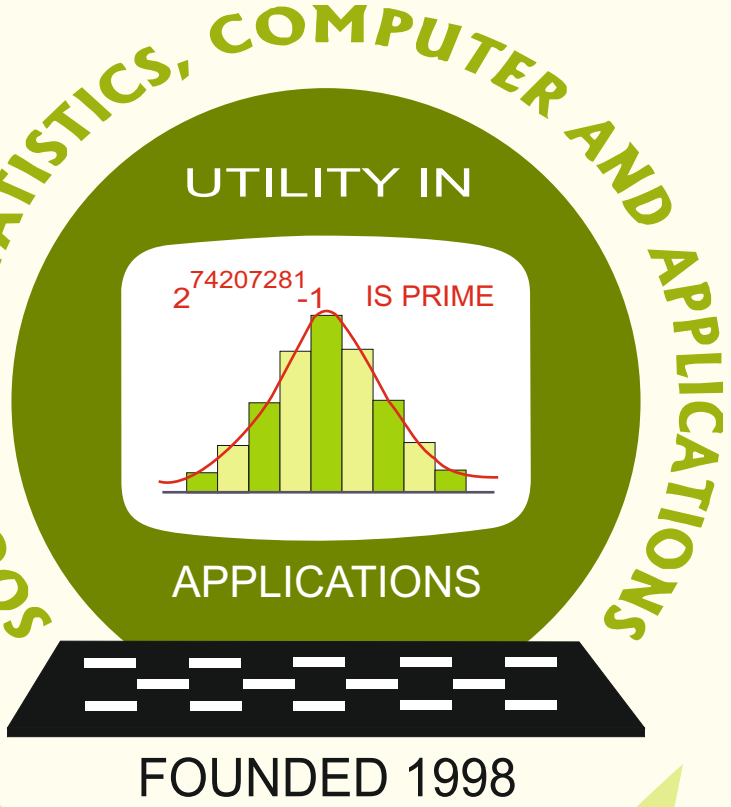
ISBN 978-81-950383-5-0

Special Proceedings (26)
(Based upon the 26th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2024 held at the Department of Mathematics and Statistics and Centre for Artificial Intelligence, Banasthali Vidyapith, Banasthali - 304022, Rajasthan, India, during February 26-28, 2024)



Society of Statistics, Computer and Applications
https://ssca.org.in/
2024

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
G.P. Samanta, K.J.S. Satyasai, Prithvi Yadav, Pankaj Mittal
R.B. Barman, R.C. Agrawal, Rahul Mukerjee, Rajpal Singh

Vice Presidents

A. Dhandapani, Manish Kumar Sharma, Manisha Pal, P. Venkatesan
Praggya Das, Ramana V. Davuluri, S.D. Sharma, V.K. Bhatia

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury, Vishal Deo

Council Members

B. Re. Victor Babu, Banti Kumar, Imran Khan, Mukesh Kumar
Parmil Kumar, Piyush Kant Rai, Rajni Jain, Rakhi Singh
Raosaheb Vikaram Latpate, Renu Kaul, Sapam Sobita Devi, Shalini Chandra
V. Srinivasa Rao, V.M. Chacko, Vishnu Vardhan, R.

Ex-Officio Members (By Designation)

Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Chair Editor, Statistics and Applications
Executive Editor, Statistics and Applications

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

ISBN 978-81-950383-5-0

Special Proceedings (26)
(Based upon the 26th (Annual) International Conference of the Society of Statistics, Computer and Applications (SSCA) - 2024 held at the Department of Mathematics and Statistics and Centre for Artificial Intelligence, Banasthali Vidyapith, Banasthali - 304022, Rajasthan, India, during February 26-28, 2024)

Editors

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

Date of Publication: October 25, 2024

Published By:
Institutional Publisher: Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2024

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

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024

CONTENTS

PREFACE

We are pleased to present the special proceedings of the twenty-sixth annual conference of the Society of Statistics, Computer, and Applications (SSCA), organized by Department of Mathematics and Statistics and the Centre for Artificial Intelligence, Banasthali Vidyapith, Banasthali – 304022, Rajasthan, India, during February 26-28, 2024. This conference was a key component of the broader international event focused on "Emerging Trends of Statistical Sciences in AI and its Applications, ETSSAA-2024".

Founded in 1998 with its inaugural gathering at Haryana Agricultural University, Hisar, SSCA has consistently organized annual conferences across various educational institutions nationwide. The society's core mission focuses on fostering research at the intersection of Statistics and Information Technology, serving both theoretical and applied statisticians committed to advancing technology for societal progress. SSCA also promotes open access to knowledge through its journal 'Statistics and Applications,' facilitating free downloads, saving, and printing of full-length papers. In addition to regular issues, the journal periodically releases special volumes addressing globally and nationally significant thematic areas.

The recent 26th conference aimed to provide a unified platform for deliberations on regional and global statistical issues. Distinguished experts in theoretical and applied statistics from India and abroad, notably from the USA, actively engaged in dissemination of knowledge. Speakers represented prestigious Indian institutions such as the Indian Statistical Institute, IITs, ICAR, RBI, senior governmental offices, and universities. The conference featured several noteworthy events, including a preconference workshop and various technical sessions. These sessions encompassed the M.N. Das Memorial Lecture and a dedicated session on Financial Statistics, where distinguished statisticians and leading practitioners in the field shared their insights on various finance related topics. Moreover, the conference included three endowment lectures: the B.K. Kale Memorial Endowment Lecture, J.K. Ghosh Memorial Endowment Lecture, and Bikas Kumar Sinha Endowment Lecture. These lectures were delivered by speakers closely associated with, or collaborated with, or students of the respective honourees. Additionally, there was the V.K. Gupta Endowment Award Lecture for Achievements in Statistical Sciences and Practice, delivered by J. Sunil Rao from USA. The SSCA introduced this year the Aditya Shastri Memorial Lecture in memory of Late Aditya Shastri, Vice Chancellor, Banasthali Vidyapith who left for heavenly abode in 2021 due to COVID - 19. The first lecture was delivered by Navin M. Shastri.

The Executive Council of SSCA resolved to compile "Special Proceedings," highlighting selected presentations, including those from the specialized Financial Statistics session. The Guest Editors appointed by the Council—V.K. Gupta, Baidya Nath Mandal, R. Vishnu Vardhan, Ranjit Kumar Paul, Rajender Parsad, and Dipak

Roy Choudhury—meticulously curated these proceedings. While constraints limited the inclusion of all invited papers, esteemed speakers were invited to submit their research contributions. Following a rigorous review process, a distinguished selection of 12 papers was accepted for publication in the special proceedings. We extend our sincere gratefulness to all the authors for prompt submission of high-quality research papers for special proceedings. We owe special debt to all the reviewers whose dedicated efforts ensured a swift and thorough review process that led to completing the review process within a short time frame. Special acknowledgment is also due to all members and office bearers of the SSCA-Executive Council for their steadfast support. Our gratitude to Ashish Das, Treasurer of the SSCA, for arranging funds for publishing these special proceedings. We are indeed grateful to Ms. Jyoti Gangwani for meticulously formatting the papers. Furthermore, our deepest appreciation goes to Prof. Ina Aditya Shastri, Vice Chancellor of Banasthali Vidyapith, and her dedicated team, including Prof. Anshuman Shastri, Director of the Centre for Artificial Intelligence, Prof. Sarla Pareek, Dean of the Faculty of Mathematics and Computing, Prof. Shalini Chandra, Head of the Department of Mathematics and Statistics, and other esteemed faculty members. Their collective efforts are a testament to the success of this intellectually enriching SSCA conference and advancing the field of Statistics and its applications in AI. The special proceedings of these sessions have been assigned the ISBN #: 978-81-950383-5-0.

We are confident that the contents encapsulated within these special proceedings will prove immensely beneficial to our readership, fostering further insights and advancements in the field of Statistics and its applications in AI and beyond. We welcome any suggestions for further improving future conferences and special proceedings, as we continually strive to serve the statistical community better.

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

New Delhi
September 2024

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 1-30

Bayesian Small Area Inference Using a Non-probability Sample

Balgobin Nandram
Department of Mathematical Sciences, Worcester Polytechnic Institute
100 Institute Road. Worcester, MA 01609

Received: 20 April 2024; Revised: 24 May 2024; Accepted: 28 May 2024

Abstract

We show how to use supplemental information from a small probability sample (ps) to do Bayesian predictive inference for finite population means of small areas using a relatively larger non-probability sample (nps). We focus on the most practical situation when there are common covariates in the nps and ps, where the nps has the study variable but no survey weights and the ps has survey weights but no study variable. We assume that the population model is correct and any functional relation between the study variable and the covariates is unspecified. Data preparation is necessary, and there are three steps, which are a double mass imputation, stratification of the population and allocation of the sample to the strata (domains), and creating a spatial structure to accommodate the covariates. Our main Bayesian analysis uses the conditional auto-regressive model, which helps to accommodate the covariates without incorporating them into the model, thereby avoiding a functional relation between the study variable and the covariates. However, the actual small areas are not part of the model, but we need to keep track of them, and the strata are modeled as the "small areas". Our procedure allows a small area (not a stratum) to participate in several strata, and this helps to mitigate over-shrinkage, which is common in small area models. Using an illustrative example on body mass index data, our method appears to work better than a standard method with a linear regression of the study variable on the covariates. Our new framework allows several extensions and it avoids an approximation used in survey design analysis.

