	0.294	23
Tamil Nadu	0.387	11	0.404	20	0.208	25	0.284	24
Uttarakhand	0.420	8	0.401	21	0.189	27	0.281	25
Bihar	0.198	27	0.387	23	0.264	24	0.229	26
Rajasthan	0.276	19	0.398	22	0.178	28	0.222	27
Madhya Pradesh	0.141	29	0.266	28	0.195	26	0.166	28
Chhattisgarh	0.160	28	0.411	19	0.055	29	0.094	29
All India	0.307		0.345		0.370		0.337	

Source: *Satyasai, K.J. and Kumar, A., 2020. NAFINDEX: Measure of Financial Inclusion based on NABARD All India Rural Financial Inclusion Survey (NAFIS) Data.*

The mismatch between demand and supply for financial services, substantial differences across regions, socio-economic groups and gender, high cost, inadequate access for micro and small enterprises and inability to harness technology to the fullest extent have been cited as the major impediments for poor uptake of credit. Inadequate banking infrastructure is another cause of concern in low-income states and regions challenged by topography, inadequate infrastructure, and security issues. Further, the demand for credit and credit absorptive capacity depends upon various factors which shape opportunities like initial asset endowments (household and geographical), cultural and geographical identification, market conditions and participation, policy environment, infrastructure, quality of institutions, industrial development, political economy of growth and public investment.

6.1. Recommendations/Suggestions for improving regional spread of credit flow

On the basis of the analysis, following suggestions are offered:

a) Financial Deepening is crucial, which not only helps in garnering more resources from the state but also helps in channelizing more resources to the state. This involves more penetration of physical outlets, BCs, and digital banking facilities, especially in the states or regions with imbalance in credit flow. Also, this involves better deposit mobilization, which requires targeted financial literacy campaigns highlighting the importance of thrift.

b) Providing better infrastructure to create an investor friendly environment leads to greater credit absorption.

c) Strengthening of RFIs especially in certain regions of the country (which are credit starved regions also) are weak. This reduces their ability to lend at grass root level. Strengthening such institutions will lead to enhanced credit flow and reduce the imbalance. As it was observed with the declining share of RRBs' and cooperative banks in the amount disbursed.

d) Computerisation of land records: Farmers such as tenants and cultivators with only usufructuary rights on their land without clear titles face difficulties in accessing institutional credit and other facilities as they cannot offer collateral. It is required to computerise land records so that a transparent system for changing land records and dividing or merging plots of land may be created.

e) Strategies for Financing SF/MF -Overall: PSL guidelines by RBI stipulated that within the 18 per cent target outlined for agriculture, 8 per cent of Adjusted Net Bank Credit (ANBC) or credit equivalent amount of off-balance sheet exposure, whichever is higher, to be purveyed to SF/MF from March 2017 onwards. RBI's Internal working group on agriculture credit has suggested to increase this sub target to 10%.

f) Land Lease Markets: State governments should be encouraged to reform their legal framework on the basis of Model Land Leasing Act proposed by NITI Aayog so that formal lending to tenant farmers can improve. This would ensure that land owners have the security of ownership rights, and land tenants are secure in their tenancy. Legalisation of land tenancy would also ensure that farmers get access to formal credit, insurance, and inputs such as fertilizer.

g) Addressing Regional Disparity: Allocation of RIDF in central, eastern and north eastern states may be increased over time to strengthen the rural infrastructure in these regions.

h) Credit Guarantee: Tenants and sharecropper are unable to access formal credit due to lack of collateral. It is important to create a credit guarantee scheme on the lines of Credit guarantee fund trust for micro and small enterprises to provide collateral free loans.

i) Collectives: Farmers' Producer Organisations (FPOs) help in overcoming the challenges of high transaction costs, security stipulations of loans and also support smallholders in gaining access to markets, public services, better price etc. through collective action. Further, small farmers in collectives would get more capacity for getting inputs at reasonably lower price and quality material and also to gain from market that would provide more income and encourage farmers to access more credit.

References

Aghion, P. and Bolton, P. (1997). A theory of trickle-down growth and development with debt-overhang. *Review of Economic Studies*, **64**, 151-172

Aportela, F. (1999). *Effects of Financial Access on Savings by Low-Income People*. Banco de México Research Department. Chapter 1 of Doctoral Dissertation, MIT Department of Economics, Cambridge, Mass.

Basu, K. (2001). On the goals of development. In *Frontiers of Development Economics*: *The Future in Perspective*, Eds. Meier, G.M. and Stiglitz, J.E., Oxford University Press, New York.

Basu, K. (2006). Globalisation, poverty and inequality: what is the relationship? What can be done? *World Development,* **34**(**8**), 1361-1373.

Beck, T. and Levine, R. (2004). Legal institutions and financial development. *NBER Working Papers Series No. Working Paper* 10417. Cambridge, MA: National Bureau of Economic Research.

Banerjee, A. V. and Newman, A. F. (1993). Occupational choice and the process of development. *Journal of Political Economy*, **101**(**2**), 274-298.

Beck, T., Demirgüç-Kunt, A. and Martinez-Peria, S. (2008). Banking services for everyone? Barriers to bank access and use around the world. *The World Bank Economic Review*,**22**(**3**), 397-430.

Beck, T., Demirguc-Kunt, A., and Levine, R. (2007). Finance, inequality and the poor. *Journal of Economic Growth*, **12**(**1**), 27-49.

Burgess, R. and Pande, R. (2005). Do rural banks matter? Evidence from the Indian social banking experiment. *American Economic Review*, **95**(**3**), 780-795

Center for Financial Inclusion at Accion International (2018). *Financial Inclusion Hype Vs. Reality: Deconstructing the* 2017 *Findex Results*. Center for Financial Inclusion.

Chavan, P. (2020). Women's access to banking in India: Policy context, trends, and predictors. *Review of Agrarian Studies,* **10**.

Demetriades, P.O. and Hussein, K.A. (1996). Does financial development cause economic growth? Time-series evidence from 16 countries, *Journal of Development Economics*, **51**(**2**), 387-411.

Demirgüç-Kunt, A., Leora, K., Dorothe S., Saniya A. and Jake, H. (2018). *The Global Findex Database* 2017: *Measuring Financial Inclusion and the Fintech Revolution*. Washington, DC, World Bank.

Dupas, P. and Robinson, J. (2009). Savings constraints and micro-enterprise development: evidence from a field experiment in Kenya. *NBER Working Paper No.* 14693.

Galor, O. and Seira, J. (1993). Income distribution and macroeconomics. *Review of Economic Studies*, **60**(**1**), 35-52

Goldsmith, R.W. (1969). *Financial Structure and Development.* New Haven, CT: Yale University Press.

Herwadkar, S. and Ghosh, S. (2013). *What Explains Credit Inequality Across Indian States? An Empirical Analysis.* Reserve Bank of India Occasional Paper, **34(1 & 2**).

Hulme, D. and Mosley, P. (1996). Financial sustainability, targeting the poorest and income impact: are there trade-offs for micro-finance institutions? *Focus*, **5** (December), Washington D.C.

Khan, A. (2001). Financial development and economic growth. *Macroeconomic Dynamics*, **5**(**3**), 413-433.

King, R. G. and Levine, R. (1993). Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics*, **108**, 717-737.

Kumar, A., Singh, D. and Kumar, P.(2007). Performance of rural credit and factors affecting the choice of credit sources. *Indian Journal of Agriculture Economics*, **62**(**3**), July-Sept.

Levine, R., Loayza, N. and Beck, T. (2000). Financial intermediation and growth: causality and causes. *Journal of Monetary Economics*,**46**(**1**), 31–77.

Levine, R. (2005). Finance and growth: theory and evidence. In *Handbook of Economic Growth*, Eds. Philippe Aghion and Steven Durlauf, Amsterdam: North-Holland Elsevier Publishers.

NABARD (2019). *Annual Report*. National Bank for Agriculture and Rural Development, Mumbai.

Rajan, R. G. and Zingales, L. (1998). Financial dependence and growth. *The American Economic Review*,**88**(**3**), 559–586. http:/www.jstor.org.

Rajan, R. G. and Zingales, L. (2003). *Saving Capitalism from the Capitalists*: *Unleashing the Power of Financial Markets to Create Wealth and Spread Opportunity*. New York: Crown Business.

Rajesh, R. and Das, A. (2019). Drivers of credit penetration in Eastern India. *Reserve Bank of India Bulletin*, October 11, 2019, Department of Economic and Policy Research, RBI, Kolkata. https://www.rbi.org.in/Scripts/BS_ViewBulletin.aspx.

Reddy, Y. V. (2015). Financial inclusion and central banking: Reflections and issues. Keynote speech at the 14th SEACEN Executive Committee meeting, Port Moresby. http://www.bis. org/review/ r151029a.pdf.

Reserve Bank of India (RBI) (2015). *Report of the Committee on Medium-term Path on Financial Inclusion*. December.

Rogaly, B. and Fisher, T. (1999). Introduction. In *Poverty, Social Exclusion and Microfinance in Britain*. Eds. Rogaly. B., Fisher, T. and Mayo, E., Oxford: Oxfam Publishing.

Sarma, M. (2008). *Index of Financial Inclusion*. ICRIER Working Paper No.215, Indian Council for Research on International Economic Relations, New Delhi, India.

Satyasai, K. J. (2012). Access to rural credit and input use: An empirical study. *Agricultural Economics Research Review*, **25**, 461-471.

Satyasai, K. J. and Kumar, A. (2020).NAFINDEX: *Measure of Financial Inclusion based on NABARD All India Rural Financial Inclusion Survey (NAFIS) Data*. NABARD Working Paper 1.

Shetty, S. L. (1997). Financial sector reforms in India: an evaluation. *Prajnan*, **25**(3-4). 253-287.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 161-176

OCD for D-optimal CRD with Odd Replication

K. K. Singh Meitei
Department of Statistics, Manipur University, India,

Received: 25 June 2021; Revised: 16 August 2021; Accepted: 22 August 2021

Abstract

Optimum allocation with an Orthogonal Covariate Design (OCD) with only entries ± 1, can not be accommodated to CRD with at least one of the replications odd, such that treatment-effects and covariate-effects are independently estimable. A CRD with at least one of replications odd, accommodated with an OCD of entries ± 1, can be D-optimal, but it cannot permit the independent estimation of treatment-effects and covariate-effects. Thus, usefulness of covariates to CRD is weakened. The present paper gives (i) the construction of corresponding OCD-component to treatment with replications, 3, 5 and q (odd) where $\mathbf{H}_{q-1}$ exists, separately, for D-optimal CRD and (ii) the construction of OCD for D-optimal CRD among the class of competent design permitting the independent estimation of treatment-effects and covariate-effects, when at least one of the replications is either 3, 5 or q (odd) where $\mathbf{H}_{q-1}$ exists.

Key words: Orthogonal covariate design; Restricted Hadamard matrix; Completely randomised design; Closures of order q to m; Covariate design-component.

1. Introduction

Lopes Troya (1982) initiated the problem of finding optimum covariate designs. In the same spirit, Wierich (1984) , Chadjiconstantinidis and Moyssiadis (1991), Liski *et al.* (2002), Das *et al.* (2003), Rao *et al*. (2003), Dutta *et al*. (2014) and many others have contributed to in the field of covariate design and its set-ups. Our problem for optimum allocation of OCD to CRD is that when one of the replications of treatments of CRD is odd and OCD accommodated to the CRD is with entries ± 1, the CRD cannot be D-optimal, permitting to estimate the treatment-effects and covariate-effects independently. If the CRD with a covariate set-up is D-optimal, it does not permit the independent estimation of treatment-effects and covariate-effects [*cf.* Example No. 5, Dey and Mukerjee (2006)] which can be seen from the relation (1.8). Thus, usefulness of covariates to CRD is weakened. Even when all replications of CRD are even, the optimum allocations of OCD with entries ± 1 for CRD to be D-optimal may not exist. As an example, consider a CRD (v = 3, n_1= n_2= 2, n_3= 4) accommodated with 4 covariates whose respective covariate design-components to the 1st, the 2nd and the 3rd treatments are given by

$$\mathbf{Z}_1=\begin{bmatrix}+1 & +1 & +1 & +1\\ -1 & -1 & -1 & -1\end{bmatrix},\ \mathbf{Z}_2=\begin{bmatrix}-1 & +1 & +1 & -1\\ +1 & -1 & -1 & +1\end{bmatrix} \text{ and } \mathbf{Z}_3=\begin{bmatrix}+1 & +1 & -1 & -1\\ +1 & -1 & +1 & +1\\ -1 & +1 & -1 & +1\\ -1 & -1 & +1 & -1\end{bmatrix}$$

Corresponding Author: K.J.S. Satyasai
Email: satyasaik@outlook.com

The general FEALM of CRD (v, $n_1, n_2, \ldots, n_v$), assuming every treatment is exposed to the same number of covariates c, is

$$y_{ij} = \mu_i + \sum_{p=1}^{c} z_{ij}^{(p)} \gamma_p + e_{ij};\ i = 1, 2, \ldots, v;\ j = 1, 2, \ldots, n_i;\ \sum n_i = n, \text{ say.} \quad (1)$$

where y_{ij}, μ_i, γ_p, $z_{ij}^{(p)}$ and e_{ij} are the j-th observation receiving the i-th treatment with common variance σ^2, the i-th treatment-effect, the p-th covariate- effect, the value of the p-th covariate exposed to the j-th observation receiving the i-th treatment and the random error component of y_{ij} with common variance σ^2 respectively.

Denoting $\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_v, \gamma_1, \gamma_2, \ldots, \gamma_c)'$, the information matrix for $\boldsymbol{\theta}$ is given by [*cf.* (2.3.8) of Das *et al.* (2015)]

$$\sigma^{-2}\mathrm{I}(\boldsymbol{\theta}) = \sigma^{-2} \begin{bmatrix} \mathbf{N} & \mathbf{T} \\ \mathbf{T}' & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \quad (2)$$

where $\mathbf{N} = diag(n_1, n_2, \ldots, n_v)$ (3)

$$\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2', \ldots, \mathbf{T}_v')';\ \mathbf{T}_i = \mathbf{1}_{n_i}'\mathbf{Z}_i \quad (4)$$

$$\mathbf{Z} = \mathbf{Z}^{(\mathrm{n}\times\mathrm{c})} = (\mathbf{Z}_1', \mathbf{Z}_2', \ldots, \mathbf{Z}_v')' \quad (5)$$

such that no column of Z is zero-column.

$$\mathbf{Z_i} = \mathbf{Z}_i^{(n_i \times c)} = \begin{bmatrix} z_{i1}^{(1)} & z_{i1}^{(2)} & \cdots & z_{i1}^{(c)} \\ z_{i2}^{(1)} & z_{i2}^{(2)} & & z_{i2}^{(c)} \\ \vdots & & \ddots & \vdots \\ z_{in_i}^{(1)} & z_{in_i}^{(2)} & \cdots & z_{in_i}^{(c)} \end{bmatrix} = [\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \ldots, \mathbf{Z}_i^{(\mathrm{c})}] \quad (6)$$

which will be known as the covariate design (CD)-component to the i-th treatment. Then, $\mathbf{Z}'\mathbf{Z} = \sum \mathbf{Z}_i'\mathbf{Z}_i$ and $\det(\mathbf{I}(\boldsymbol{\theta})) = \det(\mathbf{N}) \times \det(\mathbf{C}) = (\prod_{i=1}^{v} n_i) \times \det.(\mathbf{C})$, by (3) to (6)

$$\text{where } \mathbf{C} = \mathbf{Z}'\mathbf{Z} - \sum n_i^{-1}\mathbf{T}_i'\mathbf{T}_i \text{ and } \mathbf{T}_i = \mathbf{1}_{n_i}'\mathbf{Z}_i \quad (7)$$

The maximization of $\det(\mathbf{I}(\boldsymbol{\theta}))$ can be done in two stages: (i) the maximization of $\prod_{\mathrm{i}=1}^{v} n_i$ and (ii) that of $\det(\mathbf{C})$, [*cf.* Dey and Mukerjee (2006)] which were generalised and more broadened by Dutta *et al.* (2014). From (2) it can be seen that $\mathbf{Z}'\mathbf{Z}$ is a diagonal matrix iff the covariates are orthogonal to one another and $\mathbf{1}_{n_i}'\mathbf{Z}_i = \mathbf{0}'\ \forall\ i$, *i.e.* $\mathbf{X}'\mathbf{Z} = \mathbf{0}'$ iff the covariates are orthogonal to treatments *i.e.* the covariate-effects are independently estimable iff $\mathbf{Z}'\mathbf{Z}$ is a diagonal matrix and further, the covariate-effects and the treatment-effects are independently estimable iff

$$\mathbf{1}_{n_i}'\mathbf{Z}_i = \mathbf{0}'\ \forall\ i\ ,\ i.e.\ \mathbf{X}'\mathbf{Z} = \mathbf{0}'. \quad (8)$$

Thus, when $\mathbf{1}_{n_i}'\mathbf{Z_i} = \mathbf{0}'\ \forall$ i, each covariate matrix $\mathbf{Z_i}$ behaves independently in the sense that a covariate design-component (covariate matrix) $\mathbf{Z}_i$ does not affect another covariate design-

component $\mathbf{Z}_{i'}$ $(i \neq i')$. When at least one of n_i's is odd and $z_{ij}^{(p)} = \pm 1$, at least one of the relations (i) $\mathbf{T} = \mathbf{O}$, (ii) $\mathbf{Z}'\mathbf{Z} = n\mathbf{I}_c$ and (iii) n is multiple of v, is violated in obtaining c-Orthogonal Covariate Design (OCD) for the CRD $(v, n_1, n_2, \ldots, n_v)$; $\Sigma n_i = n$. Now, we will look for the situations where the relation (i) holds but the relation (ii) does not. Further, by (7) if $\|\mathbf{Z}_i^{(p)}\|$ are at maximum $\forall$ i, p, then det.($\mathbf{I}(\boldsymbol{\theta})$) obtains its maximum for D-optimal CRD. Now, the focus is on maximization of $\|\mathbf{Z}_i^{(p)}\|$ ∋ $\mathbf{T}_i = \mathbf{0}'$ $\forall$ i, p.

Definition 1: Given a set S (whether finite or infinite) = $\{\mathbf{x}_i / \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'; x_{ij} \in [-1, 1]; i = 1,2, \ldots\}$, $\mathbf{x}_{\alpha_1}, \mathbf{x}_{\alpha_2}, \ldots, \mathbf{x}_{\alpha_q}$ are said to be closures of order q to m among $\mathbf{x}_i$'s (simply, closures of order q) if $\|\mathbf{x}_{\alpha_1}\| \geq \|\mathbf{x}_{\alpha_2}\| \geq \ldots \geq \|\mathbf{x}_{\alpha_q}\| \geq \|\mathbf{x}_\beta\|$ $\forall$ $\mathbf{x}_\beta \in S$ where $\mathbf{x}_{\alpha_1}, \mathbf{x}_{\alpha_2}, \ldots, \mathbf{x}_{\alpha_q} \in S$ and other vectors are said to be non-closures of order q among $\mathbf{x}_i$'s.

When $q = 1$, $\mathbf{x}_{\alpha_1}$ is said to be closure to m among all other vectors (simply, a closure to m).

Remark 1:

(a) For a given set S, closures of a particular order is not unique. E.g., letting $S = \{(-1, 1, -1, 1)', (-1, -1, 1, 1)', (-1, 1, 1, -1)', (1, -1, 1, -1)', (1, 1, -1, -1)', (1, -1, -1, 1)'\} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_6\}$, say, any $q(\leq 4)$ out of 6 $\boldsymbol{s}_i$'s are closures of order $q(\leq 4)$.

(b) A vector $\boldsymbol{x} \in S$ may be one of the closures of an order and may not be of other smaller order.

(c) A vector $\mathbf{x} \in S \subseteq R^m$, which is one of the closures of an order, may not be closure of the same order if there exists at least another vector $\mathbf{y}$ $(\in S)$ ∋ $\|\mathbf{x}\| = \|\mathbf{y}\|$. That is, if there exist some vectors $\mathbf{x}_{\alpha_1}, \mathbf{x}_{\alpha_2}, \ldots, \mathbf{x}_{\alpha_u}$ ∋ $\|\mathbf{x}_{\alpha_1}\| = \|\mathbf{x}_{\alpha_2}\| = \ldots = \|\mathbf{x}_{\alpha_u}\| = \|\boldsymbol{x}\|$, each of them can be treated as one of the closures of the order q to m and at the same time can be treated as one of the non-closures of the order q to m.

(d) By the definition of closure of an order to m, a vector $\boldsymbol{x}_\alpha$ can be treated as a closure of any order to m if $\mathbf{x}_{\alpha j} = \pm 1$ $\forall$ j =1, 2, …, m i.e. $\|\mathbf{x}_{\boldsymbol{\alpha}}\| = m$.

Definition 2: An OCD-component $\mathbf{Z}_i = [\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \ldots, \mathbf{Z}_i^{(c)}]$ is said to be optimal if $\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}, \ldots, \mathbf{Z}_i^{(c)}$ are closures of order c to n_i.

Definition 3: An $m \times n$-matrix $\mathbf{A} = [a_{ij}]$; $a_{ij} \in [-1, 1]$, is said to be restricted Hadamard matrix if $\mathbf{A}'\mathbf{A}$ is diagonal and $\mathbf{1}'\mathrm{A} = \mathbf{0}'$.

For finding the number of orthogonal covariate components to n_i observations receiving the i-th treatment (OCD-component to the i-th treatment) in CRD, a lemma is given at below:

Lemma 1: For a restricted Hadamard matrix $\mathbf{A}_{m \times n}$, n is less than or equal to $m - 1$.

Proof of the lemma is given in APPENDIX-A.

Some propositions whose proof are straight forward, are given below, for future use.

Proposition 1: If $n_i \times m$-H$_i$; i = 1, 2, …, s, are s restricted Hadamard matrices, then $\mathbf{H} = (\mathbf{H}_1', \mathbf{H}_2', \ldots, \mathbf{H}_s')'$ is again so.

Obviously, $\mathbf{H}'\mathbf{H} = \text{diag.}(\sum_i a_{i1}, \sum_i a_{i2}, \ldots, \sum_i a_{im})$ where $\mathbf{H}_i'\mathbf{H}_i = \text{diag.}(a_{i1}, a_{i2}, \ldots, a_{im})$.

Proposition 2: If $n \times m$- $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$ is a restricted Hadamard matrix, then $\mathbf{H}^* = (\mathbf{h}_{\alpha_1}, \ldots, \mathbf{h}_{\alpha_s})$; $s < m$, is again so, where $\mathbf{h}_{\alpha_i}$'s are distinct and columns of $\mathbf{H}$.

Obviously, $\mathbf{H}^{*'}\mathbf{H}^* = \text{diag.}(a_{\alpha_1}, \ldots, a_{\alpha_s})$ where $a_{\alpha_i} = \| \mathbf{h}_{\alpha_i} \|$.

Proposition 3: If $n \times m$- $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$ is a restricted Hadamard matrix, then $\mathbf{H}^* = (\mathbf{h}_1, \ldots, \mathbf{h}_m, \mathbf{0}, \ldots, \mathbf{0})$ of order $n \times (m + s)$ is again so.

Obviously, $\mathbf{H}^{*'}\mathbf{H}^* = diag.(a_1, \ldots, a_m, 0, \ldots, 0)$, where $\mathbf{H}'\mathbf{H} = diag.(a_1, \ldots, a_m)$.