Key words: BHF baseline model; Finite population mean; Gibbs sampler; Inverse probability weighting; Mass imputation; Robustness; Stratification; Surrogate samples.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

We have data from a non-probability sample, nps (1), and a probability sample, ps (2), relatively much smaller. The nps and ps have common covariates, $\underline{x}$, a p-vector including an intercept. The nps has the study variable (response), y, but no survey weights, W, and the

Corresponding Author: Balgobin Nandram
Email: balnan@wpi.edu & balgobin.nandram@usda.gov

ps has survey weights, W, but no study variable, y. We know the small areas (*e.g.,* counties) in the nps, but we do not need to know these small areas in the ps. The population size, N, may be unknown and the nonsampled covariates from the nps (or the ps) are unknown. These two quantities can be constructed from the ps sampled data. The population has $(\underline{x}_i, y_i), i = 1, \ldots, N$. Letting $y_{ij}, j = 1, \ldots, N_i, i = 1, \ldots, \ell$, denote the values of the study variable from the ℓ small areas, we want to make inference about the finite population mean of the i^{th} area,

$$\bar{Y}_i = \frac{1}{N_i} \sum_{i=1}^{N_i} y_i, i = 1, \ldots, \ell.$$

Again the N_i may be unknown, and these can be constructed from the ps.

Non-probability sampling has become important over the past decade. Many surveys are done without proper survey designs. Observational studies, which are very useful, have no probabilistic structure. A probability sample is the gold standard, but many data collection procedures are done by walking around a mall or standing at the door of a library. Clearly, these data are non-probability samples, and they are generally biased. On the other hand the response rates for many probability samples have declined to situations, where the sample can more reasonably be called a non-probability sample. Therefore, it is important to do research into the topic of non-probability sampling; indeed, this is a "hot" topic and it is difficult.

From the nps, we have $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, which are respectively the covariates and the study variable (response) but we do not have survey weights, $W_i, i = 1, \ldots, n_1$. From the ps, we have $(W_{2i}, \underline{x}_{2i}), i = 1, \ldots, n_2$, but not the study variable, $y_{2i}, i = 1, \ldots, n_2$. This is a fairly general and practical set up; see Li *et al.* (2020).

In our new approach, no participation (selection mechanism) model is needed. However, our standard assumption is that the population model for the study variable is correct and the sample model is derived from it. Therefore, robustness to its assumptions plays a key role in this paper.

In Table 1, we show different scenarios of massively missing data and what action can be taken, where all the data are missing for a specific variable, and mass imputation is needed; see Kim *et al.* (2021). For example, for the nps, all survey weights are missing.

Sakshaug *et al.* (2019), Wisniowski *et al.* (2020) and Salvatore *et al.* (2023) worked with Scenario 2a to provide a full Bayesian approach. They supplemented a probability sample with a much larger non-probability sample. They mainly studied superpopulation parameters, not finite populations. There is no penalty on the non-probability sample, except for Salvatore *et al.* (2023) where they used the idea of Nandram and Rao (2021, 2023) on discounting historical data for binary data via logistic regression (*e.g.*, see Ibrahim and Chen, 2002, for the power prior). Except for Wisniowski *et al.* (2020), who used survey weights as a covariate which is a bit controversial, the others did not do so. Finite population parameter estimation was not done, and inference was not required about small areas.

Chen *et al.* (2020), henceforth sometimes CLW, studied Scenario 1 with propensity scores. They presented a doubly robust inference with non-probability samples mainly within the design-based approach. They estimated propensity scores via logistic regression (parametric modeling for the participation variable) using the probability sample without

Table 1: Data scenarios with massively missing data

	nps (1)			ps (2)			
Scenarios	w	$\underline{x}$	y	w	$\underline{x}$	y	Action
1	–	+	+	+	+	–	Supplement nps by ps
2 a	–	+	+	+	+	+	Supplement ps by nps
2 b	–	+	+	+	+	+	Data integration
3	–	+	–	+	–	+	Supplement ps by nps

NOTE: Here "+" means observed and "–" means missing.

study variable, and inverse probability weighting (IPW) for the finite population mean.

Based on the method of Chen, Li and Wu (2020) to estimate propensity scores, Nandram and Rao (2021, 2023) studied Scenario 2b. They argued that the nps should be used to construct the prior for the parameters (the study variable is observed in both nps and ps), a less practical situation. In addition, thinking of the nps as historical data, the prior of the parameters, obtained from it, should be partially discounted. For Bayesians, normalized weighted density should be used to model the study variable. Nandram, Choi and Liu (2021) discussed some mixed analyzes.

Rafei, Elliot and Flannagan (2022) and Marellla (2023) used Scenario 2b. Rafei *et al.* (2022) used Bayesian additive regression trees (BART), which has its own problems. Marella (2023) used empirical likelihood, and the study variable for the ps is manufactured and assumed real.

There has also been some activities in small area estimation for non-probability samples. Beaumont (2020) used Scenario 2a and Beaumont and Rao (2021) used covariates from nps in ps (*e.g.*, Fay-Herriot model) in Scenario 3; Nandram and Rao (2024) used Scenario 2a with unit level data and used the method of CLW to estimate propensity scores for the nps; they also discussed Scenario 2b. Rao (2021) discussed many scenarios and how the ideas of probability samples can be used to study non-probability samples. See Elliott and Valliant (2017) for an earlier review, where they discussed quasi-randomization, used in CLW and others.

Following CLW, in our current work under Scenario (1) without using propensity scores, we do not have a participation model for the selection mechanism, and the non-Bayesian notion of double robustness is null and void; in our case only the model for the study variable (response) is needed. However, this situation may not be fully practical because even though a nps is available, we may still need to plan and field a small probability sample (an unnecessary burden). Instead it is possible to obtain the necessary information in our set up using web-scraping. For example, it is possible to obtain population sizes and

population total covariates from the web for many variables.

In this paper, we have a nps (1) and a ps (2) from California (NHANES III). Body mass index (BMI) is the study variable, and for adults the normal range of BMI is $[20, 25]$; see Nandram and Choi (2010) for more details about the survey design and a discussion about a much larger BMI data set. The population size is about 5000 times the sample size (.02% sampling). The nps has about 80% of the data and the ps has about 20%. Our data are of the form, $(W_{sj}, \underline{x}_{sj}, y_{sj}), s = 1, 2, j = 1, \ldots, n_s$, where W_{sj} are original survey weights, $\underline{x}_{sj}$, a p-vector of covariates including an intercept, and y_{sj} is the study variable; but W_{1j} and y_{2j} are unknown. It is worth noting some of the features of the data:

a. There are three covaariates :Age (20-90 years' old), race (white, non-white), sex (male, female) are covariates but interactions are not significant;

b. The data are partitioned into eight (8) counties (small areas), an area in the ps may not have data, and interest is on the finite population mean of each area (county);

c. The population is stratified and the data from the nps and ps are allocated to 56 strata (areas) with some discretization: age (20-24, 25-29,. . . , 85-90), race $(0, 1)$, sex $(0, 1)$, where other covariates are redacted or missing, and the partition of age is normally used at the National Center for Health Statistics;

d. The study variable (response) and covariates in the nps, $(\underline{x}, y)$, provide the spatial relation (both $\underline{x}$ and $\underline{y}$ are used);

e. A double mass imputation is used to get the survey weights in the nps (W, $\underline{x}$ and y are used), and the weights are trimmed, calibrated to the population size, and adjusted to the effective sample size.

Our procedure differs from all others in the literature. We stratify the population into distinct values of the covariates in the nps, each distinct covariate vector represents a stratum. Note that these strata are obtained using basic knowledge about the population. We may need to discretize some covariates; in survey sampling many of the covariates are usually discrete.

In our example on BMI, age, race and sex are covariates; age (20-90 years old), race (white, non-white) and sex (male, female). So there are $71 \times 2 \times 2 = 484$ possible strata. The nps and ps data are then allocated to the strata, of course, after the sample data are observed. Then some strata will be empty and some discretization and regular imputation needs to be done. The covariates and study variable from the nps will be used to obtain a spatial structure among the strata via an incidence matrix, V. Then, we construct Table 2 for BMI data with some discretization on age to get 56 strata and 8 counties to avoid sparseness and empty strata. These are the data we actually analyze. Note that an area may have data in different strata, and each stratum will have a different parameter in our models. That is, a small area (county) may have several parameters associated with it, and this helps to mitigate over shrinkage, which is very common in small area estimation. In Section 3, we will describe how to prepare Table 2 because our new Bayesian analysis is based on it.

As a summary, we are using the nps data to make inference about the finite population mean of each small area (county for the BMI data). We have data from the nps and some information from the ps that we used to supplement the nps. This paper has six sections, including the current one. In Section 2, we review the main ideas, relevant to our current work, Nandram and Rao (2021, 2023) and Chen *et al.* (2020). In Section 3, we show how to prepare the data for a robust analysis. Specifically, we discuss stratification of the population by distinct covariate values, mass imputation to construct the survey weights for the nps, and a spatial structure, which we use to replace any functional relation between the study variable and the covariates, and Bayesian predictive inference for the finite population mean of each area (county); see Table 2. In Section 4, we discuss the hierarchical Bayesian models, where robustness is mostly based on a two-component mixture model. These models are primarily based on the Scott-Smith model; see Scott and Smith (1969) and Nandram *et al.* (2011). This section also has detailed discussion of an example of body mass index (BMI) in parallel to the rest of the technical discussion. In Section 5, we discuss some improvements and possible extensions. In Section 6, we present concluding remarks.

Table 2: Structurally complete BMI data for nps with $G = 56$ strata and $\ell = 8$ counties

Stratum	nps	Size
$\underline{x}_1$	$(W_{1j}, y_{1j}), j = 1, \ldots, n_1$	N_1
.	.	.
.	.	.
$\underline{x}_g$	$(W_{gj}, y_{gj}), j = 1, \ldots, n_g$	N_g
.	.	.
.	.	.
$\underline{x}_G$	$(W_{Gj}, y_{Gj}), j = 1, \ldots, n_G$	N_G

NOTE: There are G distinct values of $\underline{x}_g, g = 1, \ldots, G$; $G = 56$ for the BMI data. In addition, for the spatial analysis, we have the **incidence matrix**, V, among the strata. Weights may be equal or unequal and adjusted weights are used in the sampling process (study variable). An actual area (county) may be represented in several strata. The normal range of BMI for adults should be $[20, 25]$.