2. Construction of OCD-component

As covariate-effects and treatment-effects are to be independently estimable, $\mathbf{T} = (\mathbf{T}_1', \ldots, \mathbf{T}_v')'$ i.e. $(\mathbf{1}_{n_1}'\mathbf{Z_1}, \ldots, \mathbf{1}_{n_v}'\mathbf{Z}_v)'$ i.e. $\mathbf{X}'\mathbf{Z} = \mathbf{0}_{v \times c}$. Each of $\mathbf{Z}_i$'s has its own complete structure in the sense that no entry in $\mathbf{Z}_i$ does not depend on any entry in $\mathbf{Z}_{i'}$ ($i \neq i'$) concerning the independent estimation of covariate-effects and treatment-effects. In other words, none of $\mathbf{Z}_i$'s depends on the remaining others concerning the independent estimation of covariate-effects and treatment-effects. So, in this section, we focus on construction-methods of OCD-components (which are closures of some order) to the β-th treatment of CRD to be D-optimal among all possible OCD-components of the competent design, permitting the independent estimation of covariate-effects and treatment-effects for which n_β= 3, 5 and q (odd) provided Hadamard matrix $\mathbf{H}_{q-1}$ exists. The need of such OCD-components lies when at least one of n_i's is 3, 5 or q odd. Meanwhile, for the remaining n_i's, apart from 3, 5 and q (odd) provided $\mathbf{H}_{q-1}$ exists, their corresponding OCD-components are assumed to be known and existed. The construction-methods are dealt with through induction of solving restrictions imposed by the orthogonal conditions among covariates and that of covariates to the treatments.

2.1. Construction of OCD-component for $n_\beta = 3$ in CRD.

Let $n_\beta = 3$ be one of the n_i's and $\mathbf{Z}_\beta^{(p)} = (z_{\beta 1}^{(p)}, z_{\beta 2}^{(p)}, z_{\beta 3}^{(p)})'$. Then, by Lemma 1, p equals to 2 at most *i.e.* c = 2.

Case I. Considering c = 1, by the condition of orthogonal of covariate to the β-th treatment,

$$\sum_{j=1}^{3} z_{\beta j}^{(1)} = 0. \tag{9}$$

WOLG, assume $\widehat{z_{\beta 1}^{(1)}} = a_1$, $\widehat{z_{\beta 2}^{(1)}} = a_2 \ni a_1, a_2 \in [-1, 1]$ for (9) becomes a non-homogeneous linear equation involving one unknown variable so as $\widehat{z_{\beta 3}^{(1)}} = -(a_1 + a_2) \in [-1, 1]$.

Therefore, $\widehat{\mathbf{Z}_\beta^{(1)}} = (a_1, a_2, -(a_1 + a_2))'$ (10)

Case II. Considering $c = 2$, by the condition of orthogonal of covariates to the β-th treatment and that among covariates, we have

$$\sum_{j=1}^{3} z_{\beta j}^{(1)} = \sum_{j=1}^{3} z_{\beta j}^{(2)} = 0 \tag{11}$$

$$\sum_{j=1}^{3} z_{\beta j}^{(1)} z_{\beta j}^{(2)} = 0. \tag{12}$$

Using (10) of the Case I: $c = 1$, for estimating $\mathbf{Z}_{\boldsymbol{\beta}}^{(\mathbf{2})}$, (11) and (12) give

$$\sum_{j=1}^{3} z_{\beta j}^{(2)} = 0 \tag{13}$$

$$a_1 z_{\beta 1}^{(2)} + a_2 z_{\beta 2}^{(2)} - (a_1 + a_2)\, z_{\beta 3}^{(2)} = 0. \tag{14}$$

WOLG, assume $\widehat{z_{\beta 1}^{(2)}} = b_1 \in [-1, 1]$ for (13) and (14) become a system of 2 linearly independent non-homogeneous linear equations involving two unknown variables (independent is due to orthogonal between $\mathbf{Z}_{\beta}^{(1)}$ and $\mathbf{Z}_{\beta}^{(2)}$)

$$b_1 + z_{\beta 2}^{(2)} + z_{\beta 3}^{(2)} = 0 \tag{15}$$

$$a_1 b_1 + a_2 z_{\beta 2}^{(2)} - (a_1 + a_2) z_{\beta 3}^{(2)} = 0 \tag{16}$$

so as $\widehat{z_{\beta 2}^{(2)}} = -(a_2 + 2a_1) b_1 / (2a_2 + a_1) \in [-1, 1]$, $\widehat{z_{\beta 3}^{(2)}} = (a_1 - a_2) b_1 / (2a_2 + a_1) \in [-1, 1]$, by (15) and (16).

Consequently, $\widehat{\mathbf{Z}_{\beta}^{(2)}} = (b_1, -(a_2 + 2a_1) b_1/(2a_2 + a_1), (a_1 - a_2) b_1/(2a_2 + a_1))'$. (17)

Therefor, $\widehat{\mathbf{Z}_{\boldsymbol{\beta}}} = \begin{bmatrix} a_1 & b_1 \\ a_2 & -(a_2 + 2a_1) b_1/(2a_2 + a_1) \\ -(a_1 + a_2) & (a_1 - a_2) b_1/(2a_2 + a_1) \end{bmatrix}$, by (10) and (17) (18)

It is seen from (18) that infinite number of $\mathbf{Z}_{\boldsymbol{i}}$'s exist for various values of a_1, a_2, and b_1. All these $\mathbf{Z}_{\boldsymbol{i}}$'s are restricted Hadamard matrices of order 3×2. Searching the largest value of $\left\| \mathbf{Z}_{\beta}^{(p)} \right\| \ni z_{\beta j}^{(p)} \in [-1, 1]$ *i.e.* searching the closures of order 2 to 3, the possible forms of $\mathbf{Z}_{\beta}$ are given in the following table.

Table 1:

Sr. No.	1		2		3		4		5		6		7		8	
$\mathbf{Z}_{\beta}$	½	1	−½	1	½	−1	−½	−1	½	1	−½	1	½	−1	−½	−1
	½	−1	−½	−1	½	1	−½	1	−1	0	1	0	−1	0	1	0
	−1	0	1	0	−1	0	1	0	½	−1	−½	−1	½	1	−½	1
Sr. No.	9		10		11		12		13		14		15		16	
$\mathbf{Z}_{\beta}$	−1	0	1	0	−1	0	1	0	1	½	−1	½	1	−½	−1	−½
	½	−1	−½	−1	½	1	−½	1	−1	½	1	½	−1	−½	1	−½
	½	1	−½	1	½	−1	−½	−1	0	−1	0	−1	0	1	0	1
Sr. No.	17		18		19		20		21		22		23		24	
$\mathbf{Z}_{\beta}$	1	½	−1	½	1	−½	−1	−½	0	−1	0	−1	0	1	0	1
	0	−1	0	−1	0	1	0	1	−1	½	1	½	−1	−½	1	−½
	−1	½	1	½	−1	−½	1	−½	1	½	−1	½	1	−½	−1	−½

A lemma to be used later on, is proposed here.

Lemma 2. The $\mathbf{Z}_{\beta}$'s given in the Table 1 are the possible optimal OCD-components of CRD (ν, $n_1, n_2, \ldots, n_v$) to the β-th treatment with replication 3 *i.e.* $n_{\beta} = 3$.

Proof of the lemma is given in APPENDIX-B.

2.2. Construction of OCD-component for $n_{\beta} = 5$ in CRD.

Let $n_{\beta} = 5$ be one of the n_i's and $\mathbf{Z}_{\beta}^{(p)} = (z_{\beta 1}^{(p)}, z_{\beta 2}^{(p)}, z_{\beta 3}^{(p)}, z_{\beta 4}^{(p)}, z_{\beta 5}^{(p)})'$.

For c orthogonal covariates, by the condition of orthogonal of the covariates to the β-th treatment and that among themselves, we have $c(c+1)/2$ relations

$$\left.\begin{aligned} \textstyle\sum_j z_{\beta j}^{(p)} &= 0 \ \ \forall\, p = 1, 2, \ldots . c \\ \textstyle\sum z_{\beta j}^{(p)} z_{\beta j}^{(p')} &= 0 \ \ \forall\, p \neq p' = 1, 2, \ldots \ldots c \end{aligned}\right\} \tag{19}$$

By the Lemma 1, there exist at most 4 orthogonal covariates *i.e.* p equals to 4 at most, which will be enumerated in the following. Considering $c = 1, 2, 3, 4$ successively and assuming 4, 3, 2, 1 values of $z_{\beta j}^{(p)}$ as follows (for c = 1) $z_{\beta 1}^{(1)} = a_1$, $z_{\beta 2}^{(1)} = a_2$, $z_{\beta 3}^{(1)} = a_3$, $z_{\beta 4}^{(1)} = a_4$, (for c = 2) $z_{\beta 1}^{(2)} = b_1$, $z_{\beta 2}^{(2)} = b_2$, $z_{\beta 3}^{(2)} = b_3$, (for c = 3) $z_{\beta 1}^{(3)} = c_1$, $z_{\beta 2}^{(3)} = c_2$, (for c = 4) $z_{\beta 1}^{(4)} = d_1$ in (19) correspondingly, ∋ a_1 to a_4, b_1 to b_3, $c_1, c_2, d_1 \in [-1, 1]$, there are q independent non-homogenous linear equations involving q unknown variables for c = q; q = 1, 2, 3, 4 [independent is due to orthogonal among $\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{p})\prime}s$] which give the solutions $\widehat{z_{\beta 5}^{(1)}}, \widehat{z_{\beta 4}^{(2)}}, \widehat{z_{\beta 5}^{(2)}}, \widehat{z_{\beta 3}^{(3)}}, \widehat{z_{\beta 4}^{(3)}}, \widehat{z_{\beta 5}^{(3)}}, \widehat{z_{\beta 2}^{(4)}}, \widehat{z_{\beta 3}^{(4)}}, \widehat{z_{\beta 4}^{(4)}}, \widehat{z_{\beta 5}^{(4)}}$. These solutions are in terms of a_1 *to* a_4, b_1 *to* b_3, c_1, c_2, and d_1. Thus, the covariate matrix $\mathbf{Z}_{\boldsymbol{\beta}}$ to the β-th treatment is given by

$$\mathbf{Z}_\beta = ((a_1,\ a_2,\ a_3, a_4, \widehat{z_{\beta5}^{(1)}})', (b_1, b_2, b_3, \widehat{z_{\beta4}^{(2)}}, \widehat{z_{\beta5}^{(2)}})', (c_1, c_2, \widehat{z_{\beta3}^{(3)}}, \widehat{z_{\beta4}^{(3)}}, \widehat{z_{\beta5}^{(3)}})', (d_1, \widehat{z_{\beta2}^{(4)}}, \widehat{z_{\beta3}^{(4)}}, \widehat{z_{\beta4}^{(4)}}, \widehat{z_{\beta5}^{(4)}})') \quad (20)$$

It is seen from (20) that infinite number of $\mathbf{Z}_i$'s exist for various values of a_1 to a_4, b_1 to b_3, c_1, c_2 and d_1. Any OCD-component to a treatment having replication 5, is given by (20). All these $\mathbf{Z}_i$'s are restricted Hadamard matrices of order 5×4. Searching the largest value of $\left\|\mathbf{Z}_\beta^{(p)}\right\| \ni z_{\beta j}^{(p)} \in [-1,1]$ i.e. searching the closures of order 4 to 5, the possible forms of $\mathbf{Z}_\beta$ are given in the following table.

Table 2:

Sl. No.	a_1	a_2	a_3	a_4	b_1	b_2	b_3	c_1	c_2	d_1
1	1	-1	1	-1	1	1	-1	1	-1	$\pm 1/4$
2					1	-1	-1	1	1	$\pm 1/4$
3					-1	-1	1	-1	1	$\pm 1/4$
4					-1	1	1	-1	-1	$\pm 1/4$
5					-1	-1	1	1	-1	$\pm 1/4$
6					1	-1	-1	-1	-1	$\pm 1/4$
7					1	1	-1	-1	1	$\pm 1/4$
8					-1	1	1	1	1	$\pm 1/4$
9	1	1	1	1	1	1	-1	1	-1	$\pm 1/4$
10					1	-1	1	1	1	$\pm 1/4$
11					-1	-1	1	-1	1	$\pm 1 / 4$
12					-1	1	-1	-1	-1	$\pm 1 / 4$
13					-1	-1	1	1	-1	$\pm 1 / 4$
14					1	-1	1	-1	-1	$\pm 1 / 4$
15					1	1	-1	-1	1	$\pm 1 / 4$
16					-1	1	-1	1	1	$\pm 1 / 4$
17	1	1	-1	-1	1	-1	1	1	-1	$\pm 1 / 4$
18					1	-1	-1	1	-1	$\pm 1 / 4$
19					-1	1	-1	-1	1	$\pm 1 / 4$
20					-1	1	1	-1	1	$\pm 1 / 4$
21					-1	1	-1	1	-1	$\pm 1 / 4$
22					1	-1	-1	-1	1	$\pm 1 / 4$
23					1	-1	1	-1	1	$\pm 1 / 4$
24					-1	1	1	1	-1	$\pm 1 / 4$
25	-1	1	-1	1	-1	-1	1	-1	1	$\pm 1 / 4$
26					-1	1	1	-1	-1	$\pm 1 / 4$
27					1	1	-1	1	-1	$\pm 1 / 4$
28					1	-1	-1	1	1	$\pm 1 / 4$
29					1	1	-1	-1	1	$\pm 1 / 4$
30					-1	1	1	1	1	$\pm 1 / 4$
31					-1	-1	1	1	-1	$\pm 1 / 4$
32					1	-1	-1	-1	-1	$\pm 1 / 4$

33	-1	1	1	-1	-1	-1	1	-1	1	$\pm 1/4$
34					-1	1	-1	-1	-1	$\pm 1/4$
35					1	1	-1	1	-1	$\pm 1/4$
36					1	-1	1	1	1	$\pm 1/4$
37					1	1	-1	-1	1	$\pm 1/4$
38					-1	1	-1	1	1	$\pm 1/4$
39					-1	-1	1	1	-1	$\pm 1/4$
40					1	-1	1	-1	-1	$\pm 1/4$
41	-1	-1	1	1	-1	1	-1	-1	1	$\pm 1/4$
42					-1	1	1	-1	1	$\pm 1/4$
43					1	-1	1	1	-1	$\pm 1/4$
44					1	-1	-1	1	-1	$\pm 1/4$
45					1	-1	1	-1	1	$\pm 1/4$
46					-1	1	1	1	-1	$\pm 1/4$
47					-1	1	-1	1	-1	$\pm 1/4$
48					1	-1	-1	-1	1	$\pm 1/4$

A lemma follows, for use in sequel.

Lemma 3. The $\mathbf{Z}_\beta$'s given in the Table 2 are the possible optimal OCD-components of CRD (ν , $n_1, n_2, \ldots, n_v$) to the β-th treatment having replication 5 i.e. $n_\beta = 5$.

Proof of the lemma is given in Appendix-C.

2.3. Construction of OCD-component for n_β = odd in CRD where $\mathbf{H}_{n_\beta - 1}$exists

Let n_β (odd) be one of the n_i's . Under the condition that $\mathbf{H}_{n_\beta - 1}$ exists, a more general construction of OCD-component of CRD (ν, $n_1, n_2, \ldots, n_v$) to the β-th treatment, follows as a lemma.

Lemma 4. An optimum allocation of OCD-component to the β-th treatment with replication n_β (odd) where $\mathbf{H}_{n_\beta - 1}$exists, is given by

$$\mathbf{Z}_\beta = \mathbf{Z}_\beta^{n_\beta \times (n_\beta - 1)} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \ldots. & \mathbf{h}_{n_\beta - 2} & \delta / (n_\beta - 1)\mathbf{1}_{n_\beta - 1} \\ 0 & 0 & \ldots. & 0 & -\delta \end{bmatrix}$$

where $\delta = \pm 1$ and $\mathbf{h}_j$'s are given by $\mathbf{H}_{n_\beta - 1} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n_\beta - 2}, \mathbf{1}_{n_\beta - 1}]$; $j = 1, 2, \ldots, n_\beta$.

Proof of the lemma is given in Appendix-D.

3. Construction of OCD for D-Optimal CRD.

In this section there are 3 sub-sections, focusing on the constructions of c-OCD for D-optimal CRD $(v, n_1, n_2, \ldots, n_v)$, permitting the independent estimation of treatment-effects and covariate-effects among the component CRD's, when at least one of n_i's is equal to 3, 5 or q

odd such that there exists $\mathbf{H}_{q-1}$. By the definition of restricted Hadamard matrix in Definition 3, CD-components should be a restricted Hadamard matrix so as treatment-effects and covariate-effects are independently estimable. Further, to obtain the maximum det.($\mathbf{I(\theta)}$), each of OCD-components of the CRD $(v, n_1, n_2, \dots, n_v)$ to the β-th treatment are to be closures of order c to n_β for all β.

3.1. OCD for D-optimal CRD ($n_\beta = 3$)

For a 2 OCD-component of a D-optimal CRD $(v, n_1, n_2, \dots, n_v)$ D^*, permitting the independent estimation of treatment-effects and covariate-effects, let $\mathbf{Z_i}$ be the OCD-component to the i-th treatment of the D-optimal CRD D^*.

Since D^* permits the independent estimation of treatment-effects and covariates-effects, then

$$\mathbf{T}_i^* = \mathbf{0}' \ \forall \text{ i } i.e. \boldsymbol{X}^{*\prime}\mathbf{Z}^* = 0,$$
$$\mathbf{Z}^{*\prime}\mathbf{Z}^* = (\mathbf{Z}_1', \dots, \mathbf{Z}_v')(\mathbf{Z}_1', \dots, \mathbf{Z}_v')' = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}, \text{ say.}$$

And further, det.($\mathbf{I}^*(\boldsymbol{\theta})$)of $D^* = (\prod_{i=1}^{v} n_i)$ det.$(\mathbf{Z}^{*\prime}\mathbf{Z}^* - \mathbf{T}^{*\prime}\mathbf{N}^{*-1}\mathbf{T}^*)$
$= (\prod_{i=1}^{v} n_i)\alpha_1\alpha_2 \geq$ det. ($\mathbf{I}^{**}(\boldsymbol{\theta})$)of D^{**}

where D^{**} belongs to the class of all the competent CRD's $(v, n_1, n_2, \dots, n_v)$, permitting the independent estimation of treatment-effects and covariate-effects and $\mathbf{I}^{**}(\boldsymbol{\theta})$is the information matrix for $\boldsymbol{\theta}$ through D^{**}. Taking $\mathbf{Z}_i (i = 1, 2, \dots, v)$ as the OCD-component to the i-th treatment and $\mathbf{Z}_{(v+1)}$= ((1/2, 1/2, -1)$'$, (1, 1, 0)$'$) as that to the (v + 1)-th treatment of a new CRD $(v, n_i\,; i = 1, 2, \dots, v\,;\, n_{v+1} = 3)D$, say.

$$\text{Then, } \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1', \ \dots, \mathbf{Z}_v', \ \begin{pmatrix} 1/2 & 1 \\ 1/2 & 1 \\ -1 & 0 \end{pmatrix}' \end{bmatrix}' \tag{21}$$

$$\mathbf{T}_i = \mathbf{0}' \ \forall \ i = 1, 2, \dots, v+1 \text{ i. e. } \mathbf{X}'\mathbf{Z} = \mathbf{0} \tag{22}$$

$$\mathbf{Z}'\mathbf{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_v', \mathbf{Z}_{v+1}')(\mathbf{Z}_1', \dots, \mathbf{Z}_v', \mathbf{Z}_{v+1}')'$$

$$= \begin{pmatrix} \alpha_1 + 3/2 & 0 \\ 0 & \alpha_2 + 2 \end{pmatrix}, \text{ from Proposition 1, by (21) and (22).} \tag{23}$$

$$\text{Then, det.}\left(\mathbf{I(\theta)}\right)\text{of } D = (\prod_{i=1}^{v} n_i)(\alpha_1 + 3/2)(\alpha_2 + 2), \text{ by (23)} \tag{24}$$

which is the maximum of determinant of information matrix of a CRD (v+1, $n_1, n_2, \dots, n_v$, n_{v+1}=3) among all the competent CRD's accommodated with various possible OCD-components to the (v+1)-th treatment. Thus, a theorem is immediate.

Theorem 1: Existence of 2-OCD of D-optimal CRD (v,n_1,n_2, ...,n_v), permitting the independent estimation of treatment-effects and covariate-effects implies that of 2-OCD of D-optimal CRD (v+1,n_1,.....n_v,$n_{v+1} = 3$), maintaining the same estimation status.

Proof: Suppose the proposed OCD for the new resultant CRD is not an optimal allocation. Then, there exists an optimal OCD $\mathbf{Z}^{***}$ ∋ det.($\mathbf{Z}^{***\prime}\mathbf{Z}^{***}$) > ($\alpha_1$+3/2) ($\alpha_2 + 2$), by (24)

i.e. by (18) $\exists$ some $a_1, a_2, b_1 \in [-1, 1]$ $\ni$ $[\alpha_1+a_1^2+a_2^2+(a_1+a_2)^2][\alpha_2+b_1^2+\{(a_2+2a_1)\,b_1/(2a_2+a_1)\}^2 + \{(a_1-a_2)\,b_1/(2a_2+a_1)\}^2] > (\alpha_1+3/2)(\alpha_2+2)$

i.e. $[a_1^2+a_2^2+(a_1+a_2)^2]\,[b_1^2+\{(a_2+2a_1)\,b_1/(2a_2+a_1)\}^2 + \{(a_1-a_2)\,b_1/(2a_2+a_1)\}^2]>3$
since $(\delta_1+x_1)(\delta_2+x_2) > (\delta_1+y_1)(\delta_2+y_2)$ iff $x_1x_2> y_1y_2$ $\forall$ $\delta_1, \delta_2, x_1, x_2, y_1, y_2>0$

i.e. $\{2a_1^2 + 2a_2^2 + 2a_2a_1\}b_1^2\{4a_2^2 +a_1^2 + 4a_1a_2 + a_2^2 + 4a_1^2 + 4a_1a_2 + a_2^2 + a_1^2 - 2a_1a_2\}/(2a_2 + a_1)^2 > 3$ which is impossible as it is same to (B1) and (B1) is impossible.

Remark 2: By applying an induction-method for constructing OCD-components to treatments with replication number 3, the construction of 2-OCD of D-optimal CRD with replication number n_i (permitting independent estimation of treatment-effects and covariate-effects), out of which some are equal to 3, is always possible.

Note 1: By Lemma 1, $n_i \geq 3$ $\forall$ $i = 1, 2, \ldots, v$, otherwise 2-OCD of D-optimal CRD ($v, n_1, n_2, \ldots, n_v$), permitting independent estimation of treatment-effects and covariate-effects, never exists.

Hadamard matrix $\mathbf{H}_{n_i}$(if exists for $n_i \geq 3$ $\forall$ i) can be rewritten as $H_{n_i} = \left[\mathbf{h}_1^{(i)}, \ldots, \mathbf{h}_{n_i-1}^{(i)}, \ \mathbf{1}_{n_i}\right]$. The vectors $\mathbf{h}_1^{(i)}, \ldots, \mathbf{h}_{n_i-1}^{(i)}$ are closures of order n_i-1 to n_i. Therefore, a 2-OCD of D-optimal CRD ($v, n_1, n_2, \ldots, n_{v-1}$) exists, permitting independent estimation of treatment-effects and covariate-effects. Applying Theorem 1, a corollary follows.

Corollary 1: Existence of 2-OCD of D-optimal CRD ($v, n_1, \ldots, n_{v-1}, n_v$=3) permitting the independent estimation of treatment-effects and covariate-effects, is always possible, provided $\mathbf{H}_{n_i}$ exists and $n_i \geq 3$ $\forall$ i= 1, 2, …., v-1.

Using the OCD-component in Sr. No. 17, Table 1, an example of Theorem 1 is given below:

Example 1: Given a 2-OCD $\mathbf{Z}^* = (\mathbf{Z_1'}, \mathbf{Z_2'})'$ of D-optimal CRD ($v = 2, n_1 = n_2 = 4$) permitting independent estimation of treatment-effects and covariate-effects, taking the OCD-component in Sr. No. 17, Table 1 as OCD-component of the 3rd treatment of a CRD ($v = 2, n_1 = n_2 = 4, n_3 = 3$), the det. of the information matrix i.e. $|I(\boldsymbol{\theta})|$ attains maximum value $(4^2\times3)(8+2)(8+3/2)$ and the resultant CRD maintains the same estimation status where the OCD is given by $\mathbf{Z} = (\mathbf{Z_1'}, \mathbf{Z_2'}, \mathbf{Z_3'})'$; $\mathbf{Z_1} = \mathbf{Z_2} = ((-1, -1, 1, 1)', (-1, 1, -1, 1)')$, $\mathbf{Z_3} = ((1, 0, -1)', (1/2, -1, 1/2)')$.

3.2. OCD for D-optimal CRD (n_β = 5).

For a 4-OCD of a CRD ($v, n_1, n_2, \ldots, n_v$) D^*, permitting the independent estimation of treatment-effects and covariate-effects, let $\mathbf{Z}_i$ be the OCD-component to the i-th treatment of the CRD D^*. Since D^* permits the independent estimation of treatment-effects and covariates-effects, then

$$\mathbf{T}_i^* = \mathbf{0}' \ \forall \text{ i i.e. } \mathbf{X}^{*\prime}\mathbf{Z}^* = 0,$$

$$\mathbf{Z}^{*\prime}\mathbf{Z}^* = \sum \mathbf{Z}_i'\mathbf{Z}_i = \text{diag.}(\alpha_1, \ldots, \alpha_4), \text{ say}$$

and further, det. $(\mathbf{I}^*(\boldsymbol{\theta}))$ of D^* = det.$(\mathbf{I}^*(\boldsymbol{\theta}))$ of $D^* = (\prod_{i=1}^{v} n_i)$ det.$(\mathbf{Z}^{*\prime}\mathbf{Z}^* - \mathbf{T}^{*\prime}\mathbf{N}^{*-1}\mathbf{T}^*)$
$= (\prod_{i=1}^{v} n_i)(\prod_{i=1}^{4} \alpha_i) \geq$ det.$(\mathbf{I}^{**}(\boldsymbol{\theta}))$ of D^{**},

where D^{**} belongs to the class of all competent CRD's (v, n_1, n_2, ..., n_v) accommodated with 4-OCD permitting the independent estimation of treatment-effects and covariate-effects and $\mathbf{I}^{**}(\boldsymbol{\theta})$is the information matrix for $\boldsymbol{\theta}$ through D^{**}.

Taking $\mathbf{Z}_i$ (i = 1, 2, …..,v) as the OCD-component to the i-th treatment and $\mathbf{Z}_{v+1}$ = ((1, –1, 1, –1, 0)$'$, (1, 1, -1, -1, 0)$'$, (1, -1, -1, 1, 0)$'$, (1/4, 1/4, 1/4, 1/4, -1)$'$) as that to the (v+1)-th treatment of a new CRD (v+1, n_1, n_2, …, n_v, n_{v+1}= 5) D, say.

Then, for the design D, $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_v', \mathbf{Z}_{v+1}')'$ (25)

$$\mathbf{T}i = \mathbf{0}' \forall \text{ i} = 1, 2, ..., v+1 \quad \text{i.e. } \mathbf{X}'\mathbf{Z} = \mathbf{0} \tag{26}$$

Now, $\mathbf{Z}'\mathbf{Z}$ = diag. (α_1, α_2, α_3, α_4) + $\mathbf{Z}_{v+1}'\mathbf{Z}_{v+1}$, from Proposition 1

$$= \text{diag. } (\alpha_1+4, \alpha_2+4, \alpha_3+4, \alpha_4+5/4), \text{ by (25) and (26).} \tag{27}$$

Then, det.($\mathbf{I}(\boldsymbol{\theta})$) of D = $(\prod_{i=1}^{v+1} n_i)(\alpha_1+4)(\alpha_2+4)(\alpha_3+4)(\alpha_4+5/4)$, by (27) (28)

which is the maximum of determinant of information matrix of the CRD (v+1, n_1, n_2, …, n_v, n_{v+1}= 5) among all the competent CRD's accommodated with various possible OCD-components to the (v+1)-th treatment. Thus, the following theorem is immediate.

Theorem 2: Existence of 4-OCD of D-optimal CRD (v, n_1, n_2, ..., n_v) implies that of 4-OCD of D-optimal CRD (v+1, n_1, n_2, ..., n_v, n_{v+1}= 5) maintaining the same estimation status.

Proof: Suppose the proposed OCD for the new resultant CRD D is not an optimal allocation. Then, there exists an OCD $\mathbf{Z}^{***}$ ∋ det.($\mathbf{Z}^{***'}\mathbf{Z}^{***}$) > $(\alpha_1+4)(\alpha_2+4)(\alpha_3+4)(\alpha_4+5/4)$, by (28) *i.e.* by (20), ∃ some a_1 to a_4, b_1 to $b_3, c_1, c_2,\ d_1 \in [-1,1]$ ∋

$$[\alpha_1+\sum_j \left(\widehat{z_{\beta J}^{***(1)}}\right)^2][\alpha_2+\sum_j \left(\widehat{z_{\beta J}^{***(2)}}\right)^2][\alpha_3+\sum_j \left(\widehat{z_{\beta J}^{***(3)}}\right)^2][\alpha_4+\sum_j \left(\widehat{z_{\beta J}^{***(4)}}\right)^2] > (\alpha_1+4)(\alpha_2+4)(\alpha_3+4)(\alpha_4+5/4)$$

i.e. $$[\sum_j \left(\widehat{z_{\beta J}^{***(1)}}\right)^2][\sum_j \left(\widehat{z_{\beta J}^{***(2)}}\right)^2][\sum_j \left(\widehat{z_{\beta J}^{***(3)}}\right)^2][\sum_j \left(\widehat{z_{\beta J}^{***(4)}}\right)^2] > 80$$

since $\prod_{i=1}^{q}(\delta_i + x_i) > \prod_{i=1}^{q}(\delta_i + y_i)$ iff $\prod_{i=1}^{q} x_i > \prod_{i=1}^{q} y_i;\ \forall\ \delta_i's,\ x_i's,\ y_i's > 0$

$$\text{i.e.}[\sum_{k=1}^{4} a_k^2 + \left(\widehat{z_{\beta 5}^{***(1)}}\right)^2][\sum_{l=1}^{3} b_l^2 + \left(\widehat{z_{\beta 4}^{***(2)}}\right)^2 + \left(\widehat{z_{\beta 5}^{***(2)}}\right)^2][c_1^2 + c_2^2 + \left(\widehat{z_{\beta 3}^{***(3)}}\right)^2 + \left(\widehat{z_{\beta 4}^{***(3)}}\right)^2 + \left(\widehat{z_{\beta 5}^{***(3)}}\right)^2][d_1^2 + \left(\widehat{z_{\beta 2}^{***(4)}}\right)^2 + \left(\widehat{z_{\beta 3}^{***(4)}}\right)^2 + \left(\widehat{z_{\beta 4}^{***(4)}}\right)^2 + \left(\widehat{z_{\beta 5}^{***(4)}}\right)^2] > 80 \tag{29}$$

which is impossible as it is same to (C1) and (C1) is impossible. Hence proved.

Remark 3: By applying an induction-method for constructing OCD-components of treatments with replication number 5, the construction of 4-OCD of D-optimal CRD with replication number n_i's (permitting independent estimation of treatment-effects and covariate-effects), out of which some are equal to 5, is always possible.

Note 2: By Lemma 1, $n_i \geq 5\ \forall\ i$ = 1, 2, …,v, otherwise 4-OCD of D-optimal CRD (v, n_1 , ..., n_v), permitting independent estimation of treatment-effects and covariate-effects, never exists.

A Hadamard matrix $\mathbf{H}_{n_i}$(if exists for $n_i \geq 5\ \forall\ i$) can be rewritten as $\mathbf{H}_{n_i} = \left[\mathbf{h}_1^{(i)}, \ldots, \mathbf{h}_{n_i-1}^{(i)},\ \mathbf{1}_{n_i}\right]$. The vectors $\mathbf{h}_1^{(i)}, \ldots, \mathbf{h}_{n_i-1}^{(i)}$ are closures of order n_i-1 to n_i. Therefore, a 4-OCD of D-optimal CRD ($v,\ n_1,\ \ldots\ ,n_{v-1}$) exists, permitting independent estimation of treatment-effects and covariate-effects. Applying Theorem 2, a corollary follows.

Corollary 2: Existence of 4 covariate-OCD of D-optimal CRD (v, n_1, …., n_{v-1}, n_v = 5), permitting the independent estimation of treatment-effects and covariate-effects, is always possible, provided $\mathbf{H}_{n_i}$ exists and $n_i \geq 5\ \forall\ i$ = 1, 2, …, v-1.

Using the OCD-component in Sr. No. 1, Table 2, an example of Theorem 2 is given below:

Example 2: Given a 4-OCD $\mathbf{Z}^*= (\mathbf{Z}_1', \mathbf{Z}_2')'$ of D-optimal CRD (v = 2, $n_1 = n_2$ = 8) permitting independent estimation of treatment-effects and covariate-effects, taking the OCD-component in Sr. No. 1, Table 2 as OCD-component of the 3rd treatment of a CRD (v = 3, $n_1 = n_2$ = 8, n_3 = 5), the det. of the information matrix i.e. $|\mathbf{I(\theta)}|$ attains maximum value ($8^2\times 5)(16+4)^3(16+5/4)$ and the resultant CRD maintains the same estimation status where the OCD is given by $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2', \mathbf{Z}_3')'$; $\mathbf{Z}_1 = \mathbf{Z}_2$ = ((1, 1, -1, -1, -1, 1, -1, 1)′, (1, 1, 1, -1, -1, -1, 1, -1)′, (1, -1, 1, 1, -1, -1, -1, 1)′, (1, 1, -1, 1, 1, -1, -1, -1)′), $\mathbf{Z}_3$ = ((1, -1, 1, -1, 0)′, (1, 1, -1, -1, 0)′ (1, -1, -1, 1, 0)′, (1/4, 1/4, 1/4, 1/4, -1)′)

3.3. OCD for D-optimal CRD (n_β= q: $\mathbf{H}_{q-1}$ ∃)

In this sub-section, under more general case of n_β odd such that $\mathbf{H}_{n_\beta-1}$ exists, the existence of OCD for D-optimal CRD would be claimed, assuming that OCD-components of the other treatments (perhaps, with even replications) are known. Using Hadamard matrices, construction of OCD for D-optimal CRD (v, n_1, n_2, ..., n_v), permitting the independent estimation of treatment-effects and covariate-effects, is proposed.

Theorem 3: Optimum allocation of c-OCD of D-optimal CRD $(v, n_1, n_2, \ldots, n_v)$ permitting the independent estimation of treatment-effects and covariate-effects, is always possible if for some n_α's ∃ $\mathbf{H}_{n_\alpha}$and for remaining n_β's ($\neq n_\alpha$), ∃ $\mathbf{H}_{n_\beta-1}$where c = min. (n_α-1, n_β-2 ∀ α, β).

Proof of the theorem is given in Appendix-E.

Remark 4: As $\mathbf{H}_{n_\alpha}$ exists, n_α is even. Of course, $\mathbf{h}_1^{(\alpha)}, \ldots\ldots, \mathbf{h}_{n_\alpha-1}^{(\alpha)}$given in (E1), are closures of order n_α-1 to n_α and consequently, are optimal OCD-components of the α-th treatment.

Remark 5: As $\mathbf{H}_{n_\beta-1}$exists, n_β odd. The gravity of Theorem 3 is that when some of the replications of the CRD are odd, OCD of D-optimal CRD can be constructed, permitting independent estimation of treatment-effects and covariate-effects.

An example of Theorem 3 is given below.

Example 3: For a CRD (v =2, n_1 = 4, n_2= 3) to be D-optimal, as $\mathbf{H}_4$ and $\mathbf{H}_2$ exist, c = min.(n_1-1, n_2-2) = 1. Then, $\mathbf{Z}^{(1)} = (+1, +1, -1, -1)'$, $\mathbf{Z}^{(2)} = (+1, -1, 0)'$ using the Hadamard

matrices of $n = 4$ and 2 in the Table 2, Hedayat and Wallis (1978). The determinant of $\mathbf{I}(\boldsymbol{\theta})$ of the CRD with covariate design $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2')'$ obtains its maximum $2\times4(6\text{-}1)$ i.e. 40 over all possible OCD accommodated to the CRD.

References

Chadjiconstantinidis, S and Moyssiadis, C. (1991). Some D-optimal odd-equireplicated designs for covariate model. *Journal of Statistical Planning and Inference,* **28**, 83-93.

Das, K., Mandal, N. K. and Sinha, B. K. (2003). Optimal experimental designs for model with covariate. *Journal of Statistical Planning and Inference,* **115**, 273-285.

Das, P., Dutta, G.; Mandal, N. K. and Sinha, B. K. (2015). *Optimal Covariate Designs.* Springer, India (ISBN: 978-81-322-2460-0, ISBN (eBook)): 978-81-322-2461-7).

Dey, A. and Mukerjee, R. (2006). D-optimal designs for covariate models. *Statistics*, **40**(**4**), August 297-305.

Dutta, G., Das, P. and Mandal, N. K. (2014). D-Optimal designs for covariate models, *Communications in Statistics*, **43**, 165-174.

Hedayat, A and Wallis, W. D. (1978). Hadamard matrices and their applications. *Annals of Statistics,* **6**(**6**), 1184-1238.

Liski, E. P, Mandal, N. K., Shah, K. R. and Sinha, B. K. (2002). *Topics in Optimal Design*. Lectures Notes in Statistics, **163**, Springer-Verlag New York.

Lopes Troya, J. (1982). Optimal designs for covariate model. *Journal of Statistical Planning and Inference,* **6**, 373-419.

Rao, P. S. S. N. V. P., Rao, S. B., Saha, G.M. and Sinha, B. K. (2003). Optimal designs for covariates' models and mixed orthogonal arrays. *Electronic Notes in Discrete Mathematics,* **15**, 155-158.

Wierich, W. (1984). Konkrete optimale versuchspläne fűr ein lineares modell mit einem qualitativen und zwei quantitativen einflussfaktoren. *Metrika*, **31**, 285-301.

APPENDIX-A

Proof of Lemma 1:

Suppose c be the maximum number of columns in the restricted Hadamard matrix **A** with m rows. Let $\boldsymbol{a}_{(p)} = (a_{1p}, a_{2p}, \ldots, a_{mp})'$ be the p-th column of **A** By the definition of restricted Hadamard matrix,

$$\sum_i a_{ip} = 0 \;\forall\; p = 1, 2, \ldots, c \text{ and} \tag{A1}$$

$$\sum_i a_{ip}a_{iq} = 0 \;\forall\; p \neq q = 1, 2, \ldots, c. \tag{A2}$$

In (A1) and (A2) there are $c(c+1)/2$ equations involving mc unknown variables. When $c = 1$, there exists only 1 linear equation involving m unknown variables which has non-zero solutions belong to $[-1, 1]$. When $c = \beta$; $\beta = 2, 3, \ldots, m-1$ there exist $\beta(\beta+1)/2$ linearly independent equations (independent is due to (A2)) involving βm unknown variables. Using the solutions for $c = 1, 2, \ldots, \beta-1$, $\exists\ \beta(\beta+1)/2 - (\beta-1)\beta/2$, *i.e.* β linearly independent homogeneous linear equations involving m unknown variables, which have non-zero solutions $\in [-1, 1]$. If possible, when $c = m$, there exist $m(m+1)/2$ linearly independent equations involving m^2 unknown variables. Using the solutions for $c = 1, 2, \ldots, m\text{-}1$, there are $m(m+1)/2 - (m-1)m/2$ *i.e.* m linearly independent homogeneous linear equations involving m unknown variables which has no non-zero solution, because any system of linearly independent

homogeneous linear equations $\mathbf{B}_{n\times n}\mathbf{x}= \mathbf{0}$ has only zero solution, which is not of our interest. When $c = m + \gamma$ for any $\gamma = 1, 2, \ldots$,

$$\text{rank } (\mathbf{A}) \leq m \text{ since } c > m. \tag{A3}$$

By (A2), all the column vectors of $\boldsymbol{A}$ are orthogonal and consequently, linearly independent which contradicts to (A3). Thus, the case of $c = m + \gamma$ never arises. Hence proved.

APPENDIX-B

Proof of Lemma 2:

Suppose $\mathbf{Z}_\beta$ given in the Table 1 is not an optimal OCD-component of CRD to the β-th treatment with replication 3. Then, there exists an optimal OCD-component $\mathbf{Z}^*_\beta$ to the β-th treatment ∋ det.($\mathbf{Z}^{*\prime}_{\boldsymbol{\beta}}\mathbf{Z}^*_{\boldsymbol{\beta}}$) > det.($\mathbf{Z}'_{\boldsymbol{\beta}}\mathbf{Z}_{\boldsymbol{\beta}}$)

i.e. $\det.\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} > 2 \times 3/2$, by (18), where $A = a_1^2 + a_2^2 + (a_1 + a_2)^2$ and

$$B = b_1^2 + [(a_2 + 2a_1)b_1/(2a_2 + a_1)]^2 + [(a_1 - a_2)b_1/(2a_2 + a_1)]^2$$

i.e. $4(a_1^2 + a_2^2 + a_1a_2)^2\, b_1^2 > [2a_2 + a_1]^2$. (B1)

In order to get closure of order 2 to 3 for det. ($\mathbf{Z}^{*\prime}_{\boldsymbol{\beta}}\mathbf{Z}^*_{\boldsymbol{\beta}}$) to be at maximum, putting the possible largest magnitude of a_1 and a_2 , 1 and -1 (or -1 and 1) respectively in (B1) (but not 1 and 1 (or –1 and –1) because $\widehat{z^{(1)}_{\beta 3}} = -(a_1 + a_2) \in [-1, 1]$), we have $4b_1^2 > 1$ *i.e.* -1/2 > b_1 > 1/2. So, (B1) is not possible, contradicting to the possible forms of $\mathbf{Z}_{\boldsymbol{\beta}}$ given in Sl. No. 13 and 15 (or Sl. Nos. 14 and 16), Table 1, where $a_1 = 1, a_2 = -1$ (or $a_1 = -1, a_2 = 1$), $b_1 = \pm 1/2$. Hence $\mathbf{Z}_{\boldsymbol{\beta}}$'s given in the Table 1 are the optimal OCD-components of CRD to the β-th treatment with replication 3 *i.e.* the closures of order 2 to 3.

APPENDIX-C

Proof of Lemma 3:

Suppose $\boldsymbol{Z_\beta}$ given in the Table 2 is not optimal OCD-component of CRD(v, $n_1, n_2, \ldots, n_v$) to the β-th treatment with replication 5. Then, ∃ an optimal OCD-component $\mathbf{Z}^*_{\boldsymbol{\beta}}$ to the β-th treatment ∋ det.($\mathbf{Z}^{*\prime}_{\boldsymbol{\beta}}\mathbf{Z}^*_{\boldsymbol{\beta}}$) > det.($\mathbf{Z}'_{\boldsymbol{\beta}}\mathbf{Z}_{\boldsymbol{\beta}}$)

i. e. $[\sum_j \left(\widehat{z^{*(1)}_{\beta J}}\right)^2][\sum_j \left(\widehat{z^{*(2)}_{\beta J}}\right)^2][\sum_j \left(\widehat{z^{*(3)}_{\beta J}}\right)^2][\sum_j \left(\widehat{z^{*(4)}_{\beta J}}\right)^2] > 4^3 \times 5/4$, by (20)

i.e. $[\sum_{k=1}^4 a_k^2 + \left(\widehat{z^{*(1)}_{\beta 5}}\right)^2][\sum_{l=1}^3 b_l^2 + \left(\widehat{z^{*(2)}_{\beta 4}}\right)^2 + \left(\widehat{z^{*(2)}_{\beta 5}}\right)^2][c_1^2 + c_2^2 + \left(\widehat{z^{*(3)}_{\beta 3}}\right)^2 + \left(\widehat{z^{*(3)}_{\beta 4}}\right)^2 + \left(\widehat{z^{*(3)}_{\beta 5}}\right)^2][d_1^2 + \sum_{u=2}^5 \left(\widehat{z^{*(4)}_{\beta u}}\right)^2] > 80.$ (C1)

In order to get closures of order 4 to 5 for det.($\mathbf{Z}^{*\prime}_{\boldsymbol{\beta}}\mathbf{Z}^*_{\boldsymbol{\beta}}$) to be at maximum, putting the largest possible magnitude of a_1 to a_4 , b_1 to b_3, c_1, c_2 and d_1 satisfying (19), say, $a_1 = a_2$

= 1, $a_3 = a_4 = -1$, $b_1 = 1$, $b_2 = -1$, $b_3 = c_1 = 1$, $c_2 = -1$, $d_1 = 1/4$. Obviously, $\widehat{z_{\beta 5}^{*(1)}} = 0$, $\widehat{z_{\beta 4}^{*(2)}} = -1$, $\widehat{z_{\beta 5}^{*(2)}} = 0$, $\widehat{z_{\beta 3}^{*(3)}} = -1$, $\widehat{z_{\beta 4}^{*(3)}} = 1$, $\widehat{z_{\beta 5}^{*(3)}} = 0$, $\widehat{z_{\beta 1}^{*(4)}} = \cdots = \widehat{z_{\beta 4}^{*(4)}} = 1/4$, $\widehat{z_{\beta 5}^{*(4)}} = -1$. Then, (C1) gives 5/4 > 5/4. So, (C1) is not possible, contradicting to our assumption for existence of $\mathbf{Z}_{\boldsymbol{\beta}}^{*}$ ∋ det.($\mathbf{Z}_{\boldsymbol{\beta}}^{*\prime}\mathbf{Z}_{\boldsymbol{\beta}}^{*}$) > det.($\mathbf{Z}_{\boldsymbol{\beta}}^{\prime}\mathbf{Z}_{\boldsymbol{\beta}}$). Hence $\mathbf{Z}_{\boldsymbol{\beta}}$'s given in the Table 2 are the optimal OCD-components of CRD to the β-th treatment with replication 5 *i.e.* the closures of order 4 to 5.