2. Review and background information

In this section, a review of Nandram and Rao (2021, 2023, 2024) is presented. We also present the method of Chen, Li and Wu (2020) to construct the propensity scores, and

some general comments are made about this method for propensity scores to highlight and justify our new approach.

2.1. Review of Nandram and Rao (2021, 2023, 2024)

First, we describe the normalized weighted density. Nandram and Rao (2021, 2023, 2024) considered a single population, not sub-populations (*e.g.,* small areas), and they showed how to make inference for finite population mean. They assumed that sample data are available from a nps (1) and a ps (2).

We recall here that the original survey weights are $W_{si}, i = 1, \ldots, n_s, s = 1, 2$, and W_{si} is the number of units the i^{th} sampled individual represents in the finite population of size, N. Since we assume that both samples are drawn from the same population, $\sum_{i=1}^{n_s} W_{si} = N, s = 1, 2$.

Nandram and Rao (2021, 2023) defined adjusted weights,

$$w_{si} = \hat{n}_s \frac{W_{si}}{\sum_{j=1}^{n_s} W_{sj}}, i = 1, \ldots, n_s, s = 1, 2, \tag{1}$$

where the y_{si} are assumed to be independent, and the effective sample size is

$$\hat{n}_s = \frac{(\sum_{j=1}^{n_s} W_{sj})^2}{\sum_{j=1}^{n_s} W_{sj}^2}.$$

The adjusted weights in (1) is needed to construct the likelihood functions. They also assumed an estimator of the population total, based on the ps, is $\hat{N} = \sum_{i=1}^{n_2} W_{2i}$. This is actually a Horvitz-Thompson estimator, but this interpretation is not necessary. Also, a Horvitz-Thompson estimator of the total covariates, based on the ps, is

$$\widehat{\sum_{i=1}^{N} \underline{x}_{2i}} = \sum_{i=1}^{n_2} W_{2i} \underline{x}_{2i}.$$

Of course, these two estimators are based on inverse probability weighting (IPW), where the first one being natural.

In their approach, they assumed the population model, $f(y \mid \underline{x}, \underline{\theta})$, is correct, and so robustness is a serious consideration. The participation model (for selection indicators) must also be robust; in our new approach there is no participation model, a huge gain. For the sample distribution, they used the weighted density,

$$f(y \mid \underline{x}, \underline{\theta}, \underline{w}) = \frac{\{f(y \mid \underline{x}, \underline{\theta})\}^w}{\int \{f(y \mid \underline{x}, \underline{\theta})\}^w dy\}}. \tag{2}$$

In the Bayesian paradigm, the sampling process must also be proper. Because the normalization constant may be a function of $\underline{\theta}$, it must be included, and therefore, the general specification in (2) is needed.

Next, we describe Bayesian predictive inference. We have estimated the survey weights; we use Chen, Li and Wu (2020) to get nps weights. For the nps and ps we now have

$$(W_{si}, \underline{x}_{si}, y_{si}), i = 1, \ldots, \ell, s = 1, 2.$$

Therefore, the population model is

$$y_i \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_i'\underline{\beta}, \sigma^2), i = 1, \ldots, N.$$

The sample model is

$$f(y_i \mid \underline{\beta}, \sigma^2) \propto \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y_i - \underline{x}_i'\underline{\beta})^2\}\right]^{w_i}$$

and normalizing, we have

$$y_{1i} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_{1i}'\underline{\beta}, \frac{\sigma^2}{w_{1i}}), i = 1, \ldots, n.$$

The prior distribution is

$$\pi(\underline{\beta}, \sigma^2) \propto \sigma^{-2}.$$

Once the sample model is fit, we use the ps to guess the population size and covariate total and we use surrogate sampling (Nandram, 2007) to sample population (prediction).

The finite population mean is

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

and

$$\pi(\bar{Y} \mid \underline{y}_s) = \int f(\bar{Y} \mid \underline{\beta}, \sigma^2)\pi(\underline{\beta}, \sigma^2 \mid \underline{y}_s)\, d\underline{\beta}\, d\sigma^2.$$

For Bayesian predictive inference, Nandram and Rao (2021, 2023, 2024) used surrogate sampling,

$$\bar{Y} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{i=1}^{n_2} W_{2i}x_{2i}'}{\sum_{i=1}^{n_2} W_{2i}}\underline{\beta}, \frac{\sigma^2}{\sum_{i=1}^{n_2} W_{2i}}\right).$$

We do not split the finite population as $\bar{Y} = f\bar{y}_s + (1-f)\bar{y}_{ns}$, where $f = \frac{n}{N}$ is the sample fraction, $\bar{y}_s$ is the sample mean and $\bar{y}_{ns}$ is the nonsample mean, because both $\bar{y}_s$ and $\bar{y}_{ns}$ are corrupted (biased). We will use a similar procedure in our new method.

To perform Bayesian predictive inference for small areas, Nandram and Rao (2024) assumed

$$(W_{sij}, \underline{x}_{sij}, y_{sij}), j = 1, \ldots, n_{si}, i = 1, \ldots, \ell, s = 1, 2,$$

are available,

$$y_{ij} \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_{ij}'\underline{\beta} + \nu_i, \sigma^2), j = 1, \ldots, N_i, i = 1, \ldots, \ell$$

where $\Omega = (\underline{\nu}, \underline{\beta}, \sigma^2)$. We note that the population is too large to sample completely, population sizes and covariates are unknown. We use surrogate sampling (Nandram, 2007) again.

The finite population means are

$$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}, \quad \pi(\bar{Y}_i \mid \underline{y}_s) = \int f(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2)\pi(\nu_i, \underline{\beta}, \sigma^2, \mid \underline{y}_s)\, d\nu_i\, d\underline{\beta}\, d\sigma^2, i = 1, \ldots, \ell.$$

Again, Nandram and Rao (2024) used surrogate sampling for small areas.

1. For $i = 1, \ldots, \ell$,

$$\bar{Y}_i \mid \Omega \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{j=1}^{n_{2i}} W_{2ij} x'_{2ij}}{\sum_{j=1}^{n_{2i}} W_{2ij}} \underline{\beta} + \nu_i, \frac{\sigma^2}{\sum_{j=1}^{n_{2i}} W_{2ij}}\right); \bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s$$

2. For the h^{th} iterate from the Gibbs sampler,

$$\bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s \overset{ind}{\sim} \text{Normal}\left(\frac{\sum_{j=1}^{n_{2i}} W_{2ij} x'_{2ij}}{\sum_{j=1}^{n_{2i}} W_{2ij}} \underline{\beta}^{(h)} + \nu_i^{(h)}, \frac{\sigma^{2(h)}}{\sum_{j=1}^{n_{2i}} W_{2ij}}\right).$$

We note that $\sum_{j=1}^{n_{2i}} W_{2ij}$ is very large, and in this case the variance can be very small with $\bar{Y}_i^{(h)} \mid \Omega^{(h)}, \underline{y}_s$ essentially a point mass.

2.2. Propensity scores (Chen, Li and Wu, 2020)

We review the method of Chen *et al.* (2020), recall CLW, and we make some commentaries about their method; see also Wu (2022).

Denote the common covariates by $\underline{z}_i, i = 1, \ldots, N$, with nps, $\underline{z}_{1i}, i = 1, \ldots, n_1$, and ps, $\underline{z}_{2i}, i = 1, \ldots, n_2$. Again we note that nonsampled covariates are unknown.

Let $R_i, i = 1, \ldots, N$, where $R_i = 1$ if unit i is sampled and $R_i = 0$ if unit i is not sampled. With $r_i = 0, 1$, they used the parametric assumption,

$$\pi_i = P(R_i = 1 \mid \underline{z}_i) = \pi(\underline{z}_i; \underline{\theta}),$$

$$p(\underline{r} \mid \underline{\theta}) = \prod_{i=1}^{N} \{\pi(\underline{z}_i; \underline{\theta})\}^{r_i} \{1 - \pi(\underline{z}_i; \underline{\theta}\}^{1-r_i}$$

$$= \prod_{i=1}^{n} \frac{\pi(\underline{z}_i; \underline{\theta})}{1 - \pi(\underline{z}_i; \underline{\theta})} \prod_{i=1}^{N} \{1 - \pi(\underline{z}_i; \underline{\theta}\}$$

with independence over i.

Then, the population log-likelihood is

$$\ell(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{z}_{1i}; \underline{\theta})}{1 - \pi(\underline{z}_{1i}; \underline{\theta})}\right\} + \sum_{i=1}^{N} \log\{1 - \pi(\underline{z}_i; \underline{\theta})\},$$

and the pseudo-log-likelihood is

$$\ell_1(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{z}_{1i}; \underline{\theta})}{1 - \pi(\underline{z}_{1i}; \underline{\theta})}\right\} + \sum_{i=1}^{n_2} W_{2i} \log\{1 - \pi(\underline{z}_{2i}; \underline{\theta})\}.$$

The propensity scores for the nps are then $\pi(\underline{z}_{1i}; \hat{\underline{\theta}}), i = 1, \ldots, n_1$, where $\hat{\underline{\theta}}$ is the MLE of $\underline{\theta}$

Now, we make some comments about the CLW Approach.