APPENDIX-D

Proof of Lemma 4:

By the definition of Hadamard matrix,

$\| \mathbf{h}_j \| = n_\beta\text{-}1 \; \forall \, j = 1,2, \ldots, n_\beta\text{-}2$ and $\| \delta/(n_\beta - 1)\mathbf{1}_{\boldsymbol{n_\beta - 1}} \| = 1/(n_\beta\text{-}1)$

$\Rightarrow \|\mathbf{Z}_{\boldsymbol{\beta}}^{(j)}\| = n_\beta\text{-}1 \; \forall \; j$ and $\|\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}\| = n_\beta/(n_\beta - 1)$, since every $\mathbf{Z}_{\boldsymbol{\beta}}^{(j)}$ has entries "± 1" except one entry "0" which cannot be non-zero entry as $\mathbf{Z}_{\boldsymbol{\beta}}^{(j)\prime}\mathbf{1}_{(\boldsymbol{n_\beta - 1})} = 0$

$\Rightarrow \mathbf{Z}_{\boldsymbol{\beta}}^{(\mathbf{1})}, \ldots, \mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 2})}$ are closures of order n_β-2 to n_β, since $\|\mathbf{Z}_{\boldsymbol{\beta}}^{(j)}\| > \|\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}\| \; \forall$ j (D1)

and $\mathbf{Z}_{\boldsymbol{\beta}}^{(\mathbf{1})}, \ldots, \mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}$ are closures of order n_β-1 to n_β among all possible OC-components to the β-th treatment. (D2)

Suppose (D2) is not true *i.e.* $\mathbf{Z}_{\boldsymbol{\beta}}$ is not an optimal OCD-component of CRD to the β-th treatment. Then there exists a $(n_\beta - 1)$-th orthogonal covariate $\overline{\mathbf{Z}}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}$ such that

$$\|\overline{\mathbf{Z}}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}\| > \|\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}\| = n_\beta/(n_\beta\text{-}1) \quad \text{(D3)}$$

$$\text{and } \overline{\mathrm{Z}}_{\beta}^{(j)\prime}\overline{\mathrm{Z}}_{\beta}^{(n_\beta - 1)} = 0 \; \forall \, j, \text{ by the orthogonal of covariates.} \quad \text{(D4)}$$

Then, (D4) gives $\overline{\mathbf{Z}}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})} = (gd, gd, \ldots, gd, -g(n_\beta - 1)d)'$; $g = 1$ or -1; d a constant, otherwise $\mathbf{X}'\mathbf{Z} \neq 0$.

As every component of $\overline{\mathbf{Z}}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})} \in [-1, 1]$, $|-g(n_\beta\text{-}1)d|$ can take maximum 1 for $\|\overline{\mathbf{Z}}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}\|$ to be maximized (otherwise, no alternative). So, $d = 1/|g(n_\beta\text{-}1)| = 1/(n_\beta\text{-}1)$ as $g = \pm 1$.

Consequently, $\widehat{\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}} = (g/(n_\beta\text{-}1), \ldots, g/(n_\beta - 1), -g)'$

$$\Rightarrow \|\widehat{\mathbf{Z}_{\boldsymbol{\beta}}^{(\boldsymbol{n_\beta - 1})}}\| = (n_\beta - 1)/(n_\beta - 1)^2 + 1 = n_\beta/(n_\beta\text{-}1). \quad \text{(D5)}$$

Now, (D3) and (D5) gives $n_\beta/(n_\beta - 1) > n_\beta/(n_\beta - 1)$ which is absurd, contradicting to our assumption.

APPENDIX-E

Proof of Theorem 3:

WOLG, suppose for $n_1, n_2 \ldots\ldots, n_s \exists \mathbf{H}_{n_\alpha} \forall \alpha = 1, 2, \ldots, s$ and for $n_{s+1}, \ldots, n_v \exists \mathbf{H}_{n_\beta - 1} \forall \beta = s+1, \ldots, s+s_1 = v$. Then, $\mathbf{H}_{n_\alpha}$ and $\mathbf{H}_{n_\beta - 1}$ can be rewritten as follows.

$$\mathbf{H}_{n_\alpha} = [\mathbf{h}_1^{(\alpha)}, \ldots, \mathbf{h}_{n_\alpha - 1}^{(\alpha)}, \mathbf{1}_{n_\alpha}] \text{ and } \mathbf{H}_{n_\beta - 1} = [\mathbf{h}_1^{(\beta)}, \ldots, \mathbf{h}_{n_\beta - 2}^{(\beta)}, \mathbf{1}_{n_\beta - 1}]. \quad \text{(E1)}$$

$$\text{Consider } \mathbf{Z} = [\mathbf{Z}_1', \ldots, \mathbf{Z}_s', \mathbf{Z}_{s+1}', \ldots, \mathbf{Z}_v']' \quad \text{(E2)}$$

as OCD of the proposed CRD D where $\mathbf{Z}_\alpha$ and $\mathbf{Z}_\beta$ are the OCD-component to the α-th treatment and the β-th treatment of D respectively and given by $\mathbf{Z}_\alpha = [\mathbf{h}^{(\alpha)1}, \mathbf{h}^{(\alpha)2}, \ldots, \mathbf{h}^{(\alpha)c}]$ and $\mathbf{Z}_\beta = [\mathbf{h}^{(\beta)1}, \mathbf{h}^{(\beta)2}, \ldots, \mathbf{h}^{(\beta)c}]$ respectively where $\mathbf{h}^{(\alpha)l}$'s $(l = 1, 2, \ldots, c)$ are any c out of the first $n_\alpha - 1$ columns of $\mathbf{H}_{n_\alpha}$ given in (E1) and $\mathbf{h}^{(\beta)l}$'s $(l = 1, 2, \ldots\ldots, c)$ are any c out of the first $n_\beta - 2$ columns of $\mathbf{H}^*_{n_\beta - 1} = \begin{bmatrix} \mathbf{h}_1^{(\beta)}, \ldots, \mathbf{h}_{(n_\beta - 2)}^{(\beta)}, & (\delta/(n_\beta - 1))\mathbf{1}_{(n_\beta - 1)} \\ 0, \ \ldots\ldots, \ 0, & -\delta \end{bmatrix}$ and $\delta = \pm 1$.

Then, $\mathbf{Z}'\mathbf{Z} = \sum_{\alpha=1}^{s} \mathbf{Z}_\alpha'\mathbf{Z}_\alpha + \sum_{\beta=s+1}^{v} \mathbf{Z}_\beta'\mathbf{Z}_\beta$

$= (\sum_{\alpha=1}^{s} n_\alpha)\mathbf{I}_c + \left(\sum_{\beta=s+1}^{v}(n_\beta - 1)\right)\mathbf{I}_c$, from Proposition 1 and Proposition 2, by the properties of Hadamard matrices $\mathbf{H}_{n_\alpha}$ and $\mathbf{H}_{n_\beta - 1}$; $c \leq n_\alpha - 1$ and $n_\beta - 2$,

$= a\mathrm{I}_C$; $a = n - s_1$ since $(\sum_{\alpha=1}^{s} n_\alpha) + \sum_{\beta=s+1}^{v}(n_\beta - 1) = n - s_1$.

Further, $\mathbf{T}_\alpha = \mathbf{0}' \ \forall \alpha$ and $\mathbf{T}_\beta = \mathbf{0}' \ \forall \beta$ which ensure the independent estimation of treatment-effects and covariate-effects. Now, det.$(\mathbf{I}(\boldsymbol{\theta}))$ of $D = (\prod_{i=1}^{v} n_i)\det.(a\mathbf{I}_c) = (\prod_{i=1}^{v} n_i)a^c$ which is the maximum among the determinants of all information matrices of competent CRD's accommodated with any OCD.

Suppose there exist an optimal OCD $\mathbf{Z}^* = [\mathbf{Z}_1^{*\prime}, \ldots, \mathbf{Z}_s^{*\prime}, \mathbf{Z}_{s+1}^{*\prime}, \ldots, \mathbf{Z}_v^{*\prime}]'$ of the CRD D such that det.$(\mathbf{I}^*(\boldsymbol{\theta})$ accommodating with $\mathbf{Z}^*) >$ det.$(\mathbf{I}(\boldsymbol{\theta})$ accommodating with $\mathbf{Z})$

$$\textit{i.e. } (\prod_{i=1}^{v} n_i)\det.[\mathrm{diag.}(a_1, \ldots, a_c)] > (\prod_{i=1}^{v} n_i)a^c \quad \text{(E3)}$$

since $\mathbf{Z}^{*'}\mathbf{Z}^*$ is a diagonal matrix diag. $(a_1, \ldots, a_c)$, say,

$$\textit{i.e. } \prod_{p=1}^{c} a_p > a^c. \quad \text{(E4)}$$

As all the entries of covariate $\in [-1, 1]$, $n_\beta (\beta = s+1, s+2, \ldots, v)$ are odd (since $\mathbf{H}_{n_\beta - 1} \exists$) and $\mathbf{1}_{n_\beta}'\mathbf{X}_\beta = \mathbf{0}'$, each of c covariates to the α-th treatment i.e. $\mathbf{Z}_\alpha^{(1)}, \ldots, \mathbf{Z}_\alpha^{(c)}$ contains n_α entries "± 1" and that to the β-th treatment contains n_β-1 entries "± 1" and another entry "0". So, by the notation of a_p's in (E3), $a_p \leq \sum_{\alpha=1}^{s} n_\alpha + \sum_{\beta=s+1}^{v}(n_\beta - 1) \ \forall$ p

$= n - s_1$ which contradicts (E4). Hence proved.

Acknowledgements: The author is thankful to the Referee for his/her valuable comments which help me to bring the manuscript upto this stage.

Special Proceeding: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 177-184

Opportunities and Challenges in Statistics and Data Science

Yanrong Ji[1] and Ramana V. Davuluri[2]
[1]*Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*
[2]*Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.*

Received: 16 August 2021; Revised: 03 September 2021; Accepted: 06 September 2021

Abstract

Statisticians play critical roles in various fields of study that deal with data and decision making in the face of uncertainty. The rise of recent 'Data Science' offer exceptional opportunity to Statisticians because of the universal relevance of statistical methods in the interpretation of data. However, there are some challenges to become an impactful "Data Scientist" and/or "Statistician". The core of this article will be a summary of some recent research projects, through which we wish to demonstrate that statistics together with information technology makes an essential contribution to the emerging fields of 'precision medicine' and 'genomics' research. Finally, we offer our thoughts on how deep-learning methods might best be adapted for data-scarce scenarios to achieve exceptional performance.

Key words: Bioinformatics; Genomics; Machine Learning; Deep Learning.

1. Introduction

The genetic alphabet – consisting of just four letters A, C, G, and T; forms a simple and elegant language by combining redundancy and utility in the genomes. For example, a gene is a sentence of sequence of words (for example, exons and introns, like "GCTGCTGCAGAA…CTGCCTAGA", which has a certain biological meaning or function. In the most straightforward case, a gene is translated into a specific protein, made of amino acids. The information in a gene is organized into three-letter words called codons, 64 possible triplets from the 4 genetic letters, which translate into 20 different amino acids and 1 stop codon, allowing redundancy. Deciphering the language of DNA for these and other hidden instructions in translating the information from DNA to RNA and RNA to protein has been one of the ultimate goals of biological research (Portin, 2014). Statistical machine learning methods and computer systems have played critical role in interpreting the language of DNA in the human genome.

2. Deciphering the Language of Protein-coding DNA

In early days of the human genome sequencing phase, most of the gene prediction programs ((Zhang, 2002; Davuluri and Zhang, 2003)) have been mainly trained to predict coding exons, stretches of gene sequences that get translated into sequence of amino acids forming a specific protein. The crux of these programs are statistical models that calculates the

Corresponding Author: Ramana V. Davuluri;
Email: Ramana.Davuluri@stonybrookmedicine.edu

probability of a given stretch of DNA is coding exon or not. For example, GenScan (Burge and Karlin, 1997) and Genie (Kulp *et. al.*, 1996) use hidden Markov models, Grail (Xu *et. al.*, 1994) uses Neural Networks and MZEF applied quadratic discriminant functions. While the gene predictions programs were successful in accurately prediction protein-coding exonic regions, prediction of gene-promoters and non-coding first exons, remained as a highly complex problem and critical gap in gene-prediction for several years during the human genome sequencing phase. FirstEF program was developed to identify first exons and promoters by scanning human DNA sequences. It operates by finding every potential first splice-donor site (using donor QDF – *quadratic discriminant function*) and promoter (using promoter QDF) and then calculating the probability that the intervening sequence is a first exon (using exon QDF). The power of FirstEF lies in its ability to identify first exons that are either CpG related or non-CpG-related, using different sets of classification models. The details of FirstEF algorithm are described in Davuluri *et al.* (Davuluri *et. al.*, 2001) and its application on human genome sequences is explained in (Davuluri, 2003). These creative and groundbreaking approaches facilitated the prediction and annotation of Pol-II promoters in human and mouse genomes.

3. Deciphering the Language of Regulatory DNA

While the genetic code that explains how DNA is translated into proteins is universal, the gene regulatory code that determines how and when the genes are expressed varies across different cell-types and tissues (Nirenberg *et. al.*, 1965). Over the past decade, Next-Generation sequencing (NGS) technologies have accelerated our ability to generate numerous epigenomic and transcriptomic datasets (Dunham *et. al.*, 2012). These datasets collectively facilitated genome-wide identification of gene regulatory regions in an unprecedented way and unveiled the complexity of the human genes and their regulation. The exponential growth of the multi-omics data, computational analysis of large datasets has become commonplace in the study of human biology and disease.

For example, in one of our earlier studies (Gupta *et. al.*, 2010), we trained and tested different state-of-art ensemble and meta classification methods for identification of Pol-II enriched promoter and Pol-II enriched non-promoter sequences, each of length 500 bp. The classification models were trained and tested on a bench-mark dataset, using a set of 39 different feature variables that are based on chromatin modification signatures and various DNA sequence features. The best performing model was implemented in a promoter prediction algorithm that was applied on seven published ChIP-seq Pol-II datasets to provide genome wide annotation of mouse gene promoters.

Application of statistical methods successfully used in the field of linguistics: Back in 1990s, assuming nucleic acid sequences as words over the alphabet of nucleotides, computational biologists have applied statistical methods borrowed from the field of linguistics to show that the DNA indeed has all the features of a human language, ranging from alphabets and lexicons to grammar and phonetics (Brendel and Busse, 1984; Head, 1987; Searls, 1992; Ji, 1999). Searls, in his pioneering work (Searls, 2002), demonstrated that the sequential interpretation of DNA sequence to structure to function could be matched to the hierarchical model of the human language. Furthermore, it was shown that the noncoding DNA sequences were more similar to natural languages than the coding regions (Mantegna *et. al.*, 1994). One crucial feature of any language is the semantic dependency on the verbal context, which refers to the surrounding text of a particular word or sentence of interest. For example, the correct

meaning polysemous words, which spell the same but are semantically different, can only be inferred by the clues in the context (Peters *et. al.*, 2018). Similarly, in non-coding promoter region, a Transcription Factor Binding Site (TFBS) can be target of different transcription factors that belong to the same family of proteins. For example, p53 family of proteins (p53, p63, p73) share the same central DNA-binding domain structure, which binds as a tetramer to consensus response elements (RE) consisting of two decameric palindromic half-site sequences (Ethayathulla *et. al.*, 2013; Kearns *et. al.*, 2016). Many studies have suggested that p53 family members recognize the same RE sequences while exhibiting selectivity in actual binding (Osada *et. al.*, 2005; Lokshin *et. al.*, 2007; Schavolt and Pietenpol, 2007). In this case, the conserved binding site sequence is polysemous in that the semantic meaning (whether it is to be bound by any isoform of p53, p63 or p73) varies and is hard to determine from the motif sequence alone without taking the context into consideration.

Application of traditional machine-learning methods: Computational approaches that combine seemingly disparate experimental data have been successful in developing accurate classification models. For example, RandomForest (Breiman, 2001) based algorithms have been receiving increased attention in the data-science field as a means of variable selection in many classification tasks in computational biology, including the selection of a subset of genetic markers (Lunetta *et. al.*, 2004; Bureau *et. al.*, 2005) and genes in microarray data analysis (Cutler and Stevens, 2006; Pang *et. al.*, 2006) relevant for the prediction of a certain disease.

We have earlier used an integrative modeling approach that combines CART (Breiman, 1984) and RandomForest to classify different Estrogen Receptor alpha (ERα) responsive promoters (Cheng *et. al.*, 2006) and SMAD target promoters (Qin *et. al.*, 2009) with reasonably good classification accuracy and reduced instability (Qin *et. al.*, 2009). Although the main goal in classification is to build a model with minimal mis-classification error in cross-validation, in these applications we were equally interested in identifying TFBSs as highly important discriminating variables, between different groups of promoter sequences. One of the main goals of our analyses is to select potential TFBS from a large feature space (>1,000) in order to build binary classifiers. RandomForest algorithm generates internal estimates of the decrease in the classifier's overall accuracy if that particular variable was not used in building the classifier. Thus, variables (TFBSs in this case) with larger importance measures can be deemed to have more power in discriminating different groups.

This integrative microarray data-analyses and statistical modeling approach facilitated the prediction of which proteins work with estrogen to contribute to breast cancer development. The computational predictions in this study indicated that the interaction of estrogen with one of seven different partner proteins determines whether the gene is activated or suppressed in breast cancer cells. This was a noteworthy machine-learning methodology breakthrough because it allowed integrative analysis of big datasets, often consisting of expression, chromatin landscaping data and TF-binding information for thousands of genomic loci, for predicting groups of TFBS (known as cis-regulatory modules – CRMs), which then could be validated by more traditional experimental biology techniques. In addition, this novel computational methodology has been applied in several other subsequent studies within our group and outside for integrative analysis of transcriptome and genome data.

Application of deep-learning methods: In recent years, many computational tools have been developed by successfully applying deep learning techniques on genomic sequence data to

study the individual aspects of *cis*-regulatory landscapes, including DNA-protein interactions (Alipanahi *et. al.*, 2015), chromatin accessibility (Kelley *et. al.*, 2016), non-coding variants (Zhou and Troyanskaya, 2015). Most methods adopted Convolutional Neural Network (CNN)-based architecture (Zou *et. al.*, 2019). Other tools focus on the sequential characteristic of DNA and attempt to capture the state dependency using Recurrent Neural Network (RNN)-based models, such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho *et. al.*, 2014) networks. Several hybrid methods were also proposed to integrate the two model architectures (Hassanzadeh and Wang, 2016; Quang and Xie, 2016; Shen *et. al.*, 2018). To better model gene regulatory elements in non-coding DNA, an ideal computational model should (i) globally take all the contextual information into account to understand polysemous CREs; (ii) develop generic understanding transferable to various tasks; (iii) generalize well when labeled data is limited. However, both CNN and RNN architectures fail to satisfy these requirements (Bengio *et. al.*, 2013; LeCun *et. al.*, 2015). CNN is usually unable to capture long-range semantic dependency due to the limited filter size. RNN models (LSTM, GRU), although able to learn long-term dependency, greatly suffer from vanishing gradient and low-efficiency problem when it sequentially processes all past states and compresses contextual information into a bottleneck with long input sequences. In addition, a significant challenge is that most existing prediction models require massive amounts of labeled data, resulting in limited performance and applicability in data-scarce scenarios.

Both CNN and RNN network architectures have their intrinsic limitations in terms of understanding whole genome as a language. Gene regulatory components separated by hundreds or even thousands of nucleotides are often found to coordinate together, which suggests the existence of distant semantic relationship within contexts. However, CNN is usually unable to capture such long-range semantic dependency, as its capability to extract features is limited by the size of the sliding kernels. In other words, CNN can effectively learn the local features but is not efficient at understanding the sequential relationship between these segments. LSTM and GRU, in contrast, specialize in modeling time-series data and are effective at capturing the dependency. Nonetheless, a major limitation of RNN is the difficulty for parallelized computation, since sequential information at time t is intrinsically dependent upon completion of all past states. This greatly limits the efficiency of training large RNN-based models with many blocks/layers. A combination of CNN and RNN still cannot bypass the sequential processing problem of RNN and is, therefore, also suboptimal when the model size gets large.

In essence, the goal for any deep learning model is to find a good distributed representation of input that is best suited for the downstream tasks (e.g. classification) (Bengio *et. al.*, 2013; LeCun *et. al.*, 2015). In NLP, such representation is often called embeddings, which are low-dimensional numeric vector representations of words or sentences that allows for arithmetic operations (Mikolov *et. al.*, 2013b). Ng (Ng, 2017) developed dna2vec as an initial method in 2017 based on the popular word2vec model (Mikolov *et. al.*, 2013a), which computes word embeddings of variable-length k-mers of input DNA sequence with a skip-gram model architecture. Essentially, each k-length fragment of input sequence will be considered as a 'word' and converted into a dense vector representation, by predicting the words immediately surrounding it. However, one major drawback of such embedding obtained is that the resulting representation will be context-independent (Peters *et. al.*, 2018). Two subsequences "ATGCCA" will always have same vector representation, whereas in reality they may mean different things. Same limitation applies to representations learned by CNN models, which are also not contextual (Yan and Guo, 2019). Even for RNN models, it becomes difficult

for the model to compress information from all previous states into a "bottleneck" when the input sequence gets long (Bahdanau *et. al.*, 2014).

To more properly train a model that can develop contextual embedding, attention mechanism was proposed so that the model should not put equal weights on everything it learned (Bahdanau *et. al.*, 2014; Yang *et. al.*, 2016). Instead, it should learn to "attend to" the important parts by properly forming a context vector, which is essentially a weighted sum of all hidden states from the encoder (Bahdanau *et. al.*, 2014). The weights, or attention, measures the alignment between the output and every encoder-hidden state. After its initial success in machine translation, attention mechanism became widely applied in various types of tasks, including image captioning (Xu *et. al.*, 2015), speech recognition (Chorowski *et. al.*, 2015) and others. Recently, attention-based networks are also applied in bioinformatics for predicting enhancer-promoter interaction (Mao *et. al.*, 2017; Hong *et. al.*, 2020). However, most of attention networks developed typically rely on use of a RNN-based architecture, which is subjected to the drawback mentioned above (Vaswani *et. al.*, 2017).

In order to address the limitations listed above, a new type of model that simultaneously possesses the strengths of both CNN and RNN and is capable of developing contextual embedding via attention mechanisms is necessary.

Adaptation of BERT for genome sequence prediction tasks: The advent of the Bidirectional Encoder Representation Transformer (BERT) model lead the NLP research to a new era by introducing a paradigm of pre-training and fine-tuning. Major technical innovation of BERT is bidirectional training of Transformer model, a popular attention model to language modelling (Vaswani *et. al.*, 2017). This is in contrast to previous methods, which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. It was shown that bi-directionally trained language models superior performance by capturing deeper understanding of language context and flow than single-direction language models (Devlin *et. al.*, 2018). BERT enables effective use of unlabeled text by proposing novel pre-training tasks; masked language modeling and next sentence prediction. The tasks guide a model to learn contextualized representations of words and relationship between sentences. Models are first pre-trained on a massive amount of unlabeled data to learn the general rules and relationships; and then fine-tuned on task-specific labeled data to learn to perform specific classification tasks.

Based on promising results in NLP research, we hypothesized that pre-trained transformer-based neural network model offer a promising, and yet not fully explored, deep learning approach for a variety of sequence prediction tasks in the analysis of non-coding DNA. To investigate this, we recently developed a pre-trained bidirectional encoder representation, named DNABERT, for global interpretation of genomic sequences based on up and downstream nucleotide contexts (Ji *et. al.*, 2021). DNABERT out-performed most widely used programs for genome-wide regulatory elements prediction in accuracy and efficiency. Single pre-trained DNABERT model could simultaneously achieve state-of-the-art performance on prediction of promoters, splice sites, and transcription factor binding sites, after easy fine-tuning using small task-specific labelled data. Further, DNABERT enabled direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence motifs and functional genetic variant candidates. The pre-trained DNABERT with human genome could also be readily applied to other organisms with exceptional performance.

4. Conclusions

Most of the traditional bioinformatics tools develop the understanding of DNA from scratch with task-specific data. As the deep learning models become gradually deeper and wider, their demand for data is getting much more intense. Thus, simply relying on labeled data is very likely to result in poor performance when dataset size is small. In contrast, BERT-style *pre-train—fine-tune* scheme ingeniously utilizes the massive amount of unlabeled data to gain such understanding without the need for any human guidance, while such understanding it obtained is easily transferable to various downstream tasks. Therefore, the model can still achieve exceptional performance in data-scarce scenarios. Second, in comparison to CNN architecture, which only captures local context, Transformer globally capture contextual information from the entire input sequence by taking all the representations from the last layer as input and performing self-attention on them. With the self-attention mechanism, Transformer is not only straightforwardly parallelizable, but also effectively overcomes the gradient vanishing problem that RNN-based architectures usually meet. Therefore, BERT style modeling is expected to lead to many biological breakthroughs through general understanding of DNA by correctly capturing the hidden syntax

In the design of the above mentioned and other successful computational prediction programs, combining state-of-the-art statistical pattern recognition methods with computational systems design is essential. Additionally, expertise in data-curation and computer programing is key in the development of successful algorithms and bioinformatics software for solving these and several other challenging problems in mammalian genomics.

Acknowledgements

This work was supported by the National Library of Medicine of the NIH [R01LM011297 to RD].

References

Alipanahi, B., Delong, A., Weirauch, M. T. and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**, 831-838.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:14090473*.

Bengio, Y., Courville, A. and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828.

Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont, California.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5-32.

Brendel, V. and Busse, H. G. (1984). Genome structure described by formal languages. *Nucleic Acids Research*, **12**, 2561-2568.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P. and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, **28**, 171-182.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78-94.

Cheng, A. S., Jin, V. X., Fan, M., Smith, L. T., Liyanarachchi, S., Yan, P. S., Leu, Y. W., Chan, M. W., Plass, C., Nephew, K. P. et al. (2006). Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Molecular Cell*, **21**, 393-404.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:14061078*.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pp. 577-585.

Cutler, A. and Stevens, J. R. (2006). Random forests for microarrays. *Methods Enzymology*, **411**, 422-432.

Davuluri, R. V. (2003). Application of FirstEF to find promoters and first exons in the human genome. In *Current Protocols in Bioinformatics*, Vol **4.7** (ed. A Baxevanis). John Wiley & Sons, New York, NY.

Davuluri, R. V., Grosse, I. and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics*, **29**, 412-417.

Davuluri, R. V. and Zhang, M. Q. (2003). Computer software to find genes in plant genomic DNA. *Methods in Molecular Biology*, **236**, 87-108.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*.

Dunham, I. Kundaje, A. Aldred, S. F. Collins, P. J. Davis, C. Doyle, F. Epstein, C. B. Frietze, S. Harrow, J. Kaul, R. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

Ethayathulla, A. S., Nguyen, H. T. and Viadiu, H. (2013). Crystal structures of the DNA-binding domain tetramer of the p53 tumor suppressor family member p73 bound to different full-site response elements. *Journal of Biological Chemistry*, **288**, 4744-4754.

Gupta, R., Wikramasinghe, P., Bhattacharyya, A., Perez, F. A., Pal, S. and Davuluri, R. V. (2010). Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* **11 Suppl 1**: S65.

Hassanzadeh, H. R. and Wang, M. D. (2016). DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 178-183. IEEE.

Head, T. (1987). Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bulletin of Mathematical Biology*, **49**, 737-759.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computing*, **9**, 1735-1780.

Hong, Z., Zeng, X., Wei, L. and Liu, X. (2020). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*, **36**, 1037-1043.

Ji, S. (1999). The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Sciences-Paper Edition*, **870**, 411-417.

Ji, Y., Zhou, Z., Liu, H. and Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* doi:10.1093/bioinformatics/btab083.

Kearns, S., Lurz, R., Orlova, E. V. and Okorokov, A. L. (2016). Two p53 tetramers bind one consensus DNA response element. *Nucleic Acids Research*, **44**, 6185-6199.

Kelley, D. R., Snoek, J. and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, **26**, 990-999.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol*, **4**, 134-142.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, **521**, 436-444.

Lokshin, M., Li, Y., Gaiddon, C. and Prives, C. (2007). p53 and p73 display common and distinct requirements for sequence specific binding to DNA. *Nucleic Acids Research*, **35**, 340-352.

Lunetta, K. L., Hayward, L. B., Segal, J. and Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, **5**, 32.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M. and Stanley, H. E. (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters*, **73**, 3169-3172.

Mao, W., Kostka, D. and Chikina, M. (2017). Modeling enhancer-promoter interactions with attention-based neural networks. *bioRxiv*: 219667.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111-3119.

Ng, P. (2017). dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:170106279*.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. and O'Neal, C. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, **53**, 1161-1168.

Osada, M., Park, H. L., Nagakawa, Y., Yamashita, K., Fomenkov, A., Kim, M. S., Wu, G., Nomoto, S., Trink, B. and Sidransky, D. (2005). Differential recognition of response elements determines target gene specificity for p53 and p63. *Molecular Cell Biology*, **25**, 6077-6089.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E. and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028-2036.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:180205365*.

Portin, P. (2014). The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *Journal of Genetics*, **93**, 293-302.

Qin, H., Chan, M. W., Liyanarachchi, S., Balch, C., Potter, D., Souriraj, I. J., Cheng, A. S., Agosto-Perez, F. J., Nikonova, E. V., Yan, P. S. et al. (2009). An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. *BMC System Biology*, **3**, 73.

Quang, D. and Xie, X. H. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, **44**.

Schavolt, K. L. and Pietenpol, J. A. (2007). p53 and Delta Np63 alpha differentially bind and regulate target genes involved in cell cycle arrest, DNA repair and apoptosis. *Oncogene*, **26**, 6125-6132.

Searls, D. B. (1992). The linguistics of DNA. *American Scientist*, **80**, 579-591.

Searls, D. B. (2002). The language of genes. *Nature*, **420**, 211-217.

Shen, Z., Bao, W. and Huang, D.-S. (2018). Recurrent neural network for predicting transcription factor binding sites. *Scientific Reports*, **8**, 1-10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998-6008.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048-2057.

Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994). Recognizing exons in genomic sequence using GRAIL II. *Genetic Engineering*, *(N Y)* **16**, 241-253.

Yan, D. F. and Guo, S. Y. (2019). Leveraging Contextual Sentences for Text Classification by Using a Neural Attention Model. *Computational Intelligence and Neuroscience*, **2019**.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480-1489.

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Review Genetics*, **3**, 698-709.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, **12**, 931-934.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A. and Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, **51**, 12-18.

Special Proceedings: {ISBN #: 978-81-950383-0-5}
23rd Annual Conference, 24–28 February 2021; pp 185–199

Inferring the "Laws" of Finance from High-frequency Data

R K Singh[1,2,3] **and Sitabhra Sinha**[1,2]
[1] *The Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600113, India*
[2] *Homi Bhaba National Institute, Anushaktinagar, Mumbai 400 094, India*
[3] *Department of Physics, Bar-Ilan University, Ramat-Gan 5290002, Israel*

Received: 16 August 2021; Revised: 31 August 2021; Accepted: 04 September 2021

Abstract

Financial markets represent a prototypical example of complex systems in which a large number of heterogeneous agents are involved in mutual interactions (namely, trading assets with each other) each having a specific goal (maximizing their own profit). However, despite the unpredictability inherent in the individual components, the system as a whole can exhibit universal features that are invariant across different markets, asset classes and period of observation. One of the most prominent examples of such "stylized facts" is the appearance of fat tails in distributions of price fluctuations, often referred to as the *inverse-cubic law*. However, as such features have mostly been reported in studies using low-resolution data, we look for them using high-frequency data of equities trading in the National Stock Exchange of India, one of the world's leading financial markets. We find that the distribution of trade sizes (the number of stock units involved in a single transaction) possess heavy tails, decaying as a power law with a characteristic exponent. Moreover, the distribution is in general stationary for the market as a whole, even though those of individual stocks may differ significantly from one period to the next. We also investigate the distribution of waiting times between successive trades and find it to be decaying slower than exponential in the case of individual stocks. We relate this to the frequent occurrences in succession of transactions involving large returns (price changes). The correlation between the intervals separating successive trades and the magnitude of price fluctuations that we observe implies that the distributions of the waiting times and that of step lengths in the walk executed by the price of a financial asset may not be completely independent. We also find that the cumulative volatility of price movements increases linearly with time within a trading day, but with deviation from linearity at the ends. This suggests that the non-Gaussian character of the return distribution (reflected in the inverse cubic law) arises from the significant volume of end-of-day trading.

Key words: Power laws; Quantitative finance; Price returns; Inverse cubic law; Heavy-tailed distributions; Trade-size distribution.

AMS Subject Classifications: 91B80, 91G15, 62P05
Journal of Economic Literature Classification: C40, G10

Corresponding Author: Sitabhra Sinha
Email: sitabhra@imsc.res.in

1. Introduction

Financial markets are one of the best known examples of complex systems which are characterized by a large number of interacting components and exhibiting nonlinear dynamics that is inherently unpredictable (Sinha *et al.* (2010)). Even though deterministic descriptions of the time evolution of individual components may not lead easily to an understanding of how an assembly of such components will behave, paradoxically the interactions between many constituents may make it possible to obtain well-defined statistical properties of the system as a whole, i.e., the market (Sinha *et al.* (2016)). Indeed, robust statistical features have been reported for the trading dynamics of different stock markets, notable among them being the so-called "inverse cubic law" describing the nature of the distributions of fluctuations in stock prices (and market indices) in developed (Lux (1996)), as well as, developing economies (Pan & Sinha (2007) and Pan & Sinha (2008)). In general, distributions having heavy tails have been reported for price fluctuations seen across many different asset classes, including currency exchange rates (Chakraborty *et al.* (2018) and Chakraborty *et al.* (2020)) and cryptocurrencies such as bitcoin (Dixit *et al.* (2015)).

Given the price p_t of some stock at time t, price fluctuations are characterized in terms of *logarithmic returns* or *log-returns*, *viz.*,

$$r_t = \ln p_t - \ln p_{t-\Delta t}, \tag{1}$$

where Δt is the time interval separating the two prices (Figure 1). The variation in the fluctuations of price of a stock is defined in terms of *volatility*, defined as the variance of log-returns, *viz.*,

$$\sigma_t^2 = \langle r_t^2 \rangle - \langle r_t \rangle^2. \tag{2}$$

If q_i denotes the number of stocks traded in the i-th transaction (also referred to as the *trade size*) then the volume of stocks $V_{\Delta t}$ traded over a time interval $[t, t + \Delta t]$ is defined as

$$V_{t,\Delta t} = \sum_{i=1}^{N_{t,\Delta t}} q_i, \tag{3}$$

where $N_{t,\Delta t}$ is the number of trades that have occurred during the interval $[t, t + \Delta t]$. The above relation implies that quantities characterizing a trading event, viz., trade sizes, trading volumes and number of trades are not independent of each other.

Most early studies of the empirical statistical properties of markets have used daily (or end of day) trade data which does not take into consideration the dynamical behavior of intra-day trading (Vijayraghavan & Sinha (2011)). In recent times, the availability of high-frequency (HF) data containing information about every transaction taking place in the market has made it possible to uncover the properties of financial markets at the highest possible resolution (Dacorogna *et al.* (2001)). In general, empirical analysis using such data can reveal features that are not possible to observe using data collected at temporal

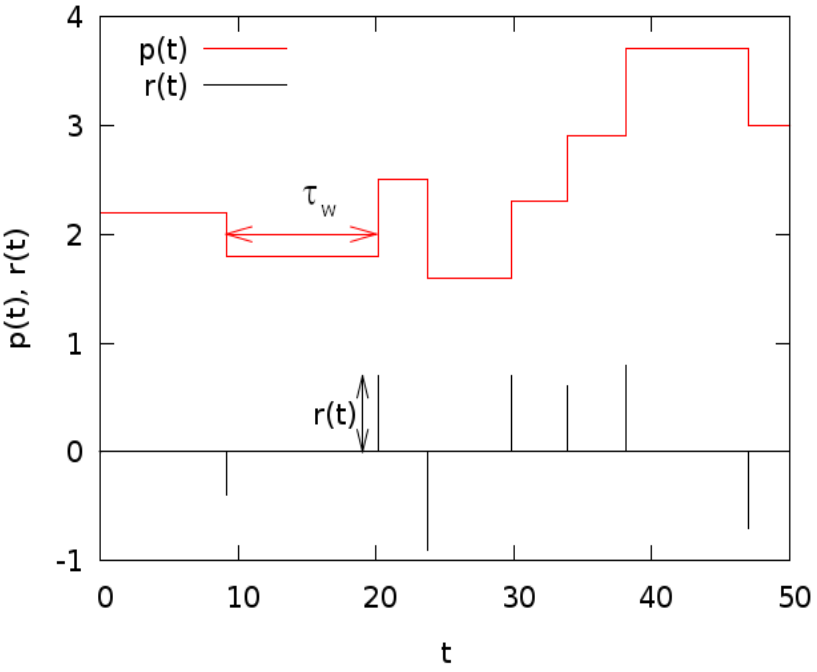


Figure 1: Schematic figure illustrating the time-evolution of price $p(t)$ and its fluctuation, measured as log-return $r(t)$, of an asset. Price changes every time a trade occurs, the time between two successive trades being denoted as τ_w.

resolutions of a single day. For example, the transactions in a stock market do not occur at regularly spaced intervals, implying that the events defining transactions of an asset could themselves be a stochastic process with a given distribution of *waiting times* τ_w, which are defined as the duration between two successive transactions. Motivated by this, we have used the HF equities trading data of the National Stock Exchange (NSE) of India (NSE (2020)) to uncover its principal statistical features. We focus on properties of the market as a whole, as well as, that of individual stocks, *e.g.*, the distribution of trade sizes, the distribution of waiting times between two successive trades, as well as the relation between price fluctuations and waiting times. Our study has implications for the current understanding of the dynamics of a developing market, in particular, the intra-day behavior of trades and price movements, as well as, in modeling such behavior using the tools of statistical physics. Our paper is organized as follows. In Section 2, we briefly describe the data used in our analysis. The results are described in Section 3, with successive subsections dealing with distributions of trade sizes, waiting times, logarithmic returns and their inter-relations. We conclude with an outline of the main findings in Section 4.

2. Description of the Data

In order to study the statistical properties of NSE, we use tick-by-tick HF data for the month of December in each year during the period 1999 to 2012 (the reason for focusing on a particular month is to enable comparison between the behavior seen in different years without the confounding factor of intra-annual seasonal variations). Note that the HF data comes with its own set of challenges, e.g., overwhelming data size, unevenly spaced time

series, etc. This is visible from a small sample of the data set for December 2003 shown below:

```
20031201|MTNL|09:56:29|122.20|10
20031201|MTNL|09:56:29|122.25|50
20031201|MTNL|09:56:29|122.30|40
20031201|SATYAMCOMP|09:56:30|335.25|1000
20031201|SAIL|09:56:30|42.70|700
20031201|M&M|09:56:30|355.95|100
20031201|SATYAMCOMP|09:56:30|335.25|500
20031201|SATYAMCOMP|09:56:30|335.25|100
20031201|RAINCALCIN|09:56:30|25.40|500
20031201|VDOCONINTL|09:56:30|78.85|70
```

where the columns separated by "|" represent respectively the date, name of the company whose equity is being traded, the time of the transaction, price per stock of the equity traded, and the number of stocks that changed hands during the transaction (i.e., the trade size q).

It can be observed from the above sample that many transactions share the same time-stamp. This is because the temporal resolution of recording the transactions is 1 second, so that if two transactions occur within a duration of less than a second of each other, they will have the same time-stamp. However, the ordering of transactions is reported in the correct time-sorted order. For the purpose of the present analysis, we assume that the transactions sharing the same time-stamp take place at the same instant. As the market is open for the trading of common stocks only between 0950 hours and 1530 hours, the results reported here are obtained by exclusively considering trades that took place between 0950 hours and 1530 hours on a given day.

In order to study the properties of individual stocks we choose four representative stocks, *viz.*, $HDFCBANK$ (Finance sector), $INFOSYS$ (Infotech sector), $RELIANCE$ (Energy sector) and $SUNPHARMA$ (Pharmaceutical sector), as these belong to a few of the most important industrial sectors in NSE in terms of market capitalization. We show in Figure 2 the variation of price $p(t)$, returns $r(t)$ and trade size $q(t)$ as a function of time for $RELIANCE$ for the first 30 minutes of trading on December 1, 2005 as a typical example of the financial time series.

3. Results

3.1. Distribution of Trade Sizes

We report the cumulative probability distribution $P(Q \geq q)$ of the trade sizes q for the entire market (i.e., aggregating over all equities traded) for the month of December for years from 1999 to 2012 in Figure 3. We see that each of the distributions (corresponding to every year between 1999-2012) when shown on a log-log graph has a substantial linear portion, implying that the tails for the distribution of trade sizes calculated over the entire market follows a power law.

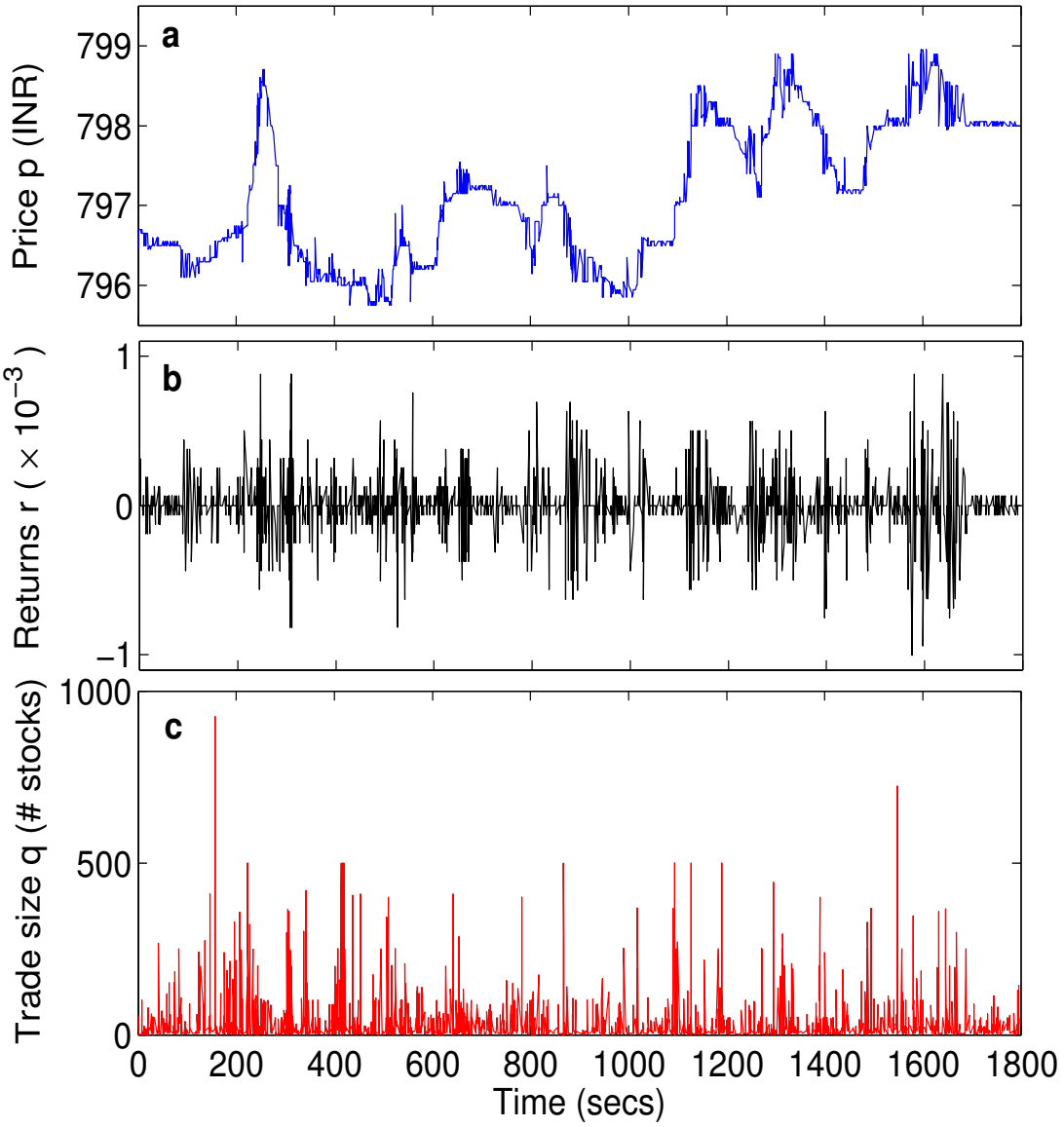


Figure 2: Representative time-series of $RELIANCE$ stock for 30 minutes beginning at 0950 hrs on December 1, 2005 showing (a) price $p(t)$ of the stock, (b) log-returns $r(t)$, and (c) trade size $q(t)$ as a function of time. Time is measured in seconds, with the origin (0) set at 0950 hrs.

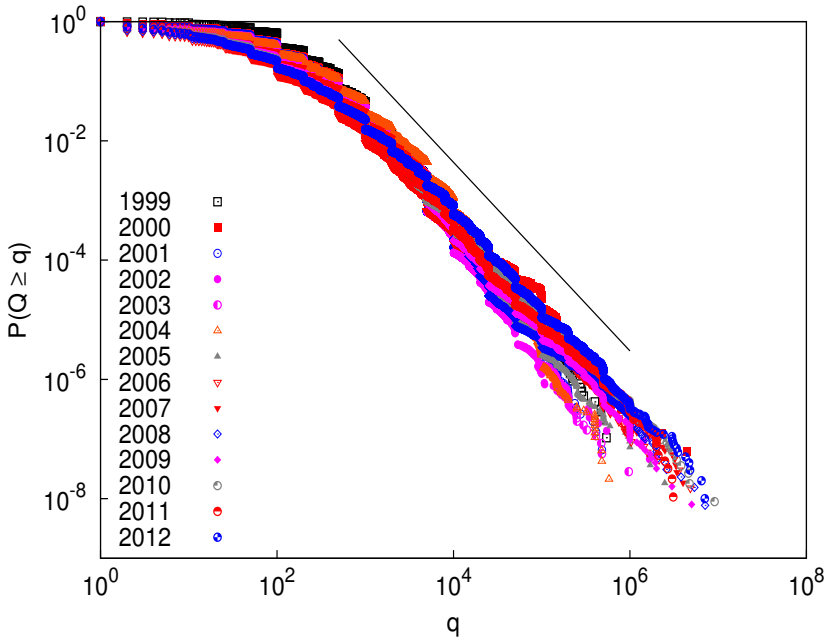


Figure 3: Cumulative probability distribution $P(Q \leq q)$ of the trade sizes q for NSE for the month of December for the years 1999 to 2012. Note the doubly logarithmic axis of the graphs, which means that a linearly decaying tail implies the existence of power-law decay of the distribution (the line represents a power-law fit with exponent 1.6).

Figure 4 shows the distribution of trade sizes q for equities of the four representative companies for the month of December in the years 2003, 2007 and 2010. It is evident from the figure that the tails of the distributions of each of the stocks also exhibit power law decay. Hence, the tails of the distribution of the trade sizes q are of the form:

$$P(Q \geq q) \sim q^{1-\alpha}, \tag{4}$$

where α is the exponent characterizing the power law distribution. The maximum likelihood estimates of the exponents (Clauset *et al.* (2009)) are shown in Figure 5. It is seen from Figure 5(a) that $\alpha \leq 3$ for almost all periods for the entire market (except 2003). This implies that for these distributions, moments other than the first do not exist. As this property holds true for almost the entire period under consideration, we can say that the distribution of trade sizes for the market is stationary, in the sense that it is Lévy stable with second and higher moments diverging. However, this is not true for the case of individual stocks, as seen from Figure 5 (b), where we see that the values taken by α for different stocks vary between 2 and 4. In addition, values taken by α for a particular stock exhibit widely different values depending on the period at which it is being observed, thus suggesting that the statistical behavior of trading dynamics for individual equities is non-stationary.