1. Horvitz-Thompson estimator may be sub-optimal with survey weights because the ratio of the study variable and the selection probabilities may not be close to a constant;

2. Propensity scores depend only on $\underline{z}$ (common variables only), and there may be other important variables;

3. Propensity scores are not selection probabilities because the entire population should be taken into consideration, and a proper quasi-randomization (see Elliot and Valiant, 2017) cannot be executed;

4. Logistic regression is not robust against its assumptions (*e.g.*, the linearity assumption on logit);

5. They assume ignorable selection, but nonignorable selection is preferred;

6. There are difficulties in optimization (convergence) especially for small samples;

7. In the Bayesian paradigm, the uncertainty in the estimation of the propensity scores must be taken into consideration, and this is a difficult problem, but a bootstrap procedure shows a 50% increase in standard deviation;

8. Need more robust participation models, perhaps a mixture of several link functions (*e.g.*, t_4, t_8, *etc.*).

Our new method has the following features: (a) We avoid using propensity scores because they are not selection probabilities in the CLW method; (b) We avoid direct link between the study variable and covariates (robustness, spatial model); (c) We do not need non-sample covariates for prediction; (d) We avoid using the Horvitz-Thompson estimator for prediction because it is sub-optimal with survey weights; (e) We consider a robust population model in which a two-component mixture model is used to accommodate outliers and non-normality.

We are particularly interested in inference about the finite population mean of each small area. The small areas (counties) are not modeled in our procedure, but rather the strata are modeled as small areas. Therefore, the phrase "small areas" is used in two ways, one for the constructed strata and one for counties. Another approach for non-probability samples with small areas is given by Nandram and Rao (2023b), where the actual small areas are modeled directly, and inference is about finite population means and percentiles (*e.g.*, 85^{th} and 95^{th} percentiles for BMI are useful). Our new approach avoids many difficult problems, except one of them is to incorporate the uncertainty in the estimated propensity scores, but we are still working on this problem.

3. Data preparation

In this section, we show how to prepare the data to construct our procedures and methods. We show how to obtain survey weights in the nps, form the strata, allocate the sample units to the strata, and obtain the the neighborhood structure among the strata. As we stated already, the strata are formed by distinct covariates. This is done for the BMI data; see Table 2 for the data we analyze in this paper.

We note the following steps in our procedure.

a. A few strata may be empty, and we can donate one unit from a stratum with at least two units (with all variables) to a "nearby" empty stratum separately for the nps (1) and the ps (2). For the nps, there was one empty stratum and for the ps there were three empty strata when the BMI data is processed.

b. The $W_{1ij}, j = 1, \ldots, n_{1i}$, are unknown, but it is true that $\sum_j^{n_{1i}} W_{1ij} = \sum_j^{n_{2i}} W_{2ij} = N_i, i = 1, \ldots, G$.

c. Assume all W_{1ij} are the same with the same $\underline{x}_i$ (i^{th} stratum); under simple random sampling (SRS) without replacement $W_{1ij} = N_i/n_{1i}$ (multiple-level regression and post-stratification (MRP) uses equal weights).

d. We have also obtained unequal weights, $W_{1ij}, j = 1, \ldots, n_{1i}$, and this can be accomplished using a double (reverse) mass imputation.

e. All weights are trimmed to mitigate the effects of outlying weights. Outliers decrease the effective sample size. Trimmed weights are calibrated to the population size and adjusted to the effective sample size for modeling, but too much trimming can lead to a false sense of security (decreased variance but increased bias).

We use double mass imputation to provide a structurally complete nps data set; see Kim *et al.* 2021. In mass imputation with two data sets, an entire variable may be missing from one data set, and both data sets are used to impute the missing data. For nps, we have $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, and for the ps we have $(W_{2i}, \underline{x}_i), i = 1, \ldots, n_2$ (no intercept). The procedure is straightforward:

a. Use $(\underline{x}_{1i}, y_{1i}), i = 1, \ldots, n_1$, to fill in $y_{2i}, i = 1, \ldots, n_2$. This is done using the Mahalanobis distance among the $\underline{x}_{1i}$ and $\underline{x}_{2i}$ via nearest neighbors.

b. Stack the y_{1i} under the $\underline{x}_{1i}$ to create a new vector, $\underline{x}^*_{1i}$. Similarly, stack y_{2i} under the $\underline{x}_{2i}$ to create a new vector, $\underline{x}^*_{2i}$.

c. Use $(W_{2i}, \underline{x}^*_{2i}), i = 1, \ldots, n_2$, to fill in $W_{1i}, i = 1, \ldots, n_1$. This is done using the Mahalanobis distance among the $\underline{x}^*_{1i}$ and $\underline{x}^*_{2i}$ via nearest neighbor again.

We compare two situations, which are equal weights within strata (different strata can have different weights) and unequal weights within strata (small areas can have different weights). Actually the situation of equal weights correspond to ignorable selection and the situation of unequal weights corresponds to the situation of non-ignorable selection. See Nandram and Choi (2010) and more recently Nandram (2022).

Next, we show how to use the nps to get the spatial structure, which is used as a surrogate of the covariates to avoid the specification of any functional relation between the study variable and the covariates. Recall robustness is very important because the population model is assumed to be correct.

We follow the steps below:

a. Use a surrogate for any functional relation between the study variable and the covariates (Lockwood 2023, PhD Dissertation). The spatial approach can provide this surrogate.

b. Use ordinary least squares to find the $G \times G$ incidence matrix, V, which has zeros every where except when two strata are neighbors. We use a distribution-free procedure, where we assume that

$$y_{gj} = \underline{x}_g' \underline{\beta} + \nu_g, j = 1, \ldots, n_g,$$
$$\hat{\nu}_g = \bar{y}_g - \underline{x}_g' \hat{\underline{\beta}}, g = 1, \ldots, G,$$
$$\bar{y}_g = n_g^{-1} \sum_{j=1}^{n_g} y_{gj}, \ \hat{\underline{\beta}} = \left(\sum_{g=1}^{G} n_g \underline{x}_g \underline{x}_g' \right)^{-1} \sum_{g=1}^{G} n_g \underline{x}_g \bar{y}_g.$$

The neighbors of stratum g are $\mathcal{N}_g = \{h : |\hat{\nu}_g - \hat{\nu}_h| < t_o\}, h = 1, \ldots, G$.

c. Choose t_o to make the Moran's I correlation coefficient strong.

d. Once the strata are obtained, we never need the covariates again. Specifically, we do not need to estimate the nonsampled covariates, a huge saving, and the uncertain Horvitz-Thompson estimator of the population total covariate is not needed.

We note that there are difficulties using the Mahalanobis distance to find the incidence matrix. The distinct covariates should not be used, and there is virtually no control over this procedure.

Finally, we show how to do Bayesian predictive inference for the finite population means of the small areas using surrogate sampling as described in the review. Again, the population model is

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}(\mu_g, \sigma^2), j = 1, \ldots, N_g, g = 1, \ldots, G.$$

Let $\mathcal{C}_{ig}, g = 1, \ldots, G, i = 1, \ldots, \ell$, denote the set of sampled units of the i^{th} area in g^{th} stratum. Define the sum of the weights associated with i^{th} area in the g^{th} stratum as

$$N_{ig} = \sum_{j \in \mathcal{C}_{ig}} W_{gj}.$$

Note that many of these N_{ig} are zeros. Then, for $N_{ig} > 0$,

$$\bar{Y}_{ig} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\mu_g, \frac{\sigma^2}{N_{ig}}\}$$

and

$$\bar{Y}_i = \frac{\sum_{g=1}^{G} N_{ig} \bar{Y}_{ig}}{\sum_{g=1}^{G} N_{ig}}, i = 1, \ldots, \ell.$$

For posterior inference, we have $(\mu_g^{(h)}, \sigma^{2(h)}), h = 1, \ldots, M, g = 1, \ldots, G$, from the sample model. Then, we have $\bar{Y}_i^{(h)}, h = 1, \ldots, M, i = 1, \ldots, \ell$.

4. Hierarchical Bayesian models and a numerical example

This is the main section of the paper in which we present the hierarchical Bayesian models to analyze the new data. In parallel, we present our numerical example on body mass index (BMI) data.

First, we fit a small area model with covariates. This is the basic model of Battese *et al.* (1988); henceforth the BHF model. A full Bayesian approach of the BHF model is given by Toto and Nandram (2011) and Molina *et al.* (2014). However, this model is notoriously non-robust to its assumptions (normality, linearity, outliers). We use the BHF model as the baseline model for comparison.

Second, we use the Scott-Smith model (Scott and Smith, 1969), which does not have covariates; this model was intended for cluster sampling, but nowadays we have been using it for small area estimation; see Nandram *et al.* (2011), henceforth SS model. Note that the SS model is a special case of the BHF model. For example, we will construct a spatial model to accommodate the covariates and we will add robustness through mostly a two-component mixture model (Chakraborty *et al.*, 2019, Goyal *et al.*, 2020) and one slightly different model with a stick-breaking prior (Ishwaran and James, 2001). So we are using the SS model to drop the linearity assumption of the study variable and the covariates, to robustify the assumption of normality, and accommodating outliers. Note all SS models have a spatial component to accommodate the covariates.

We will compare the models with (a) equal weights and (b) unequal weights using posterior inference about the finite population means by actual small areas (counties in the application on BMI). For comparison, we use posterior mean (PM), posterior standard deviation (PSD), numerical standard error (NSE), posterior coefficient of variation (PCV) and 95% highest posterior density interval (HPDI). All our data analyses will be done on the BMI data.

All computations were done on WPI's Solar Cluster. The Gibbs sampler was used for almost all models. The computations used more time (increasing complex models) as we go from Table 3 to Table 7 because longer runs are needed to ensure strong mixing in the Gibbs samplers. Apart from the data preparation, which took just a few minutes, the computational time for all models (equal & unequal weights) took just about one hour.

Baseline (BHF) model

In the BHF model we use the original 8 counties. The stratum table with the weights is reverted to the small areas (not strata). We note that BHF model is notoriously not robust to its assumptions.