3.2. Distributions of waiting times and price fluctuations

An important quantity associated with a given stock is its price $p(t)$ at any given time t and which generally fluctuates over time. A generic time-series representing price variations of a given equity is shown in Figure 6. We see in panel (a) that the price $p(t)$ at time

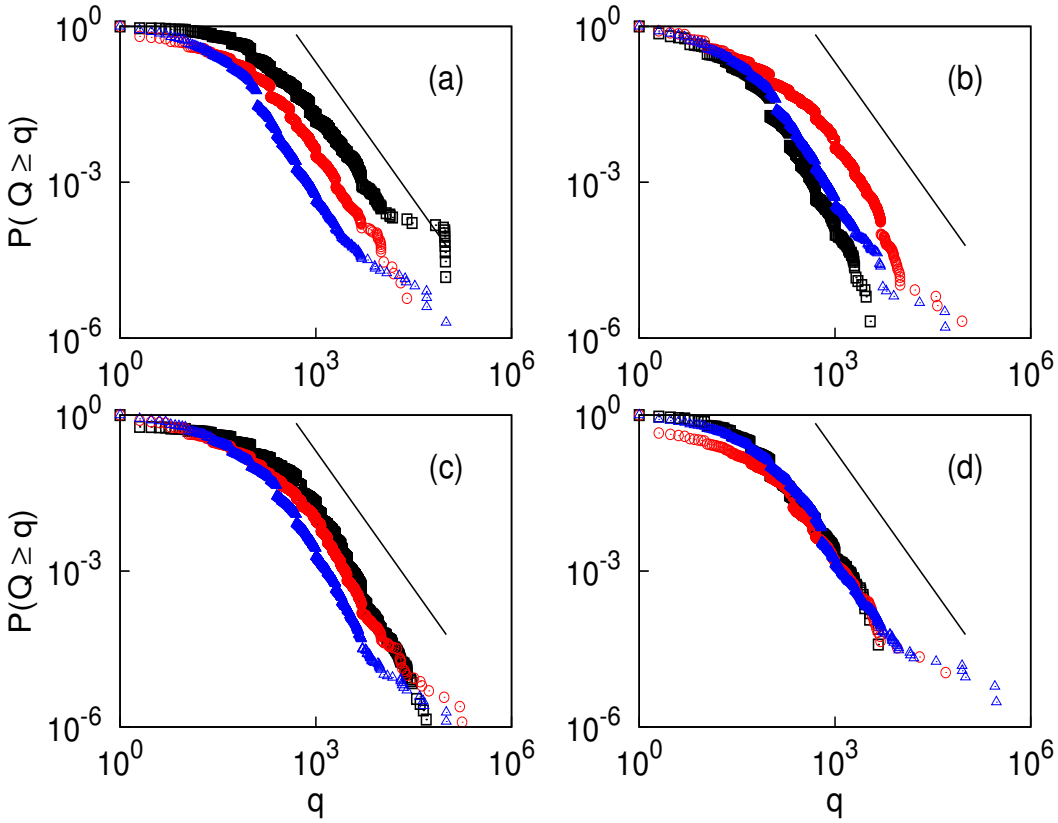


Figure 4: Cumulative probability distribution of trade sizes q for the equities of (a) $HDFCBANK$, (b) $INFOSYS$, (c) $RELIANCE$, and (d) $SUNPHARMA$. The distributions are shown for the month of December for the years 2003 (black squares), 2007 (red circles) and 2010 (blue triangles). As in the case of the entire market, the distributions of trade sizes for individual equities also exhibit a power law form.

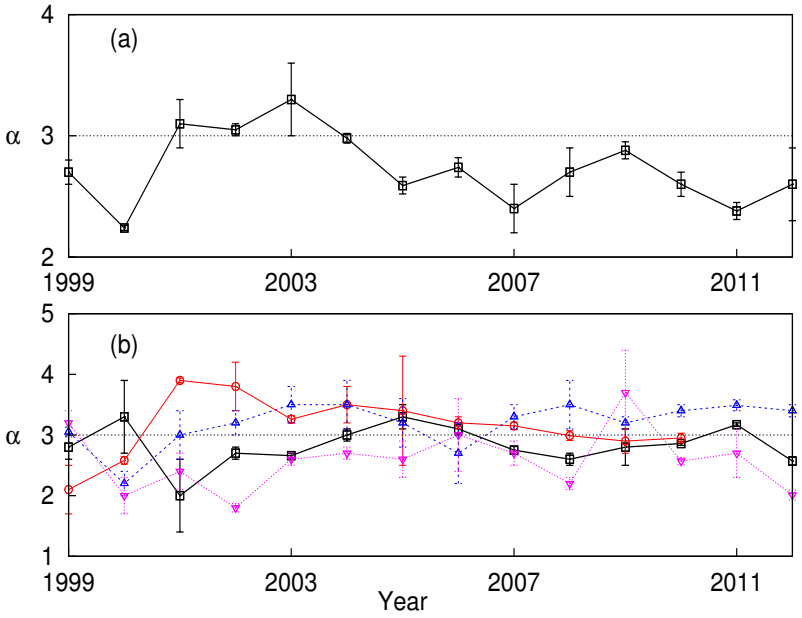


Figure 5: Maximum likelihood estimates of the exponents characterizing the power law nature of the tail for the distribution of trade sizes for (a) the entire market and (b) the equities of *HDFCBANK* (black squares), *INFOSYS* (red circles), *RELIANCE* (blue triangles), and *SUNPHARMA* (maroon inverted triangles), for the month of December for all years between 1999 to 2012. Note that the data for *INFOSYS* was not available for the years 2011 and 2012. Bars represent the error in estimating the exponents, obtained using bootstrap technique. The horizontal broken line indicates $\alpha = 3$ which demarcates distributions with Levy-stable nature from those that will eventually converge to a Gaussian.

t and at time $t + \tau_w$ can be same or different, where τ_w is the waiting time between the two transactions. The tick-by-tick log-return associated with the successive price changes is measured as

$$r(t) = \ln p(t) - \ln p(t - \tau_w) \tag{5}$$

and is shown in Figure 6 (b). The random walk nature of price changes is easily seen from the figure, and the presence of irregular waiting times τ_w makes it an effectively continuous time process. Characterizing the distribution of waiting times is of fundamental importance in order to understand the price dynamics of a given asset. For that purpose, we show the distribution of waiting times for *RELIANCE* for the month of December 2005 in Figure 7.

It can be seen that the distribution of waiting times $P(\tau > \tau_w)$ cannot be fit by a single exponential distribution having a characteristic period $\langle \tau_w \rangle$. This implies that the system has inherent long-range memory and that the occurrence of successive transactions are not independent events. This becomes clear upon fitting the empirical distribution of waiting times with a theoretical curve having the form of a sum of exponentials, *viz.*,

$$P(\tau \geq \tau_w) = \sum_i a_i \exp(-(\tau_w/b_i)). \tag{6}$$

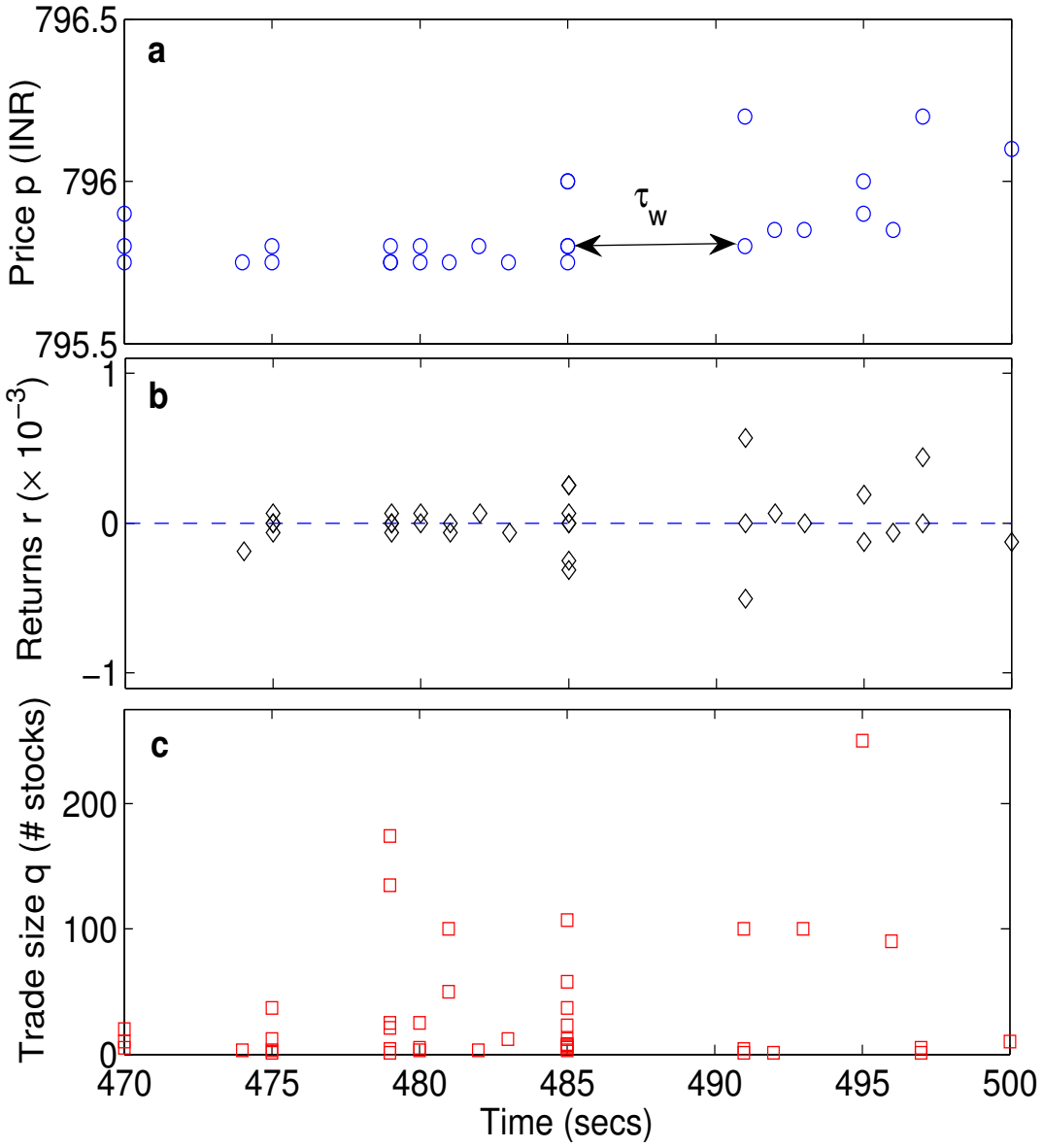


Figure 6: A magnified view of the time series shown in Figure 2 indicating that several transactions can share the same time stamp because the temporal resolution of the recording is limited to 1 second. The waiting time τ_w between two successive trades specifies an interval during which no transaction takes place.

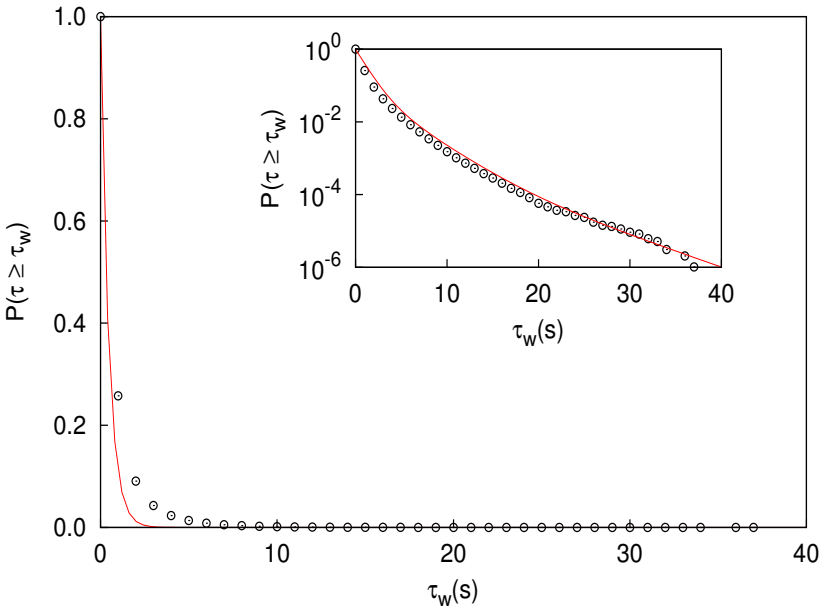


Figure 7: Cumulative probability distribution of waiting times for successive trades in *RELIANCE* stock for the month of December in 2005. The empirically obtained distribution (circles) has been fitted with an exponential distribution (solid curve) having the same mean waiting time $\langle \tau_w \rangle \approx 0.453$. The inset shows a fit with a theoretical distribution that is a sum of three exponentially decaying curves $\sum_i a_i \exp(-(\tau_w/b_i))$ having different characteristic times τ_w/b_i $(i = 1, 2, 3)$.

Such theoretical distributions have been used earlier to describe systems having memory (Goychuk (2009)). The values of the coefficients used to fit the data shown in Figure 7 are $a_1 = 0.897, a_2 = 0.1, a_3 = 0.003; b_1 = 1, b_2 = 0.4, b_3 = 0.2$.

We also see from Figure 8 that the distribution of returns $Pr(X \geq x)$ has power law decaying tails with exponent close to 3. This is in accordance with the well-known inverse cubic law of price fluctuations reported for financial markets (Lux (1996), Gopikrishnan *et al.* (1998) and Pan & Sinha (2007)).

3.3. Relation between price fluctuations and waiting times

The relation between price fluctuations and waiting times can be discerned from Figure 9 which shows a scatter plot of log-returns r measured for successive transactions against the corresponding time-interval τ_w between them for a particular equity. It can be observed from the diagram that transactions resulting in large price changes generally occur closer to each other. To further quantify this observed behavior we look at the distribution of the log-returns conditioned on waiting times, i.e., $P(r|\tau_w)$ in Figure 10. It is evident from the distribution $P(r|\tau_w)$ that larger returns generally occur close to each other in time (Scalas (2006)), thus suggesting that the waiting-times and returns may not be independent of each other. This property is also observed for different stocks and for different years. This has implications towards the modeling of price dynamics by continuous time random walks, as market transactions may be better modeled by walks whose step lengths are not chosen

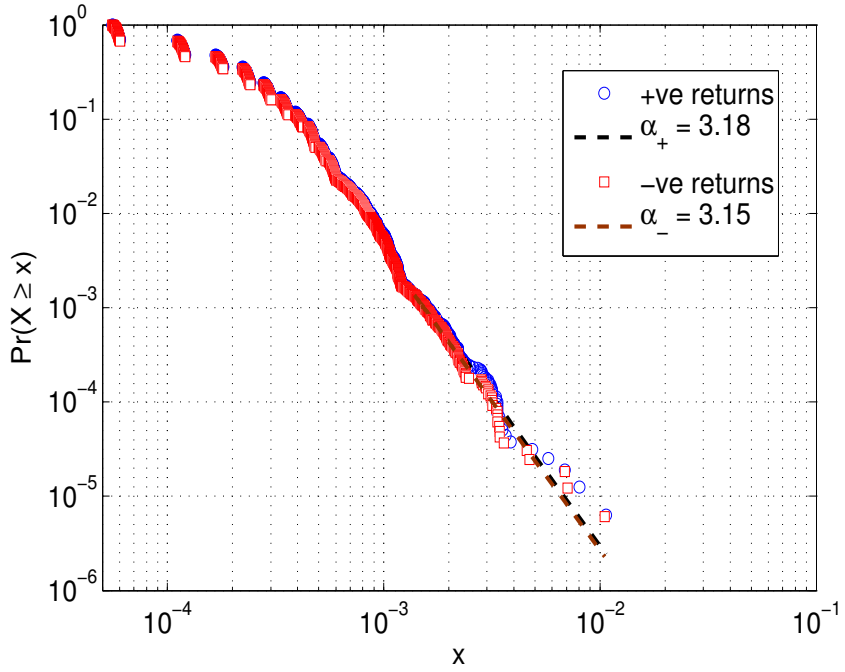


Figure 8: Distribution of log-returns of $RELIANCE$ (each return being measured over an interval of 1 tick) for the month of December 2005.

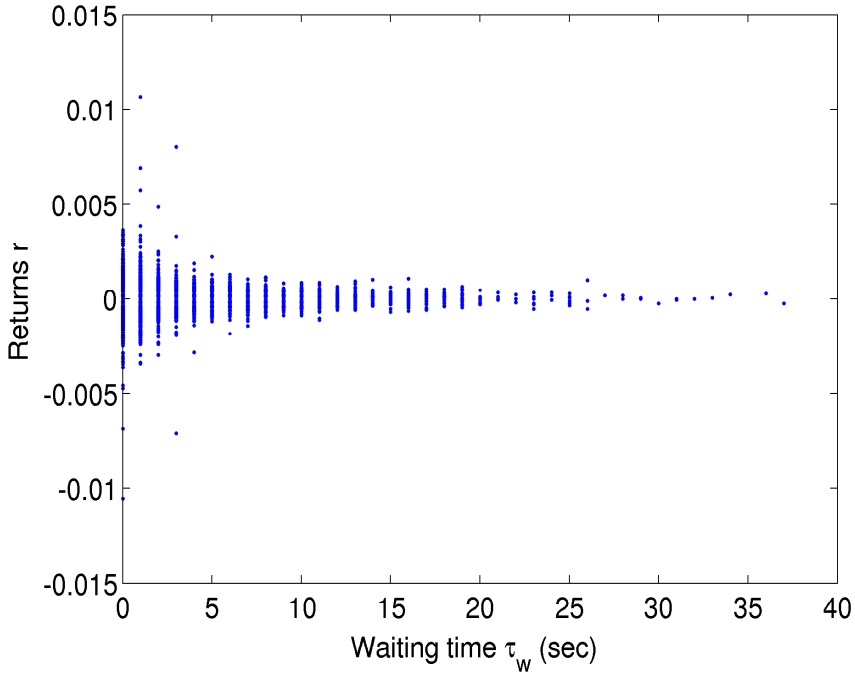


Figure 9: Scatter plot of waiting times τ_w between successive trades and the corresponding log-return r of $RELIANCE$ shares for all transactions that took place in December 2005.

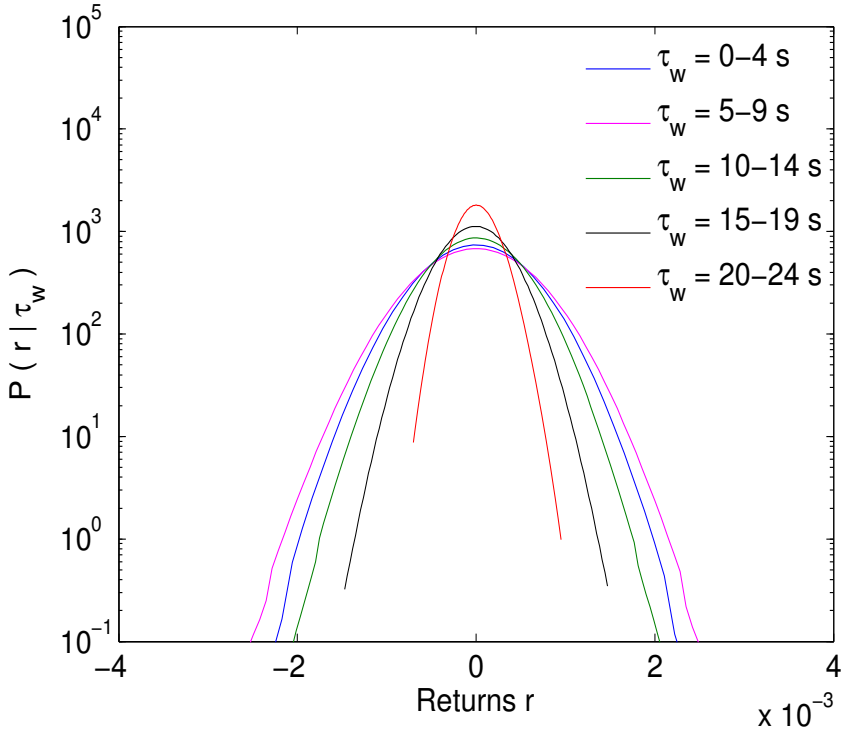


Figure 10: Conditional distribution $P(r|\tau_w)$ of log-returns $r(t)$ conditioned on the waiting times τ_w for successive trades in equities of $RELIANCE$ that were carried out in December 2005.

independently of the waiting time between successive steps, contrary to what is generally assumed (Merton (1976) and Masoliver *et al.* (2000)).

In order to characterize the dynamics of intra-day trading we define the variance of log-returns over an interval Δt as:

$$\sigma^2(t) = \frac{1}{N_{\Delta t}(t) - 1} \sum_i (r_i - \langle r_i \rangle)^2, \tag{7}$$

where $N_{\Delta t}(t)$ is the number of trades occurring in the interval of length Δt and $\langle r_i \rangle$ is the average return over the interval. We observe from Figure 11 that the variance over the intervals is nearly independent of the length of the interval Δt, and find that the scaled cumulative variance $\sigma_c^2(t)\Delta t$ is independent of Δt, as shown in Figure 12. In addition, we also find that the cumulative variance grows linearly with time, thus implying that for the major part of the day, price fluctuations of individual stocks are inherently Gaussian. We can see, however, a perceptible deviation from linearity at the beginning of the trading day, and more prominently towards its end. This suggests that the non-Gaussian nature of the return distribution, as evident from the heavy tails characterized by the inverse cubic law, is possibly an outcome of the significant volume of transactions that occur at either end of a trading day.

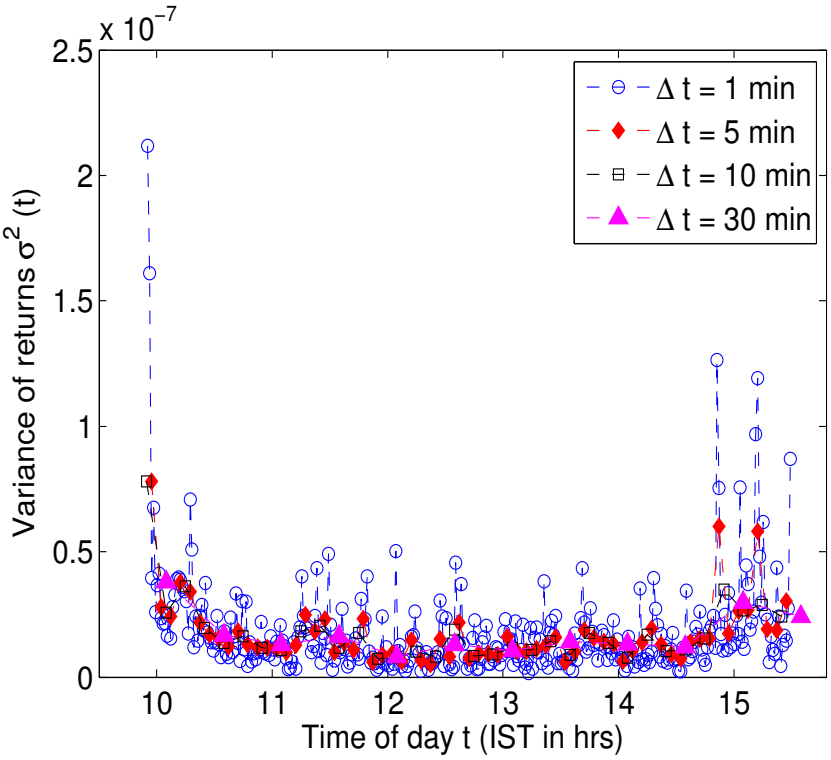


Figure 11: Intra-day volatility $\sigma^2(t)$ (measured by variance of the log-returns) during a period Δt shown as a function of time of day t for $RELIANCE$ on December 1, 2005, for different intervals Δt.

4. Conclusions

In this paper we investigate the statistical behavior of financial markets which constitute prototypical examples of complex systems having large number of components with unpredictable dynamics. Despite such unpredictability, statistically regular properties for the entire system can be observed as has indeed been reported for many different observables associated with market dynamics. Notable among such invariant features are the power laws describing the tails in the distributions of asset price fluctuations, as well as, the distributions of trade sizes. However, most early studies reporting such features have used low-resolution daily data, and thus do not take into account the information about intra-day trading. Motivated by this, in this paper we have reported our preliminary analysis of the high-frequency equities trading data obtained from the National Stock Exchange of India. Such data provides information about market movements at the highest possible resolution thereby revealing vital clues for understanding the underlying dynamics of this complex system.

Using data for the month of December for all years between 1999-2012 we show that gross statistical properties of the market as a whole are in general stationary, even though those of its constituents, i.e., equities of individual companies, may not be. In particular, we see that the distribution of trade sizes aggregated over all equities does not change its nature over time, with the exponent characterizing the power law nature of the tails taking values from the interval $(2, 3)$, i.e., it is Lévy stable with second and higher moments diverging.

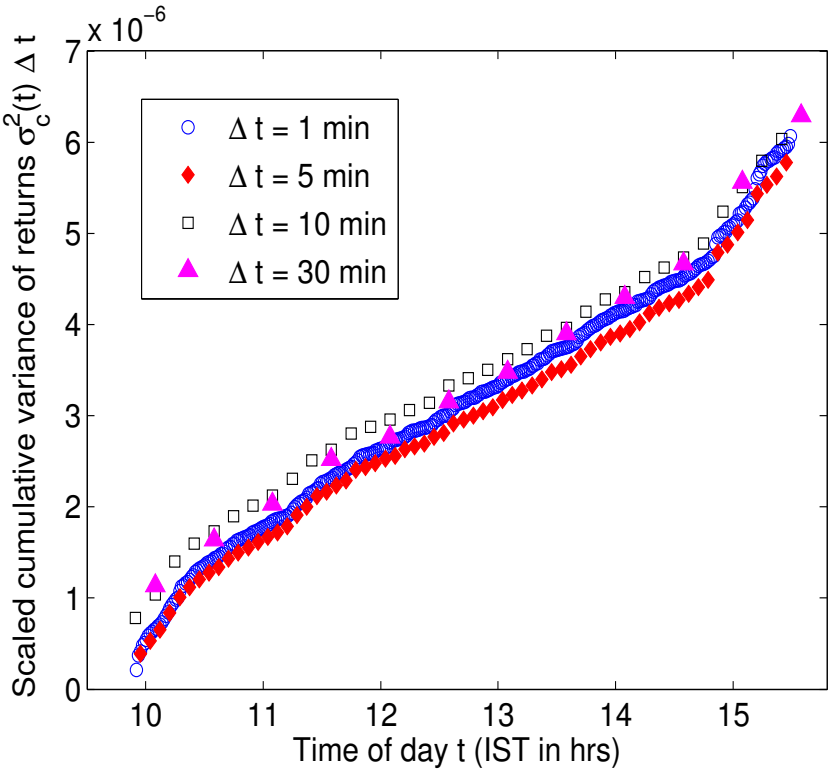


Figure 12: Scaled cumulative variance of the log-returns for intra-day trading $\sigma_c^2(t)\Delta t$ as a function of time of day t for $RELIANCE$ on December 1, 2005. It is observed that scaling by interval Δt over which the variance is calculated, the curves for cumulative variance $\sigma_c^2(t)$ for different choices of Δt overlap.

However, for individual stocks, this distribution can differ significantly between one period and another. We also find that waiting-times of transactions of individual stocks exhibit non-exponential character, and are related to log-returns in that transactions involving larger returns occur close to each other. This implies that the distribution of waiting times and step lengths of the random walk executed by a financial asset are not independent of each other, as has been assumed in many studies. We also find that cumulative volatility of returns increases linearly with time within a day. This implies that for a major part of the day, price fluctuations are Gaussian in nature. However, a significant deviation from linearity is seen towards the ends, suggesting that the genesis of the heavy-tailed nature of return distributions (reflected in the inverse cubic law) lies in the significant volume of trade that occurs at the beginning and towards the end of a trading day.

Acknowledgements

We thank Frederic Abergel, Trilochan Bagarti, Abhijit Chakraborty and Soumya Easwaran for helpful discussions. The work was supported in part by the Center of Excellence in Complex Systems and Data Science, The Institute of Mathematical Sciences, funded by the Department of Atomic Energy, Government of India.

References

Chakraborty, A., Easwaran, S. and Sinha, S. (2018). Deviations from universality in the fluctuation behavior of a heterogeneous complex system reveal intrinsic properties of components: The case of the international currency market. *Physica A*, **509**, 599–610.

Chakraborty, A., Easwaran, S. and Sinha, S. (2020). Uncovering hierarchical structure of international FOREX market by using similarity metric between fluctuation distributions of currencies. *Acta Physica Polonica A*, **138(1)**, 105–115.

Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, **51(4)** 661–703.

Dacorogna, M. M., Gençay, R., Müller, U. A., Olsen, R. B. and Pictet, O. V. (2001). *An Introduction to High-frequency Finance.* Academic Press, San Diego.

Easwaran, S., Dixit, M. and Sinha, S. (2015). Bitcoin dynamics: The inverse square law of price fluctuations and other stylized facts. In *Econophysics and Data Driven Modelling of Market Dynamics*, Springer, Cham, 121–128.

Gopikrishnan, P., Meyer, M., Amaral, L. A. N. and Stanley, H. E. (1998). Inverse cubic law for the distribution of stock price variations. *European Physical Journal B*, **3(2)**, 139–140.

Goychuk, I. (2009) Viscoelastic subdiffusion: From anomalous to normal. *Physical Review E*, **80(4)**, 046125.

Lux, T. (1996). The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. *Applied Financial Economics*, **6(6)**, 463–475.

Masoliver, J., Montero, M., Perelló, J. and Weiss, G. H. (2000). The continuous time random walk formalism in financial markets. *Journal of Economic Behavior & Organization*, **61(4)**, 577–598.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, **3(1-2)**, 125–144.

NSE (2020). *NSE India: About NSE Data & Analytics.* https://www1.nseindia.com/supra_global/content/dotex/about_dotex.htm.

Pan, R. K. and Sinha, S. (2007). Self-organization of price fluctuation distribution in evolving markets. *Europhysics Letters*, **77(5)**, 58004.

Pan, R. K. and Sinha, S. (2008). Inverse-cubic law of index fluctuation distribution in Indian markets. *Physica A*, **387(8)**, 2055–2065.

Scalas, E. (2006). The application of continuous-time random walks in finance and economics. *Physica A*, **362(2)**, 225–239.

Sinha, S., Chakrabarti, A. S. and Mitra, M. (2016). What is economics that physicists are mindful of it? *The European Physical Journal Special Topics*, **225(17-18)**, 3087–3089.

Sinha, S., Chatterjee, A., Chakraborti, A. and Chakrabarti, B. K. (2010). *Econophysics: An Introduction.* Wiley-VCH, Weinheim.

Vijayraghavan, V. S. and Sinha, S. (2011). Are the trading volume and the number of trades distributions universal ? In *Econophysics of Order-Driven Markets*, Springer, Milan, 17–30.

Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 201-214

Unreasonable Charges for Transacting Digitally in a BSBDA– Extent of SBI's Undue Enrichments through UPI and RuPay Charges

Ashish Das
Department of Mathematics, Indian Institute of Technology Bombay, Mumbai 400076, India

Received: 13 July 2021; Revised: 22 September 2021; Accepted: 25 September 2021

Abstract

Two major milestones under the financial inclusion drive of the country were the institution of the Basic Savings Bank Deposit Account (BSBDA) by RBI in 2012-13 and introduction of the Pradhan Mantri Jan Dhan Yojana (PMJDY) by the government in August 2014. The PMJDY facilitated opening of BSBDAs by unbanked households based on the guiding principles of banking the unbanked and securing the unsecured.

There had been a systematic breach in the RBI regulations when State Bank of India (SBI) imposed exorbitant and unreasonable charges onto BSBDA customers who transacted digitally. Since June 1, 2017, SBI charged @ Rs. 17.70 for every debit transaction beyond four a month. This has adversely impacted the BSBDA customers of SBI who, on the call of the government and RBI, embraced digital means of financial transactions. These relatively vulnerable, gullible and marginalized fellow-countrymen being thrust with charges @ Rs. 17.70, for transacting digitally, is not only unreasonable but unjust too.

We analyse the extent of SBI's undue enrichments through imposition of transaction charges on digital payment means, (i) the UPI and (ii) the RuPay debit card. During the period January-September 2020, SBI attributed to about 222 crore UPI transactions and about 6.8 crore RuPay (POS and eCom) debit card transactions. Of these UPI and RuPay debit card transactions, only 2.23% of the transactions, i.e., 5.1 crore transactions were charged by SBI @ Rs. 17.70 per transaction, exclusively from the BSBDA customers. In this process, SBI collected Rs. 90.2 crore, which was subsequently refunded in February-March 2021. Nevertheless, SBI made undue interest gains of Rs. 4.65 crore, which actually belongs to the BSBDA customers, who were charged for these UPI and RuPay debit card transactions. This undue enrichment of SBI, at the cost of the depositors, has not been assessed or audited yet.

More seriously, for the prior 33 months, April 2017 through December 2019, SBI has collected over Rs. 164 crore towards charges imposed on atleast 9 crore UPI and RuPay debit card transactions. Again, SBI has still not refunded this amount to the BSBDA customers.

The Board of Directors of SBI have been vested with the responsibility to ensure that charging Rs. 17.70 for every UPI/ RuPay debit card transaction is reasonable as per principles laid down by RBI. Given that it costs a bank disproportionately more to provide an ATM cash withdrawal service than to provide the UPI/ RuPay debit card transaction facility, SBI's imposition of a uniform charge of Rs. 17.70 for both the ATM cash withdrawal and the UPI/ RuPay debit card transaction is grossly unreasonable and in breach of RBI regulations. While having embraced digital means for transacting, the BSBDA customers remained an unprotected lot since SBI's actions amounted to exploitation of this marginalized section of the society through imposition of usurious service charges.

*Key word*s: Savings Account; BSBDA; PMJDY; Board of Directors; UPI; RuPay debit card.

Corresponding Author: Ashish Das
Email: ashish@math.iitb.ac.in

1. Introduction

Two major milestones under the financial inclusion drive of the country were the institution of the Basic Savings Bank Deposit Account (BSBDA) by RBI in 2012-13 and introduction of the Pradhan Mantri Jan Dhan Yojana (PMJDY) by the government in August 2014. The PMJDY facilitated opening of BSBDAs by unbanked households based on the guiding principles of banking the unbanked and securing the unsecured.

Data from RBI's Annual Reports and DFS's weekly PMJDY data shows that upon introduction of the PMJDY, there had been a significant growth of BSBDAs in the country. As of December 2020, 64.9 crore BSBDAs have been opened through branch and Business Correspondent (BC) points, of which, nearly two-thirds (41.6 crore) have been opened under the PMJDY. From the inception of the PMJDY, the thrust has been to open only BSBDAs under the yojana. Though general public may be less familiar with the nomenclature of BSBDAs, of the 173 crore savings bank accounts opened (as of March 2020), 60 crore were BSBDAs. Thus, more than a third of the savings bank accounts are BSBDAs.

1.1. BSBDA - the backdrop

RBI introduced the BSBDA in August 2012. Banks were advised to offer a 'Basic Savings Bank Deposit Account', which will offer the following minimum common facilities:

• The account shall not have the requirement of any minimum balance;
• While there will be no limit on the number of deposits that can be made in a month, account holders will be '**allowed a maximum of four withdrawals**' in a month, including ATM withdrawals; and
• The account shall provide the facility of ATM card or ATM-cum-Debit Card.

These facilities were required to be provided in a BSBDA without any charges.

Furthermore, RBI in their 2012 definition of BSBDA indicated that banks would be free to evolve other requirements including pricing structure for additional value-added services beyond the stipulated basic minimum services on **reasonable** and transparent basis and applied in a non-discriminatory manner.

As per the mandate, minimum common facilities include "account holders will be allowed **a maximum of** four withdrawals in a month". Therefore, at least in this mandate, the question of "Banks are free to levy reasonable charges in BSBDAs beyond 4 free transactions" does not arise because of the explicit and non-superfluous words "**a maximum of**" in the mandate "allowed a maximum of four withdrawals".

Nevertheless, after a year of the introduction of BSBDA, in September 2013, RBI issued detailed guidelines (clarification-circular) on how to interpret the August 2012 circular on BSBDA. While defining the features of a BSBDA, the September 2013 circular or the July 2015 Master Circular explains the characteristic features of a BSBDA unequivocally. A salient refinement in the definition of BSBDA 'allowed more than four withdrawals' in a month, **at the bank's discretion**, provided the bank does not charge for the same. RBI is clear to mention that in a BSBDA, banks cannot charge, and if they do, the account is not a BSBDA.

Thus, the regulatory requirements made it amply clear that in addition to mandatory free banking services (that included four withdrawals per month), so long as the savings deposit account is a BSBDA, banks cannot impose any charge even for value-added banking services that a bank may like to offer at their discretion.

Select FAQs of September 2013 Circular – Highlighting features of a BSBDA and **Why Banks cannot charge in a BSBDA so long as the account is a BSBDA**

Query-11: What kinds of services are available free in the 'Basic Savings Bank Deposit Account'?
Response: The services available free in the 'Basic Savings Bank Deposit Account' will **include deposit and withdrawal of cash**; receipt / credit of money through electronic payment channels or by means of deposit / collection of cheques at bank branches as well as ATMs.
Query-13: Whether banks are free to offer more facilities than those prescribed for 'Basic Savings Bank Deposit Account'?
Response: Yes. However, **the decision to allow services beyond the minimum prescribed has been left to the discretion of the banks** who can either offer additional services free of charge or evolve requirements including pricing structure for additional value-added services **on a reasonable** and transparent **basis** to be applied in a non-discriminatory manner with prior intimation to the customers. Banks are required to put in place a **reasonable pricing structure for value added services or prescribe minimum balance requirements** which should be displayed prominently and also informed to the customers at the time of account opening. Offering such additional facilities should be non-discretionary, non-discriminatory and transparent to all 'Basic Savings Bank Deposit Account' customers. **However, such accounts enjoying additional facilities will not be treated as BSBDAs.**
Query-14: If BSBDA customers have more than 4 withdrawals and request for cheque book at additional cost, will it cease to be a BSBDA?
Response: Yes. Please refer to response to the above query (Query No.13). **However, if the bank does not levy any additional charges and offers more facilities free than those prescribed under BSBDA a/cs without minimum balance then such a/cs can be classified as BSBDA.**
Query-15: Whether the existing facility available in a normal saving bank account of Five free withdrawals in a month in other banks ATMs as per IBA (DPSS) instructions will hold good for BSBDA?
Response: No. In BSBDA, banks are required to provide free of charge minimum four withdrawals, through ATMs and other mode including RTGS/NEFT/Clearing/Branch cash withdrawal/transfer/internet debits/standing instructions/EMI etc. It is left to the banks to either offer free or charge for additional withdrawal/s. However, in case the banks decide to charge for the additional withdrawal, the pricing structure may be put in place by banks on a reasonable, non-discriminatory and transparent manner by banks.
Query-24: In terms of RBI circular DPSS. CO.CHD. No. 274/03.01.02/2012-13 dated August 10, 2012, if "payable at par" / "multi-city" cheques are issued to BSBDA customers based on their request, can banks prescribe minimum balance requirements?
Response: BSBDA does not envisage cheque book facility in the minimum facilities that it should provide to BSBDA customers. **They are free to extend any additional facility including cheque book facility free of charge (in which case the account remains BSBDA) or charge for the additional facilities (in which case the account is not BSBDA).**
Query-25: What is the definition of "Basic Savings Bank Deposit Account" (BSBDA)?
Response: All the existing 'No-frills' accounts opened pursuant to guidelines issued vide circular DBOD. No. Leg. BC. 44/09.07.005/2005-06 dated November 11, 2005 and converted into BSBDA in compliance with the guidelines issued in circular DBOD.No.Leg.BC.35/09.07.005/20012-13 dated August 10, 2012 as well as fresh accounts opened under the said circular should be treated as BSBDA. **Accounts enjoying additional facilities under the reasonable pricing structure for value added services, exclusively for BSBDA customers should not be treated as BSBDAs.**

It took RBI nearly six years to recognise the serious defects in the formulation of the BSBDA as a savings deposit product. Subsequent to highlighting lacuna in the regulation on BSBDA, effective July 1, 2019, RBI further refined the definition of BSBDA. This time, for a BSBDA, RBI allowed banks to impose service charges (if they so desire) on debit transactions (beyond four a month) subject to extant laws and reasonableness of the charges. For insights into the genesis of this change in definition of BSBDA, see Das (2017, 2018).

Such a refinement in the definition of BSBDA made much sense. Banks were earlier prohibited to provide value-added services for a fee (including debit transactions beyond four a month) so long as the account is a BSBDA. Therefore, in order to comply with the regulation then, many banks had resort to either debarring debits beyond four a month in a BSBDA or, continued providing fee debit transactions beyond four a month. For more details, see Das (2017, 2018).

The new BSBDA regulation made it amply clear that there should arise no rationale for banks to anymore debar debit transactions beyond four a month, since banks now had the freedom to charge for debit transactions beyond four a month in a BSBDA (so long as the charges are reasonable). Moreover, RBI also clearly indicated that the BSBDA shall be considered a normal banking service available to all.

1.2. Mandates set by RBI towards ensuring reasonableness in service charges

RBI in their extant July 2015 notification on 'Master Circular on Customer Service in Banks' sets mandates onto banks towards ways and means of Fixing Service Charges and Ensuring Reasonableness of Bank Charges. The actions required to be taken by banks is indicated under the column 'action points for banks' in the Annex I to the above-mentioned master circular. The actions include:

A. While **Fixing Service Charges for *various types* of services** like charges for cheque collection, etc., banks should ensure that the **charges are reasonable and are not out of line with the average cost of providing *these* services**. The Bank's Board of Directors has been vested with the responsibility to ensure the reasonableness of such charges.

B. Regarding **Ensuring Reasonableness of Bank Charges,** in order to guarantee fair practices in banking services, RBI had constituted a Working Group to formulate a scheme for ensuring reasonableness of bank charges. Based on the recommendations of the Group, action required to be taken by banks is indicated in the Master Circular. The actions include:

i) Identification of basic banking services, where the prime parameter for identifying the basic banking services relates to deposit accounts and remittance services.
Telegraphic Transfer, ECS, NEFT and EFT are among the then identified basic remittance services, and would additionally include services considered appropriate towards basic services for deposit accounts and remittance services. Accordingly, IMPS, UPI, BHIM-UPI, and debit cards (merchant payments), each fall under basic banking services.
When transactions occur in different delivery channels, for the purpose of pricing, they are to be treated on a separate footing. (This had been in breach.)

ii) Offering basic banking services outside the scope of bundled products. Here, some of the banks do not levy charges on each individual product or service. Products and services are bundled and offered to a customer as a composite offering. The bank recovers the cost of these operations through net interest income. The bank achieves break-even levels through higher average balances in customer accounts which yield healthy interest margins or by imposing charges for keeping inadequate balances.
In so far as the basic services are concerned, the banks' objective should be to ensure that these are made available to the users at reasonable prices/charges and towards this, the basic services should be delivered outside the scope of the bundled products. (This had been in breach.)

iii) Principles for ensuring reasonableness in fixing the service charges include
a. **For basic services rendered to special category of individuals (such as individuals in rural areas, pensioners and senior citizens), banks will levy charges on more liberal terms than the terms on which the charges are levied to other individuals.** (This had been in breach.)
b. For the basic services rendered to individuals, **banks will levy charges only if the charges are just and supported by reason.** (This had been in breach.)
c. For the basic services to individuals, **the banks will levy service charges ad-valorem only to cover any incremental cost and subject to a cap.**

RBI has emphasised that banks have to adhere to these guidelines. Under the above principles mandated by RBI, as an illustration, when a bank imposes a charge of Rs. 20 for an unassisted online- and mobile-based digital transaction, the bank has to establish that such charges are not out of line with the average cost of providing the unassisted digital services. Moreover, if the bank charges the same Rs. 20 even for a cash withdrawal transaction carried out at an ATM, or microATM or over-the-counter, the question of ensuring the RBI's fundamentals on reasonableness of the charges towards digital transactions becomes more difficult to meet. More so, since it is well established that for banks, cash is an expensive mode to transact than the relatively cheaper digital transfer of funds. Also, the RBI's mandate had always been to encourage BSBDA customers (including PMJDY account holders) to adopt digital payment through the prescribed digital modes vis-à-vis the cash transactions.

The Board of Directors is vested with the responsibility to ensure the reasonableness of such service charges and are liable for the same. The board's dereliction, if any, in ensuring reasonableness of charges for debit transactions via digital means like NEFT, IMPS, UPI, BHIM-UPI, and debit cards (merchant payments) needs an in-depth understanding.

1.3. The theme of the present article

While defining the features of a BSBDA (for the period September 2013 through June 2019), the regulatory requirements made it amply clear that in addition to mandatory free banking services (that included four withdrawals per month), so long as the savings bank account is a BSBDA (rather than a non-BSBDA savings bank account), the banks cannot impose any charge even for value-added banking services that a bank may like to offer at their discretion; and RBI considers a withdrawal, beyond four a month, a value-added service.

There had been a systematic breach in the RBI regulations on BSBDAs by few banks, most notably by State Bank of India (SBI) that hosts the maximum number of BSBDAs, when it charged @ Rs. 17.70 for every debit transaction beyond four a month. Such transactions

comprised even the non-cash digital transactions done through NEFT, IMPS, UPI, BHIM-UPI and the debit card for merchant payments.

Such imposition of service charges resulted in undue collections to the tune of over Rs. 300 crore from among nearly 12 crore BSBDA customers of SBI during the period 2015-20, of which the period 2018-19 alone saw collection of Rs. 72 crore and the period 2019-20, Rs. 158 crore.

We assess the dereliction in SBI's duty towards the PMJDY when the BSBDA users were unduly (and against the extant regulations) forced to part with such high charges for their day-to-day (non-cash) digital debit transactions that the bank determinedly imposed.

The core contention that is highlighted here has no intention of disrupting or creating any hindrance towards the stability of SBI but rather to support those affected due to non-adherence of a technicality by SBI. This involves over 12 crore savings account holders of SBI who under the PMJDY were brought into the reach of financial inclusion. These relatively vulnerable, gullible and marginalized fellow-countrymen being thrust with charges @ Rs. 17.70, every time they transacted digitally (unassisted non-cash means), is shown to be grossly unreasonable, exploitative and unjust.

2. PMJDY Depositors – the Unprotected Prey

2.1. The contribution of Public Sector Banks towards PMJDY

The public sector banks (PSBs) have significantly contributed towards the Prime Minister's mission on financial inclusion. The PMJDY mission of the government, in their FAQs on PMJDY explains that PMJDY accounts are BSBDAs in nature with additional facility of RuPay Debit card with accident insurance coverage and an overdraft facility. As per DFS's weekly PMJDY data, at the end of calendar year 2020, of the PMJDY accounts opened, 97% is attributed to the 12 PSBs in India. The contribution from the 21 private sector banks is a bare minimum of 3% of the PMJDY accounts opened.