The population model is

$$y_{ij} \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, \sigma^2\}, j = 1, \ldots, N_i, i = 1, \ldots, \ell.$$

The sample model is

$$y_{ij} \mid \nu_i, \underline{\beta} \overset{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{ij}}\}, j = 1, \ldots, n_i,$$

$$\nu_i \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(0, \frac{\rho}{1-\rho}\sigma^2), i = 1, \ldots, \ell,$$

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}. \tag{3}$$

The joint posterior density for the sample model described in (3) is proper if the design matrix is full rank and $\ell \geq 2$. It can be sampled using a random sampler; a noisy Gibbs sampler is not necessary.

Bayesian predictive inference is now standard using surrogate sampling,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\underline{\bar{X}}_i'\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}\},$$

$$\pi(\bar{Y}_i \mid \underline{y}_s) = \int f(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2)\pi(\nu_i, \underline{\beta}, \sigma^2, \mid \underline{y}_s)d\nu_i \; d\underline{\beta} \; d\sigma^2, i = 1, \ldots, \ell.$$

In Table 3 we compare the BHF models, with equal weights and unequal weights. We observe that the PMs are rougher, PSDs are larger and PMs are closer to 25 (but not as close as we want) under the model with unequal weights than the model with equal weights. Note that counties 4 & 8 are mostly different from the others. But can we do better on these three measures?

SS model with spatial effects

We next delete the covariates and replace them with the spatial covariance matrix. This is the SS model with spatial effects.

Now the population model is

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\{\mu_g, \sigma^2\}, j = 1, \ldots, N_g, g = 1, \ldots, G.$$

The sample model is

$$y_{gj} \mid \mu_g, \rho \overset{ind}{\sim} \text{Normal}\{\mu_g, \frac{\sigma^2}{w_{gj}}\}, j = 1, \ldots, n_g,$$

$$\mu_g \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(\theta, \frac{\rho}{1-\rho}\sigma^2), g = 1, \ldots, G,$$

$$\pi(\theta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}. \tag{4}$$

The model, described in (4) and the few lines above it, is the basic SS model, which can be fit using a random sampler.

However, we need to add the spatial structure and we use the simultaneous conditional auto-regressive (CAR) model (*e.g.*, He and Sun, 2000; Chung and Datta, 2022). The sample model is now given by

$$y_{gj} \mid \mu_g, \sigma^2 \overset{ind}{\sim} \text{Normal}\left\{\mu_g, \frac{\sigma^2}{w_{gj}}\right\}, j = 1, \ldots, n_g, g = 1, \ldots, G,$$

Table 3: BHF model is fit directly to the small areas

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.629	0.365	0.004	0.013	(26.905, 28.351)
2	138	28.277	28.025	0.398	0.004	0.014	(27.293, 28.847)
3	667	27.340	27.629	0.192	0.002	0.007	(27.248, 27.995)
4	133	25.980	27.051	0.405	0.004	0.015	(26.221, 27.814)
5	96	27.075	27.547	0.406	0.004	0.015	(26.767, 28.368)
6	119	27.313	27.318	0.395	0.004	0.014	(26.538, 28.108)
7	100	27.518	27.881	0.449	0.005	0.016	(27.002, 28.777)
8	137	26.698	27.470	0.370	0.004	0.013	(26.741, 28.205)
b. Unequal weights							
1	140	27.324	27.234	0.440	0.004	0.016	(26.359, 28.085)
2	138	28.277	28.594	0.436	0.005	0.015	(27.728, 29.442)
3	667	27.340	27.193	0.197	0.002	0.007	(26.821, 27.590)
4	133	25.980	25.901	0.463	0.005	0.018	(25.002, 26.794)
5	96	27.075	27.855	0.497	0.005	0.018	(26.838, 28.790)
6	119	27.313	27.372	0.457	0.005	0.017	(26.489, 28.302)
7	100	27.518	27.394	0.514	0.005	0.019	(26.443, 28.439)
8	137	26.698	26.266	0.460	0.005	0.017	(25.355, 27.136)

NOTE: $\bar{y}_1 = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}, i = 1, \ldots, \ell$, is the ordinary sample average. Posterior inference is based on 1,000 iterates that provide posterior mean (PM), posterior standard deviation (PSD), numerical standard error (NSE), posterior coefficient of variation (PCV), and 95% highest posterior density interval (HPDI). Only random samplers, no Gibbs samplers, are used.

$$\underline{\mu} \mid \theta, \rho, \psi, \sigma^2 \sim \text{Normal}\left\{\underline{j}\theta, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\right\},$$

($\underline{j}$ is a vector of ones),

$$\pi(\theta, \rho, \psi, \sigma^2) \propto \frac{1}{\sigma^2},$$

$$0 < \rho < 1, \frac{1}{\lambda_1} < \psi < \frac{1}{\lambda_G},$$

where $\lambda_1 < \ldots < \lambda_G$, are eigenvalues of $R^{-1}V$, and $R = \text{diagonal}(r_g, g = 1, \ldots, G)$ with r_g the g^{th} row (column) sum of V. We note that the joint posterior density is proper and it can be fit using the Gibbs sampler. This is our first SS model.

In Table 4 we notice that the PSDs for some of the areas are smaller under unequal weights (not desirable). The PMs are smoother under the unequal weights, but closer to 25.

We need to improve this model, which can be done by robustification of either the sampling process or the area means or both. All these models require the use of the Gibbs sampler. Details of number of iterations used are described in the notes of the tables. To come up with those numbers, we have used the Geweke test of stationarity and the effective sample, and in all cases the Gibbs sampler mixed strongly.

Table 4: Scott-Smith model is fit to the strata with spatial effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.197	0.314	0.009	0.012	(26.638, 27.832)
2	138	28.277	26.741	0.258	0.008	0.010	(26.233, 27.276)
3	667	27.340	27.104	0.204	0.007	0.008	(26.709, 27.485)
4	133	25.980	27.050	0.269	0.010	0.010	(26.602, 27.591)
5	96	27.075	26.829	0.300	0.009	0.011	(26.256, 27.435)
6	119	27.313	27.082	0.356	0.012	0.013	(26.378, 27.812)
7	100	27.518	27.125	0.411	0.013	0.015	(26.358, 27.867)
8	137	26.698	27.076	0.279	0.008	0.010	(26.592, 27.632)
b. Unequal weights							
1	140	27.324	26.640	0.299	0.008	0.011	(26.113, 27.229)
2	138	28.277	26.730	0.253	0.006	0.009	(26.232, 27.208)
3	667	27.340	26.546	0.178	0.005	0.007	(26.225, 26.895)
4	133	25.980	26.687	0.424	0.016	0.016	(25.831, 27.375)
5	96	27.075	26.149	0.310	0.008	0.012	(25.575, 26.758)
6	119	27.313	26.412	0.293	0.010	0.011	(25.779, 26.954)
7	100	27.518	25.877	0.372	0.012	0.014	(25.257, 26.632)
8	137	26.698	25.951	0.289	0.010	0.011	(25.399, 26.538)

NOTE: The Gibbs sampler is run $11,000$ times with a "burn-in" of $1,000$ and a systematic sample of every tenth is taken.

SS model with spatial effects and robust study variable

The population model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma < 1$.

The sample model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}),$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$

$$\underline{\mu} \overset{ind}{\sim} \text{Normal}\{\theta \underline{j}, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\},$$

$$\pi(\theta, \sigma^2, p, \rho, \gamma, \psi) \propto \frac{1}{\sigma^2}, \frac{1}{\lambda_1} < \psi < \frac{1}{\lambda_G}.$$

The joint posterior density is

$$\pi(\underline{z}, \underline{\mu}, \theta, \sigma^2, \gamma, \rho, \psi \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2}\prod_{g=1}^{G}\prod_{j=1}^{n_g}[(1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}})]^{1-z_{gj}}[p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}})]^{z_{gj}}$$

$$\times\text{Normal}\{\theta \underline{j}, \frac{\rho}{1-\rho}\sigma^2(R - \psi V)^{-1}\},$$

and this posterior can be sampled using the Gibbs sampler; see note to Table 5. The results are better than those in BHF model and the SS model with only spatial effects.

SS Model with spatial effects, robustness on study variable and random effects

The population model is now

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma_0 < 1$.

The sample model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{5}$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$,

$$\mu_g \overset{ind}{\sim} (1-q)\text{Normal}(\nu_g, \gamma_1\frac{\rho_1}{1-\rho_1}\sigma^2) + q\text{Normal}(\nu_g, \frac{\rho_1}{1-\rho_1}\sigma^2), \tag{6}$$

$0 < q < 1/2, 0 < \gamma_1 < 1$,

$$\underline{\nu} \sim \text{Normal}\{\theta \underline{j}, \frac{\rho_2}{1-\rho_2}\sigma^2(R - \psi V)^{-1}\},$$

$$\pi(\theta, \sigma^2, p, q, \rho_1, \rho_2, \psi, \gamma_0, \gamma_1) \propto \frac{1}{\sigma^2}.$$

The assumptions in (5) and (6) express a form of Bayesian double robustness.

The joint posterior density is

$$\pi(\underline{z}, \underline{t}, \underline{\mu}, \underline{\nu}, \theta, \sigma^2, p, q, \rho_1, \rho_2, \psi, \gamma_0, \gamma_1 \mid \underline{y}) \propto$$

Table 5: Scott-Smith model is fit to the strata with spatial effects and robust study variable

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.257	0.309	0.008	0.011	(26.701, 27.839)
2	138	28.277	27.100	0.248	0.010	0.009	(26.579, 27.541)
3	667	27.340	27.315	0.202	0.006	0.007	(26.969, 27.667)
4	133	25.980	27.206	0.252	0.007	0.009	(26.715, 27.674)
5	96	27.075	26.956	0.358	0.011	0.013	(26.341, 27.538)
6	119	27.313	27.286	0.233	0.008	0.009	(26.821, 27.712)
7	100	27.518	27.040	0.425	0.014	0.016	(26.251, 27.731)
8	137	26.698	27.235	0.238	0.008	0.009	(26.838, 27.628)
b. Unequal weights							
1	140	27.324	27.021	0.345	0.011	0.013	(26.349, 27.668)
2	138	28.277	27.084	0.311	0.009	0.011	(26.464, 27.584)
3	667	27.340	26.835	0.186	0.005	0.007	(26.493, 27.203)
4	133	25.980	26.752	0.381	0.011	0.014	(26.095, 27.410)
5	96	27.075	26.503	0.367	0.011	0.014	(25.784, 27.121)
6	119	27.313	27.003	0.341	0.011	0.013	(26.340, 27.619)
7	100	27.518	26.163	0.457	0.016	0.017	(25.311, 26.969)
8	137	26.698	26.417	0.281	0.010	0.011	(25.823, 26.902)

NOTE: The Gibbs sampler is run 40,000 times with a "burn-in" of 10,000 and a systematic sample of every thirtieth.

$$\frac{1}{\sigma^2}\prod_{g=1}^{G}\prod_{j=1}^{n_g}[(1-p)\text{Normal}_{y_{gj}}(\mu_g,\gamma_0\frac{\sigma^2}{w_{gj}})]^{1-z_{gj}}[p\text{Normal}_{y_{gj}}(\mu_g,\frac{\sigma^2}{w_{gj}})]^{z_{gj}}$$
$$\times\prod_{g=1}^{G}[(1-q)\text{Normal}_{\mu_g}(\nu_g,\gamma_1\frac{\rho_1}{1-\rho_1}\sigma^2)]^{1-t_g}[q\text{Normal}_{\mu_g}(\nu_g,\frac{\rho_1}{1-\rho_1}\sigma^2)]^{t_g}$$
$$\times\left(\frac{1-\rho_2}{\rho_2\sigma^2}\right)^{G/2}|R-\psi V|^{1/2}\exp\left\{-\frac{1-\rho_2}{2\rho_2\sigma^2}(\underline{\nu}-\theta\underline{j})'(R-\psi V)(\underline{\nu}-\theta\underline{j})\right\},$$

and this can be sampled using the Gibbs sampler; see note to Table 6. Again the results look better than the previous ones. This model appears to be the best: The PSDs for unequal weights are larger than those for equal weights, and the PMs for unequal weights are smaller than those for equal weights (therefore closer 25).

Table 6: Scott-Smith model is fit to the strata with spatial effects and robust study variable and robust random effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	26.826	0.268	0.008	0.010	(26.283, 27.321)
2	138	28.277	26.798	0.270	0.007	0.010	(26.305, 27.324)
3	667	27.340	27.096	0.221	0.006	0.008	(26.679, 27.523)
4	133	25.980	27.166	0.198	0.006	0.007	(26.778, 27.550)
5	96	27.075	26.722	0.239	0.007	0.009	(26.280, 27.172)
6	119	27.313	26.997	0.238	0.006	0.009	(26.539, 27.451)
7	100	27.518	26.814	0.227	0.007	0.008	(26.379, 27.238)
8	137	26.698	27.185	0.202	0.006	0.007	(26.815, 27.597)
b. Unequal weights							
1	140	27.324	26.228	0.355	0.010	0.014	(25.479, 26.839)
2	138	28.277	26.523	0.333	0.010	0.013	(25.889, 27.167)
3	667	27.340	26.367	0.270	0.008	0.010	(25.851, 26.899)
4	133	25.980	26.460	0.236	0.007	0.009	(26.000, 26.915)
5	96	27.075	25.949	0.283	0.008	0.011	(25.389, 26.477)
6	119	27.313	26.432	0.317	0.010	0.012	(25.746, 26.953)
7	100	27.518	25.917	0.274	0.008	0.011	(25.376, 26.412)
8	137	26.698	26.171	0.245	0.008	0.009	(25.709, 26.653)

NOTE: The Gibbs sampler is run $60,000$ times with a "burn-in" of $15,000$ and a systematic sample of every forty-fifth.

SS model with spatial effects, robust study variable and stick-breaking priors on random effects

The population model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\sigma^2) + p\text{Normal}(\mu_g, \sigma^2),$$

$j = 1, \ldots, N_g, g = 1, \ldots, G$, $0 < p < 1/2$ and $0 < \gamma_0 < 1$.

The sampling model is

$$y_{gj} \mid \mu_g \overset{ind}{\sim} (1-p)\text{Normal}(\mu_g, \gamma_0\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{7}$$

$j = 1, \ldots, n_g, g = 1, \ldots, G$,

$$\underline{\mu} \sim \text{Normal}\{\theta\underline{j} + \underline{\eta}, \frac{\rho_1}{1-\rho_1}\sigma^2(R - \psi V)^{-1}\},$$

where $\underline{j}$ is a vector of ones, and the Pitman-Yor two-parameter process is

$$\eta_g \mid \underline{t} \overset{ind}{\sim} \sum_{s=1}^{G_0} p_s \text{Normal}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2), G_0 \leq G, g = 1, \ldots, G, \tag{8}$$

$$p_1 = \nu_1, p_2 = \nu_2(1-\nu_1), \ldots, p_{G_0} = \prod_{s=1}^{G_0-1}(1-\nu_s),$$

$$\nu_s \mid \delta_1, \delta_2 \overset{ind}{\sim} \text{Beta}\{1-\delta_1, \frac{1-\delta_2}{\delta_2} + (s-1)\delta_1\}, s = 1, \ldots, G_0,$$

$$t_s \overset{ind}{\sim} \text{Normal}(0, \frac{\rho_3}{1-\rho_3}\sigma^2), s = 1, \ldots, G_0.$$

The assumptions in (7) and (8) are a form of Bayesian double robustness; these are more flexible than the assumptions in (5) and (6). We have used the prior,

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}, \theta \sim \text{Normal}(\theta_o, \sigma_o^2),$$

where θ_o and σ_o^2 must be specified. Also, for computational stability, we have kept $0 < \delta_1 < \frac{1}{2} < \delta_2 < 1$ with uniform priors on δ_1, δ_2, $\gamma_o, p, \psi, \rho_1, \rho_2, \rho_3$.

In Table 7, this model is similar to the model in which there is robustness on both the study variable and the area effects. But there are some aberrations as some of the PSDs under the stick-breaking prior are smaller with unequal weights, which is a bit concerning.

As a summary, we compare all the models in Table 8. We use the summaries,

$$RD = \{1 - \frac{AVG_{UW} - 25}{AVG_{EW} - 25}\}100\%, \quad RP = \{\frac{GM_{UW}}{GM_{EW}} - 1\}100\%, \tag{9}$$

where AVG_{UW} and AVG_{EW} are the arithmetic means of the 8 PMs for respectively the unequal weights (UW) case and the equal weights (EW) case, and GM_{UW} and GM_{EW} are the geometric means of the 8 PSDs for respectively the unequal weights (UW) case and the equal weights (EW) case. Here RD is the percent AVG_{UW} is closer to 25 than AVG_{EW} (expected to be positive), and RP is the percent increase of GM_{UW} over GM_{EW} (expected to be positive).

Under the RD measure, three models stand out with robust study variable. Under the RP measure, three models stand out. For both measures the model that wins is the model in which study variable and the area effects are both robust; indeed, this is a novel model. While the stick-breaking of the area effects is robust, it is a bit concerning that RP is negative, but it is possible to overcome this problem. Detailed results, like the other models, were not shown for the first and second SS model in Table 8. These results show that robustness on the study variable is important, because for the first SS model, without robustness on the study variable, there are artificially low PSDs relative to the BHF model and the second SS model with robust study variable.

Table 7: Scott-Smith model is fit to the strata with spatial effects and robust study variable and stick-breaking priors on random effects

County	n_1	$\bar{y}_1$	PM	PSD	NSE	PCV	95% HPDI
a. Equal weights							
1	140	27.324	27.286	0.347	0.012	0.013	(26.708, 28.092)
2	138	28.277	26.674	0.330	0.010	0.012	(25.989, 27.256)
3	667	27.340	27.172	0.214	0.006	0.008	(26.762, 27.580)
4	133	25.980	27.105	0.303	0.009	0.011	(26.573, 27.689)
5	96	27.075	26.880	0.367	0.010	0.014	(26.152, 27.533)
6	119	27.313	27.226	0.575	0.021	0.021	(26.194, 28.372)
7	100	27.518	27.174	0.378	0.010	0.014	(26.403, 27.824)
8	137	26.698	27.133	0.311	0.009	0.011	(26.534, 27.739)
b. Unequal weights							
1	140	27.324	26.738	0.347	0.009	0.013	(26.144, 27.448)
2	138	28.277	26.738	0.282	0.008	0.011	(26.193, 27.296)
3	667	27.340	26.581	0.201	0.006	0.008	(26.213, 26.943)
4	133	25.980	26.715	0.424	0.012	0.016	(25.872, 27.429)
5	96	27.075	26.076	0.310	0.011	0.012	(25.522, 26.747)
6	119	27.313	26.453	0.410	0.011	0.016	(25.642, 27.223)
7	100	27.518	25.856	0.372	0.009	0.014	(25.189, 26.543)
8	137	26.698	25.914	0.324	0.009	0.013	(25.283, 26.503)

NOTE: The Gibbs sampler is run $100,000$ times with a "burn-in" of $25,000$ with a systematic sample of every seventy-fifth, and it took twenty-five minutes.

5. Improvements and extensions

In this section, we show what improvements can be made to our new procedure and possible extensions. We also show how to make inference about all propensity scores, not just those associated with the non-probability sample, but its non-sampled part of the population as well.