Among the 12 PSBs, SBI has significantly contributed in the financial inclusion drive of the PMJDY. The contribution towards opening of PMJDY accounts by SBI is over one-third of the PMJDY accounts opened by the PSBs. As of end-December 2020, SBI had 12.8 crore PMJDY accounts to their credit (**with average balances of Rs. 2700**), which contribute to about 39% accounts among PSBs and 31% among all banks. The four banks, Bank of Baroda, Punjab National Bank, Bank of India and Union Bank of India (henceforth called the *big four*) together have a total of 13.3 crore PMJDY accounts (just a bit more than SBI's tally of 12.8 crore).

RBI publishes the BSBDA data in their annual report, where they categorise such accounts as "through branches" and "through BCs". As of December 2020, data shows that 55% of the BSBDAs are categorised as "through BCs".

2.2. Overall collections toward service charges from BSBDA depositors by PSBs

In order to remain informed on the approach taken by PSBs towards extracting a fee for day-to-day debit transactions in a BSBDA, we collected information from the 12 PSBs. The specific queries related to the amount of service charges collected from BSBDA customers during the period 2014-20. In response to the specific queries, most of the PSBs responded, which is summarised in Table 1.

Among the 12 PSBs, seven major banks like Bank of Baroda, Bank of India, Union Bank of India, Central Bank of India, UCO Bank, Bank of Maharashtra and Punjab & Sind Bank, in compliance of the RBI regulation, *did not* impose any service charges on BSBDAs and accordingly their service charge collection had been Nil. However, among the rest, the information prominently highlights that SBI collected over Rs. 300 crore (Rs. 265 crore) over the past five 2015-20 (three 2017-20) financial years. The four others banks collected bare minimum amounts.

Table 1: Service charges collected by PSBs in BSBDAs

Charges collected (Rs Cr)	2014-20	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
SBI	308.32	4.73	12.44	26.31	**34.74**	**72.07**	**158.03**
Bank of Baroda	**Nil**	Nil					
Punjab National Bank	9.90	0.66	1.40	1.17	1.44	2.18	3.05
Bank of India	**Nil**	Nil					
Union Bank of India	**Nil**	Nil					
Indian Bank	?						
Canara Bank	?						
Central Bank of India	**Nil**	Nil					
UCO Bank*	**Nil**	Nil					
Bank of Maharashtra	**Nil**	Nil					
Indian Overseas Bank	5.28	0	0	0.01	0.42	2.46	2.39
Punjab & Sind Bank	**Nil**	Nil					

*There is a certain lack of clarity in UCO Bank's response
? Complete data not available
Source: Written response from banks based on RTI queries

Notes:

- SBI data pertains to PMJDY accounts. SBI did not provide data on BSBD - branch channel accounts.
- Punjab National Bank data pertains to BSBDAs other than PMJDY accounts. No charges were collected for accounts under PMJDY.
- Indian Bank data in explicit form not available for BSBDA. However, no charges were collected for accounts under PMJDY.
- Canara Bank data in explicit form not available.
- Before merger, Oriental Bank of Commerce had also responded as 'Nil'.
- Before merger, United Bank of India had also responded as 'Nil'.

While the *big four* together served more PMJDY accounts (13.3 crore) than what SBI did (just about 12.8 crore), in terms of service charge collections, SBI collected over Rs. 300 crore against near nil collections by the *big four*. In Table 2, the columns corresponding to

'*Proportion relative to SBI*', SBI is 100%. For other 11 PSBs, the revenue collected (in %) is negligible, if not zero, whereas the number of accounts served (in %) is significantly large. Therefore, it transpires that there exists a significant bias to the disadvantage of the 11 PSBs (vis-à-vis SBI), just because, unlike SBI, majority of the PSBs had been compliant to the BSBDA regulations that prohibited imposition of service charges.

Table 2: Service charges collected by PSBs in BSBDAs relative to SBI

Public Sector Banks	**Revenue collected 2014-20 (Rs Cr)**	Proportion relative to SBI revenue (%)	**No. of Accounts (Cr)**	Proportion relative to SBI Accounts (%)
SBI	308.32	-	12.8	-
Bank of Baroda	**Nil**	0	4.8	37.6
Punjab National Bank	9.90	3.2	3.9	30.9
Bank of India	**Nil**	0	2.5	19.5
Union Bank of India	**Nil**	0	2.0	16.0
Indian Bank	?	?	1.8	13.9
Canara Bank	?	?	1.4	10.7
Central Bank of India	**Nil**	0	1.4	11.0
UCO Bank*	**Nil**	0	1.0	7.6
Bank of Maharashtra	**Nil**	0	0.7	5.2
Indian Overseas Bank	5.28	1.7	0.5	4.1
Punjab & Sind Bank	**Nil**	0	0.1	1.0

Source: DFS's weekly PMJDY data as of 30/12/2020 and Table 1

3. Unreasonable and Undue Charges Imposed by SBI

Among the PSBs, SBI is credited to be the major contributor towards the country's financial inclusion mission. SBI's contribution in the country's drive towards financial inclusion is noteworthy – SBI did a commendable job. Alongside, it is also pertinent to keep track of issues surrounding customer centricity and protection of such a vulnerable group against acts not in line with the letter and spirit behind regulations framed for this very group of bank depositors.

As of June 2019, SBI had about 11.2 crore BSBDAs under the PMJDY. SBI currently has over 12 crore BSBDAs categorized as "through BCs", which are mostly the PMJDY accounts. Additionally, SBI has nearly two crore BSBDAs categorised as "through branches".

Given the size of such BSBDAs opened by SBI, our primary focus is on how SBI mishandled such accounts in breach of the fundamentals laid down by RBI to protect people brought under the ambit of the government's and the Prime Minister's financial inclusion drive.

In breach of RBI's extant regulations framed under the Banking Regulation Act, 1949, SBI charged the gullible lot – the BSBDA customers. This imposition of service charges resulted in undue collections to the tune of over Rs. 300 crore from among nearly 12 crore BSBDA customers of SBI during the period 2015-20, of which the period 2018-19 alone saw collection of Rs. 72 crore and the period 2019-20, Rs. 158 crore.

To begin with, a regulation had clearly been set (August 2012) that a monthly maximum of 4 debit transactions are allowed in a BSBDA, which are to be provided free of any charge. Thereafter, just to accommodate the bank's desire to offer more than 4 debits in a month, RBI

refined the regulation (September 2013) and provided for a minimum of 4 free debits in a month for a BSBDA. This allowed the banks, if they so desired, to allow more than 4 free debits in a month. That become the definition of a BSBDA, which inherently intends to protect such accounts against being charged for debit transactions. It was only from July 2019 that RBI allowed banks to reasonably charge a BSBDA beyond four free debits in a month.

3.1. SBI's reasonableness in fixing a charge of Rs. 17.70 for every digital transaction

Even if we keep aside the technicalities involving RBI's September 2013 circular, the question still remains as to how SBI could have been compliant with respect to the RBI's August 2012 and July 2015 mandates on "Principles for ensuring reasonableness in fixing the service charges"? Such reasonableness of charges was to be ensured by the banks' Board of Directors based on the regulatory principles as mandated by RBI (see, Section 1.2). Under such principles mandated by RBI, when SBI imposes a charge of Rs. 17.70 for every unassisted debit card-, online- and mobile-based digital transaction, are such charges just? **The charge of Rs. 17.70 for an unassisted digital debit transaction, is neither reasonable nor just.**

3.2. Government makes SBI liable to refund over Rs. 90 crore for charges imposed

Effective January 1, 2020, the Payment and Settlement Systems (PSS) Act, 2007, prohibited SBI to charge on any debit transaction done using UPI (BHIM-UPI) and RuPay debit card (for merchant payments). In breach of this law, SBI charged @ Rs. 17.70 for UPI and RuPay debit card digital transactions from the gullible BSBD-BC channel account holders during the period January-September 2020. NPCI disseminates monthly UPI and Rupay POS/eCom digital transaction volumes.

Table 3 provides the number of monthly UPI transactions, with SBI as the remitter bank. In the year 2020, there were 1888.08 crore overall UPI transactions, of which a total of 533.55 crore UPI debit-transactions were carried out by SBI alone. This indicates that 28.26% of UPI transactions in the year 2020 can be attributed to UPI debit-transactions of SBI. The monthly % contribution of SBI Volume is provided for 12 months in Table 3. The monthly average % Volume is 28.27%, with a standard deviation σ of 1.14%.

Table 3: UPI Transactions – Overall and SBI (Remitter Bank)

Month	Total Volume (Crore)	SBI Volume (Crore)	% SBI Volume
Jan-20	130.50	34.75	26.62
Feb-20	132.57	34.98	26.39
Mar-20	124.68	33.82	27.12
Apr-20	99.96	29.85	29.86
May-20	123.45	36.56	29.62
Jun-20	133.69	39.06	29.22
Jul-20	149.74	43.93	29.34
Aug-20	161.88	44.70	27.61
Sep-20	180.01	51.02	28.34
Oct-20	207.16	59.92	28.93
Nov-20	221.02	60.79	27.50
Dec-20	223.42	64.16	28.72
Monthly Avearge	157.34	44.46	28.27
Year 2020	1888.09	533.55	28.26

Source: NPCI and author's computation

Similarly, Table 4 provides the number of monthly RuPay POS and eCom transactions. The share of SBI in the overall RuPay transaction figures is taken a 10%. Accordingly, in Table 4, for the year 2020, we have derived a column representing the monthly contribution of SBI Volume towards RuPay POS and eCom transactions.

Table 4: RuPay POS and eCom Transactions – Overall and SBI (Issuer Bank)

Month	Total Volume (Crore)	SBI Volume (Crore)	% SBI Volume
Jan-20	14.04	1.40	10
Feb-20	13.16	1.32	10
Mar-20	11.67	1.17	10
Apr-20	7.25	0.73	10
May-20	9.11	0.91	10
Jun-20	10.09	1.01	10
Jul-20	10.71	1.07	10
Aug-20	11.54	1.15	10
Sep-20	12.21	1.22	10
Oct-20	13.36	1.34	10
Nov-20	12.79	1.28	10
Dec-20	12.91	1.29	10
Monthly Avearge	11.57	1.16	10
Year 2020	138.84	13.88	10

Source: NPCI and author's computation

Central Board of Direct Taxes (CBDT) during August 2020, advised banks to refund the charges collected, if any, on or after January 1, 2020 on transactions carried out using the prescribed digital mode and not to impose charges on future transactions carried out through such modes. In terms of CBDT directives, starting February 17, 2021, SBI initiated the refunds @ Rs. 17.70 in respect of UPI and RuPay debit card digital transactions to the BSBDA customers. The refund process was completed on March 31, 2021. **SBI refunded the charges imposed during January-September 2020 for 5.1 crore UPI and RuPay debit card digital transactions, amounting to Rs. 90.2 crore.**[1]

Irrespective of what CBDT invoked under the PSS Act, 2007, SBI is still silent on the question of ensuring reasonableness in fixing a charge @ Rs. 17.70 for every digital transaction, as per RBI's July 2015 mandates. The board of directors of SBI, who is supposed to have ensured the reasonableness in fixing the service charges found it appropriate to consider an ATM/microATM cash withdrawal cost for the bank to be at par with an unassisted digital transfer through means like UPI, BHIM-UPI, RuPay debit card, NEFT, etc. This is sheer oversight in assessing reasonableness.

3.3. Extent of SBI's undue interest gains from charges imposed on UPI/ RuPay debit card digital transactions

The marginalized section of the BSBD-BC Channel account customers of the bank have been charged in a discriminatory fashion. SBI charged for nearly six months in 2020, i.e.,

[1] As provided by SBI, the exact figure of refunds for service charges imposed on 5,09,53,806 UPI and RuPay debit card digital transactions is Rs. 90,19,04,466.

during January 1 - April 6, 2020 and during July 1 - September 14, 2020. We work out the enrichment derived by SBI due to the charges imposed and subsequent delayed refund of the same. SBI enriched themselves by depriving the customers of the savings rate of interest on the undue collections made and the advantage taken by SBI in form of interest earned of the collected funds at Reverse Repo (bare minimum). The extent of this enrichment alone is Rs. 4.65 crore. Table 5 provides the corresponding calculations, considering (5a) the Reverse Repo and the Savings Bank rates, (5b) the number of days that generated interest at the various rates, and finally (5c) the interest gains from the undue UPI/ RuPay debit card digital charges collected during January-September 2020. **SBI has not passed on the undue interest gains of Rs. 4.65 crore to their depositors.**

Table 5: SBI's gains through interest from undue charges imposed during Jan-Sep 2020

5a: Reverse Repo and the Savings Bank rates

R Repo Rate in 2020-21		SBI's Savings Bank Rate in 2020-21	
Rate prevailing until		Rate prevailing until	
26 March 2020	4.9	13 March 2020	3.25
16 April 2020	4	18 April 2020	3
21 May 2020	3.75	30 May 2020	2.75
31 March 2021	3.35	31 March 2021	2.7

Source: RBI and SBI

5b: Number of days that generated interest at the various rates

Month	Savings Bank Rate	Reverse Repo Rate	SB Days	R Repo Days	Savings Bank Rate	Reverse Repo Rate	SB Days	R Repo Days	Savings Bank Rate	Reverse Repo Rate	SB Days	R Repo Days	Savings Bank Rate	Reverse Repo Rate	SB Days	R Repo Days
Jan-20	3.25	4.9	58	71	3	4	36	21	2.75	3.75	42	35	2.7	3.35	285	294
Feb-20	3.25	4.9	28	41	3	4	36	21	2.75	3.75	42	35	2.7	3.35	285	294
Mar-20	3.25	4.9	7	13	3	4	27	19	2.75	3.75	42	35	2.7	3.35	285	294
1-6 Apr 2020	-	-	-	-	3	4	15	13	2.75	3.75	42	35	2.7	3.35	285	294
Jul-20	-	-	-	-	-	-	-	-	-	-	-	-	2.7	3.35	239	239
Aug-20	-	-	-	-	-	-	-	-	-	-	-	-	2.7	3.35	208	208
1-14 Sep 2020	-	-	-	-	-	-	-	-	-	-	-	-	2.7	3.35	185	185

5c: Interest gains from the undue UPI/ RuPay debit card digital transaction charges collected

Month	No. UPI Debit Txn (Crore)	No. RuPay Digital Txn (Crore)	No. UPI/RuPay Txn (Crore)	Total amount refunded (Rs Crore)	Savings interest (Rs Crore)	R Repo interest (Rs Crore)	Total enrichment (Rs Crore)
Jan-20	34.75	1.40	36.15	14.25	0.46	0.60	1.07
Feb-20	34.98	1.32	36.30	14.31	0.42	0.55	0.97
Mar-20	33.82	1.17	34.99	13.79	0.37	0.47	0.85
1-6 Apr 2020	5.97	0.15	6.11	2.41	0.06	0.08	0.14
Jul-20	43.93	1.07	45.00	17.74	0.31	0.39	0.70
Aug-20	44.70	1.15	45.86	18.08	0.28	0.35	0.62
1-14 Sep 2020	23.81	0.57	24.38	9.61	0.13	0.16	0.29
Total Txns	**221.96**	**6.83**	**228.79**	90.19	2.04	2.60	**4.65**
Charges imposed on	**5.10**						
% of Txns charged on	**2.23**						
Charges per Txn	**Rs 17.70**						

Source: Tables 3, 4 and author's computation

Though 5.1 crore is the number of UPI and RuPay debit card digital transactions on which SBI imposed charges during January-September 2020, and subsequently refunded @ Rs. 17.70 per transaction, SBI could not provide the breakup for each of UPI and RuPay debit card digital transactions separately. In 2020, SBI attributed to about 222 crore UPI transactions and about

6.8 crore RuPay (POS and eCom) debit card digital transactions during the periods when SBI had imposed the charges. Of these UPI and RuPay debit card digital transactions, only 5.1 crore transactions, i.e., 2.23% of the SBI's UPI/ RuPay debit card digital transactions were charged by SBI @ Rs. 17.70.

3.4. Extent of SBI's exploitation from UPI/ RuPay debit card digital transactions

SBI's imposition of Rs. 17.70 for an unassisted digital debit transaction cannot be considered reasonable, and thus is in breach of the RBI's July 2015 mandates. Accordingly, SBI should technically refund such charges recovered in respect of digital debit transactions to the BSBDA customers with effect from April 1, 2017, rather than only for the period January 1, 2020 to September 14, 2020, for which SBI has already refunded Rs. 90.2 crore.

We assess the amount of money collected by SBI during financial years FY18, FY19 and FY20, towards charges imposed on UPI and RuPay debit card digital transactions. To derive the same, we use the fact that the monthly average percentage of SBI volumes for UPI transactions is 28.27% of the overall UPI volumes, with a standard deviation σ of 1.14% (Table 4). Thus, the 2σ lower bound for the monthly average percentage of SBI volumes for UPI transactions is 25.99%. Similarly, for RuPay debit card digital transactions, as a conservative estimate, we have taken the SBI's share of transaction volume for POS and eCom combine as 10% of the overall RuPay debit card digital transaction volume. We use these statistics to arrive at the amount of service charges collected by SBI of UPI transactions from the BSBDA customers during the three financial years 2017-20.

Table 6 shows the workout of the undue and unreasonable UPI/ RuPay debit card digital transaction charges collected by SBI that has still not been refunded. In FY18, FY19 and FY20, SBI collected an undue sum of about Rs. 10.1 crore, Rs. 60.1 crore and Rs. 137.1 crore respectively, when they charged @ Rs. 17.70 per UPI transaction. Of the UPI charges collected in FY20, SBI refunded Rs. 40.8 crore only, while retaining nearly Rs. 96.3 crore. Thus, SBI did not refund Rs. 166.5 crore, which is the undue sum collected for the three financial years 2017-2020. Even if we give a benefit of doubt to SBI and consider the 2σ lower limits of 25.99% as the SBI's UPI contribution, then too the withheld undue amount is atleast Rs. 153 crore. In order words, we can say with more than 95% confidence that SBI has collected more than Rs. 153 crore towards charges collected for UPI transactions that it has still not returned back to our marginalized countrymen while they used their accounts to transact digitally.

The bottom half of Table 6 derives the undue collections of RuPay debit card digital transactions carried out during FY18, FY19 and FY20, as Rs. 2.4 crore, Rs. 4.4 crore and Rs. 5.8 crore respectively. Of the RuPay debit card digital transaction charges collected in FY20, SBI refunded Rs. 1.5 crore only, while retaining nearly Rs. 4.3 crore. It may be noted that charges imposed for transacting digitally using a RuPay card is a gross violation in several aspects. More so, since for every RuPay transaction, SBI as an issuer bank earned payment card interchange in the range of 0.15% - 0.65% of the transaction amount.

From Table 6, we see that during the 33 months April 2017 through December 2019, SBI has collected anywhere between Rs. 164 crore and Rs. 177 crore towards charges imposed on at least 9 crore UPI and RuPay debit card digital transactions done by BSBDA customers. SBI has still not refunded this improperly collected amount to the BSBDA customers.

Table 6: Extent of charges imposed for UPI/ RuPay debit card digital transactions

Period →	Jan-Mar 2020	Apr-Dec 2019	FY19	FY18	FY18-FY20 Undue Collections (Rs Cr)
Overall UPI Txn (Cr)	387.76	864.11	539.15	91.52	
SBI's UPI Txn @ 28.27% of overall Txn (Cr)	103.55	244.31	152.44	25.88	
SBI's UPI Txn @ 25.99% of overall Txn (Cr)	103.55	224.61	140.14	23.79	
2.23% of SBI's UPI Txn @ 28.27% (Cr)	2.31	5.44	3.39	0.58	
Average : SBI's UPI Txn charges collected (Rs Cr)	40.82	96.31	60.09	10.08	166.47
2.23% of SBI's UPI Txn @ 25.99% (Cr)	2.31	5.00	3.12	0.53	
2σ Lower Limit : SBI's UPI Txn charges collected (Rs Cr)	40.82	88.54	55.24	9.26	153.05
Overall RuPay POS and eCom Txn (Cr)	38.87	109.20	112.71	66.77	
SBI's RuPay Txn @ 10% of overall Txn (Cr)	3.89	10.92	11.27	6.68	
2.23% of SBI's RuPay Txn @ 10% (Cr)	0.09	0.24	0.25	0.15	
Average : SBI's RuPay POS and eCom Txn charges collected (Rs Cr)	1.53	4.30	4.44	2.41	11.15
SBI's Overall Charges collected (Rs Cr)	158.03		72.07	34.74	

Source: NPCI and author's computation

4. Concluding Remarks

The Board of Directors of SBI have been vested with the responsibility to ensure that charging Rs. 17.70 for every UPI and RuPay debit card digital transaction is reasonable as per principles laid down by RBI. Given that it costs a bank disproportionately more to provide an ATM cash withdrawal service than to provide the UPI/ RuPay debit card digital transaction facility, SBI's imposition of a uniform charge of Rs. 17.70 for both ATM cash withdrawal and UPI transaction is grossly unreasonable and in breach of RBI regulations.

Though on the one hand the country envisages a less-cash society while on the other hand the BSBDA customers had been disincentivised in their digital transactions for day-to-day payments. While having embraced digital means for transacting, the BSBDA customers remained an unprotected lot since SBI's actions amounted to exploitation of this marginalized section of the society through imposition of usurious service charges.

References

Annual Report of the RBI for the Year 2020-21. RBI, May 27, 2021.

Basic Statistical Return (BSR-2) – *Deposits with Scheduled Commercial Banks in India* - March 2020. RBI, October 14, 2020.

Das, A. (2021b). *Statement of Facts - Relating to SBI's 'Statement of clarification on media reports based on the IIT Bombay Study'*. IIT Bombay, April 18, 2021.

Das, A. (2021a). *Regulating Basic Savings Bank Deposit Accounts - Do we need to care for these marginalized depositors?* IIT Bombay Technical Report, April 11, 2021.

Das, A. (2020). *Deviating from the BHIM-UPI Law*. IIT Bombay Technical Report, August 24, 2020.

Das, A. (2018). *Can BSBDA Depositors have Long Innings? – Be Aware and Remain Vigilant*. IIT Bombay Technical Report, May 26, 2018.

Das, A. (2017). *Basic Savings Bank Deposit Account – Complex Design, Faulty Implementation*. IIT Bombay Technical Report, June 26, 2017.

Financial Inclusion- Access to Banking Services – Basic Savings Bank Deposit Account (BSBDA). RBI/2018-19/206 DBR.LEG.BC.No.47/09.07.005/2018-19 dated June 10, 2019.

Financial Inclusion- Access to Banking Services – Basic Savings Bank Deposit Account (BSBDA) – FAQs. DBOD.No. Leg.BC.52/09.07.005/2013-14 dated September 11, 2013.

Financial Inclusion- Access to Banking Services – Basic Savings Bank Deposit Account. DBOD.No. Leg. BC.35/09.07.005/2012-13 dated August 10, 2012.

Financial Inclusion. DBOD.No.Leg.BC. 44/09.07.005/2005-06 dated November 11, 2005.

Master Circular on Customer Service in Banks. RBI/2015-16/59 DBR No. Leg. BC. 21/09.07.006/2015-16 dated July 1, 2015.

Statement on Developmental and Regulatory Policies - October 4, 2017- Banking Facility for Senior Citizens and Differently abled Persons. RBI/2017-18/89 DBR.No.Leg.BC.96/09.07.005/ 2017-18 dated November 9, 2017.

Publisher
Society of Statistics, Computer and Applications
B-133, Ground Floor, C.R. Park, New Delhi - 110019
Tele: 011 - 40517662
https://ssca.org.in/
statapp1999@gmail.com
2021

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203