5.1. Improvements, Mahalanobis distance

We discuss how to replace the Mahalanobis distance because it is not appropriate with discrete variables. Leon and Carriere (2005) introduced a generalized Mahalanobis distance for mixed data but this method is cumbersome; so we seek a simpler mathod that avoids the Mahalanobis distance completely. Basically we use the nps (1) and the ps (2) to get propensity scores, and matching to get surrogates for $y_{2j}, j = 1, \ldots, n_2$. Then, we use the nps (1) and ps (2) with the surrogates, to get propensity scores again, and matching to

Table 8: Comparison of the six models by summaries of the PMs and PSDs over the eight counties

	PM			PSD		
Models	EW	UW	RD (%)	EW	UW	RP (%)
Covariates (BHF)	27.569	27.226	15.391	0.363	0.418	15.295
NR, NS-RE (SS)*	27.540	27.174	16.840	0.143	0.152	5.844
RS, NS-RE (SS)*	27.186	26.732	26.243	0.265	0.329	24.381
RS, S-RE (SS)	27.034	26.372	48.264	0.294	0.298	1.200
RS & S-R-RE (SS)	26.917	26.215	57.703	0.233	0.284	22.009
RS & S-SB-RE (SS)	27.081	26.384	50.402	0.341	0.326	-4.423

NOTE: Scott-Smith (SS) spatial models have the CAR prior and the non-spatial models* replace $R - \psi V$ by the identity matrix. The summaries are respectively the arithmetic mean of the eight PMs and geometric mean of the eight PSDs. The first three models are non-spatial (NS), and last four models have robust study variable (RS).

get surrogates for $W_{1j}, j = 1, \ldots, n_1$. The y_{2j} are not used for further analysis.

Our procedure has two steps. First, we massively impute $y_{2i}, i = 1, \ldots, n_2$, using propensity scores (i.e., matching via nearest neighbors). Second, we match (nearest neighbor) propensity scores conditional on the nps and ps data now available.

In the first step, we define $I_i = 0$ if $i = 1, \ldots, n_1$ (nps) and $I_i = 1$ if $i = n_1+1, \ldots, n_1+n_2 = n$ (ps); $\underline{x}_i = \underline{x}_{1i}, i = 1, \ldots, n_1$ and $\underline{x}_{n_1+i} = \underline{x}_{2i}, i = 1, \ldots, n_2$. We then assume logistic regression model,

$$I_i \overset{ind}{\sim} \text{Bernoulli}(\frac{e^{\underline{x}_i'\underline{\beta}}}{1 + e^{\underline{x}_i'\underline{\beta}}}), i = 1, \ldots, n = n_1 + n_2.$$

Let $\hat{\underline{\beta}}$ denote the maximum likelihood estimator of $\underline{\beta}$, we have propensity scores,

$$\pi_i = \frac{e^{\underline{x}_i'\hat{\underline{\beta}}}}{1 + e^{\underline{x}_i'\hat{\underline{\beta}}}}, i = 1, \ldots, n.$$

We fill in the missing $y_{2i}, i = 1, \ldots, n_2$, using matching on the π_i. For each $i = n_1+1, \ldots, n_1+n_2$, we find which $j, j = 1, \ldots, n_1$, minimizes $|\pi_i - \pi_j|$, say j^*; j^* may not be unique. Then, the value of the study variable given to unit i is $y_{2j^*}, i = n_1 + 1, \ldots, n_1 + n_2$.

In the second step, we define $\underline{x}_i = \underline{x}_{1i}, i = 1, \ldots, n_1, \underline{x}_i = \underline{x}_{2i}, i = 1, \ldots, n_2$. Similarly, we define $y_i = y_{1i}, i = 1, \ldots, n_1, y_i = y_{2i}, i = 1, \ldots, n_2$. Also, define $I_i = 1, i = 1, \ldots, n_1$, for the nps (1) and $I_i = 0, i = n_1 + 1, \ldots, n_1 + n_2 = n$, for the ps (2). Note that $\underline{x}_i$ has p components, including an intercept. For the nonignorable model, we assume logistic regression,

$$I_i \mid \underline{\beta}, y_i \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{(\underline{x}_i'\underline{\beta}_{(p)} + y_i\beta_{p+1})}}{1 + e^{(\underline{x}_i'\underline{\beta}_{(p)} + y_i\beta_{p+1})}}\right\}, i = 1, \ldots, n.$$

We optimize the likelihood function to obtain the maximum likelihood estimator of $\underline{\beta}$, which we now denote by $\hat{\underline{\beta}}$. The propensity scores are then

$$\pi_i = \frac{e^{\underline{x}_i'\hat{\underline{\beta}}+y_i\hat{\beta}_{p+1}}}{1+e^{\underline{x}_i'\hat{\underline{\beta}}+y_i\hat{\beta}_{p+1}}}, i = 1, \ldots, n.$$

For each $i = 1, \ldots, n_1$, we find which $j, j = n_1 + 1, \ldots, n$, minimizes $|\pi_i - \pi_j|$, say j^*; j^* may not be unique. Then, the weight given to unit i is $W_{2j^*}, i = 1, \ldots, n_1$. Denote these weights by $W_{1i}, i = 1, \ldots, n_1$. Letting $N = \sum_{i=1}^{n_2} W_{2i}$, the design estimate of the population size, our final weights for the nps (1) are

$$W_{1i} \equiv N\frac{W_{1i}}{\sum_{i=1}^{n_1} W_{1i}}, i = 1, \ldots, n_1.$$

Note that the y_{2i} are discarded and are not used in any further analysis.

5.2. Extensions

Unfortunately, in our new procedure sample sizes and the sub-population sizes of the strata (domains) are random variables. The uncertainty in their values should be taken into consideration.

Let $\underline{j} = (1, \ldots, 1)'$, a vector of ones, and $\underline{O} = (0, \ldots, 0)'$, a vector of zeros, $N_g >> n_g, g = 1, \ldots, G$ denote G-vectors.

a. Population sizes

Letting $\underline{N} = (N_1, \ldots, N_G)'$, domain sizes from ps (1), we assume

$$\underline{N} - \underline{j} \sim \text{Multinomial}(N - G, \underline{P}), \quad \underline{P} \sim \text{Dirichlet}(\underline{O}),$$

$$\underline{P} \mid \underline{N} \sim \text{Dirichlet}(\underline{N} - \underline{j}), \quad \underline{T} - \underline{j} \mid \underline{P}, \underline{N} \sim \text{Multinomial}(N - G, \underline{P})$$

b. Sample sizes

Letting $\underline{n} = (n_1, \ldots, n_G)'$, observed domain sizes, we assume

$$\underline{n} - \underline{j} \sim \text{Multinomial}(n - G, \underline{p}), \quad \underline{p} \sim \text{Dirichlet}(\underline{O}),$$

$$\underline{p} \mid \underline{n} \sim \text{Dirichlet}(\underline{n} - \underline{J}), \quad \underline{t} - \underline{j} \mid \underline{p}, \underline{n} \sim \text{Multinomial}(n - G, \underline{p}).$$

c. Posterior inference

Using Bayes' theorem, the joint posterior density is

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}),$$

where $\underline{z}$ is a vector of latent variables; $z_{gj}, j = 1, \ldots, t_g, g = 1, \ldots, G$. There are three cases, which we must consider,

a. If $t_g < n_g$, take a simple random sample without replacement from $y_{gj}, j = 1, \ldots, n_g$ (carry on W_{gj});

b. If $t_g = n_g$, retain $(y_{gj}, W_{gj}), j = 1, \ldots, n_g$;

c. If $t_g > n_g$, take all $(y_{gj}, W_{gj}), j = 1, \ldots, n_g$ and draw a simple random sample with replacement from them to get the others.

We use the following decomposition of the joint posterior density,

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}) =$$

$$\pi_1(\Omega \mid \underline{z}, \underline{t}, \underline{T}, \underline{y}, \underline{W}, \underline{n}, \underline{N})\ \pi_2(\underline{z} \mid \underline{t}, \underline{T}, \underline{y}, \underline{W}, \underline{n}, \underline{N})\ \pi_3(\underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}).$$

We assume

$$\pi(\Omega, \underline{z}, \underline{t}, \underline{T} \mid \underline{y}, \underline{W}, \underline{n}, \underline{N}) =$$

$$\pi_1(\Omega \mid \underline{z}, \underline{t}, \underline{y}, \underline{W}, \underline{n})\ \pi_2(\underline{z} \mid \underline{t}, \underline{y}, \underline{W}, \underline{n})\ \pi_{30}(\underline{t} \mid \underline{n})\ \pi_{31}(\underline{T} \mid \underline{N}).$$

5.3. Inference for all propensity scores

As usual, there are two cases, ignorable selection and non-ignorable selection. For ignorable selection in our new procedure, logistic regression is null and void. Basically, we have a simple random sample without replacement from each stratum. So the selection probabilities for the g^{th} stratum are $n_g/N_g, j = 1, \ldots, N_g$; obviously these vary with the sizes of the strata/domains. It does not affect the modeling of the study variable. It is trivial to deal with random sample sizes of the strata/domains; see Section 5.2. Now the method of Chen, Li and Wu (2020) is completely useless.

For nonignorable selection, we do not need the Horvitz-Thompson estimator (pseudo-likelihood). However, we will obtain the propensity scores for the entire population of N individuals, and therefore these propensity scores can be interpreted as selection probabilities. We use the following logistic regression model,

$$\prod_{g=1}^{G} \left[\prod_{j=1}^{n_g} \frac{e^{\underline{x}_g'\underline{\beta} + y_{gj}\beta_{p+1}}}{1 + e^{\underline{x}_g'\underline{\beta} y_{gj}\beta_{p+1}}} \left\{ \frac{1}{e^{\underline{x}_g'\underline{\beta} + y_{gj}\beta_{p+1}}} \right\}^{N_g/n_g - 1} \right].$$

Note that we are assuming each individual in the g^{th} stratum in the sample is reproduced N_g/n_g with the same value of the study variable. Maximum likelihood estimators, $\hat{\underline{\beta}}$, can be obtained for $\underline{\beta}$. Hence, the propensity scores, π_{gs}, are given by

$$\pi_{gs} = \frac{e^{\underline{x}_g'\hat{\underline{\beta}} + y_{gs}\hat{\beta}_{p+1}}}{1 + e^{\underline{x}_g'\hat{\underline{\beta}} y_{gs}\hat{\beta}_{p+1}}}, s = (j-1)\frac{N_g}{n_g} + 1, \ldots, j\frac{N_g}{n_g},\ j = 1, \ldots, n_g.$$

Some comments are in order.

(a) It is possible to provide a full Bayesian method to obtain the propensity scores.

(b) The π_{gs} are very variable, and they will not add up to n, and we can rake them up to n. String out the π_{gs} as $\pi_i, i = 1, \ldots, N$.

(c) The raking procedure can be a bit problematic because $0 < \pi_i < 1, i = 1, \ldots, N$ and $\sum_{i=1}^{N} \pi_i = n$ for sampling without replacement. Of course, if $\sum_{i=1}^{n} \pi_i \geq n$, raking does not cause any problems. If $\sum_{i=1}^{n} \pi_i < n$, there are difficulties because raking up will make some of the raked $\pi_i > 1$, which we do not want. Keep the π_i in the fourth quartile unchanged. Suppose $\sum_{i \in Q_4} \pi_i = n_o$, where Q_4 is the fourth quartile; then rake up the π_i in the first three quartiles to $n - n_o$.

(d) It is not much more difficult to make inference about the study variable with random sample sizes of the strata/domains.

6. Concluding remarks

A structurally complete probability sample is obtained from the non-probability sample using double mass imputation with supplemental data from a relatively small probability sample (no study variable). The population is stratified by distinct covariates and the nps and ps are allocated to the strata. The study variable and the covariates are used to construct an incidence matrix (spatial structure), which is used to accommodate the covariates. The covariates are never used in the proposed models.

We have used the Scott-Smith model to avoid specifying the uncertain relationship between the study variable and covariates for unit level data without consideration of the participation variable. Robust models are specified for both the study variable and the random effects, and in one model, the Pitman-Yor two-parameter stick-breaking process is used. This is needed because the population model, used for prediction, is assumed to be correct. The nps data are used to construct an incidence matrix with the neighboring strata.

The small areas are not modeled directly, rather the model is placed on the strata with the non-probability samples. Inference about the actual small areas is obtained in the output analysis. This helps with over-shrinkage.

A simulation study to assess the predictive power of the proposed models will be useful. Sensitivity of Bayesian predictive inference of the finite population mean to the size of the ps sample needs to be investigated. Further robustification can be done using the stick-breaking process for the sampling process (study variable) instead of the two component mixture model for the study variable in the Scott-Smith model. An alternative approach, using structural error models, is presented by Nandram (2023).

The method is promising, and the non-Bayesian notion of double robustness is null and void. The probability sample plays a fairly minor role, and it can be eliminated if the required population information can be obtained using web-scraping, an emerging science. There is a good comparison with the BHF (baseline) model. Future work will be focused in this direction (stratification and matching).

At one of my talks, a participant asked if the means of the two components in mixture model can be different. Of course, they can be different; the mixture model in the stick-breaking process has a different mean for each cluster. The model of Goyal *et al.* (2020) is limited, and so in Appendix B, we describe how one might use different means for a single area; thereby adding flexibility to this two-component mixture model.

Acknowledgements

This work was presented as invited lectures at the IISA 2023 Conference, Colorado, USA, at the ISI 2023 Conference, Ottawa, Canada, and in India, Febrauary 2024, as a Special Invited Lecture at the SSCA Conference, Banasthali Vidyapith, Rajasthan, and in March 2024, at the Indian Institute of Management (IIM), Ahmedabad, as a more extended lecture. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram). Balgobin also thanks five of his PhD students (Dr. Ashley Lockwood, Dr. Lingli Yang, Dr. Dilli Bhatta, Yang Liu and Zihang Xu) for a final reading of the paper.

Appendices

Appendix A: Extension of the Scott-Smith model with stick breaking process

We extend the Scott-Smith model to have a robust model on the study variable, spatial effects and random effects, where the random effects have a stick-breaking prior (Ishwaran and James, 2001).

For $g = 1, \ldots, G$, $j = 1, \ldots, N_g$, again we have the mixture population model (robustness),

$$y_{gj} \mid \mu_g \sim (1-p)\text{Normal}(\mu_g, \gamma\sigma^2) + p\text{Normal}(\mu_g, \sigma^2), \tag{A.1}$$

where $0 < p < \frac{1}{2}, 0 < \gamma < 1$.

For $g = 1, \ldots, G$, $j = 1, \ldots, n_g$, again we have the sample model,

$$y_{gj} \mid \mu_g \sim (1-p)\text{Normal}(\mu_g, \gamma\frac{\sigma^2}{w_{gj}}) + p\text{Normal}(\mu_g, \frac{\sigma^2}{w_{gj}}), \tag{A.2}$$

where $0 < p < \frac{1}{2}, 0 < \gamma < 1$.

Then,

$$\underline{\mu} \mid \underline{\eta} \sim \text{Normal}\{\underline{j}\theta + \underline{\eta}, \frac{\rho_1}{1-\rho_1}\sigma^2(R - \psi V)^{-1}\}. \tag{A.3}$$

We consider the Pitman-Yor two parameter stick-breaking process for $\eta_g, g = 1, \ldots, G$.

The stick-breaking process is

$$\pi(\eta_g \mid \underline{t}, etc.) = \sum_{s=1}^{G_o} p_s \text{Normal}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2),$$

where G_o is the number of distinct clusters with independence over $g = 1, \ldots, G$. Using the latent variables, d_g, it can be expressed in a computationally convenient form,

$$\pi(\eta_g, d_g \mid \underline{t}, etc.) = \prod_{s=1}^{G}[p_{d_g}\text{Normal}_{\eta_g}\{t_{d_g}, \frac{\rho_2}{1-\rho_2}\sigma^2\}]^{I(d_g=s)}$$

with independence over $(\eta_g, d_g), g = 1, \ldots, G$. The number of clusters, G_o, is the number of distinct d_g and d_g informs which cluster η_g belongs. Note that the limit of the product is ℓ

and it cannot be larger; this is different from a Dirichlet process, where the upper limit in the product goes to infinity. Here $p_1 = \nu_1, p_2 = \nu_2(1-\nu_1), \ldots, p_{G_o} = \prod_{s=1}^{G_o-1}(1-\nu_s)$ and

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

Then, for $\underline{\eta}$, we have

$$\underline{\eta} \mid \underline{t}, \underline{d} \sim \text{Normal}\{P\underline{t}, \frac{\rho_2}{1-\rho_2}\sigma^2 I\}, \tag{A.4}$$

where P is an incidence (partition) matrix (i.e., it consists of zeros and ones), mapping the areas to the clusters. Finally, we assume

$$t_s \overset{ind}{\sim} \text{Normal}\{0, \frac{\rho_3}{1-\rho_3}\sigma^2\}, s = 1, \ldots, G_o. \tag{A.5}$$

Note that $\underline{\eta}$ and $\underline{t}$ are G_o-vectors.

Also, it is clear that

$$Pr(d_g = s \mid \underline{t}, etc.) \propto p_s \text{Normal}_{\eta_g}(t_s, \frac{\rho_2}{1-\rho_2}\sigma^2), s = 1, \ldots, G, g = 1, \ldots, G.$$

Once $d_g, g = 1, \ldots, G$, are sampled, the incidence matrix, P is obtained, and all other parameters can be sampled. That is, draw the d_g first, and all other parameters can be sampled easily. Note at this moment, if we do not have enough z_s or ν_s, we must sample their priors.

We need to specify a proper prior for θ. Actually, $\mu_g = \phi_g + \eta_g, g = 1, \ldots, G$, where the ϕ_g are spatial effects and the η_g are clustering effects. Therefore, there is weak identifiability and we must take care of this issue. Now, $\text{E}(\mu_g \mid \theta, \eta_g) = \text{E}(\phi_g \mid \theta, \eta_g) + \eta_g$, and $\text{E}(\phi_g \mid \theta, \eta_g) = \theta$. Also, note that $\text{E}(\underline{\eta}) = \text{E}\{\text{E}(\underline{\eta} \mid \underline{t})\} = \text{E}(P\underline{t}) = \underline{0}$. Therefore, we have centered the μ_g on θ by taking $\text{E}(\phi_g \mid \eta_g) = \theta$ and $\text{E}(\underline{\eta}) = \underline{0}$ This centering together with proper diffused priors on θ can overcome the weak identifiability in this hierarchical Bayesian model. It is possible to improve this model further.

However, there are some additional problems that are likely to occur with stick breaking. First, there can be a single cluster, but at least three clusters are needed in any partition. One can use a random grouping with three clusters if this happens (this is rare). Second, to allow a relatively larger number of clusters, we take $1 - \delta_1 > \frac{1}{2}$, better than Jeffrey's prior for the first parameter of the beta density, and $\frac{1-\delta_2}{\delta_2} < 1$, thereby assisting the second beta parameter from getting too large a priori. These two conditions give $0 < \delta_1 < \frac{1}{2}$ and $\frac{1}{2} < \delta_2 < 1$. Third, an informative prior for θ is required. This can be obtained by using a small sub-sample of the data to avoid double using all the data. These three things help to obtain a more efficient Gibbs sampler.

To obtain the prior for θ, we can take a random sample of 10% of the data, $\bar{y}_g, g = 1, \ldots, G$. Now calculate the average, θ_o, and the variance, σ_o^2. Then, we take $\theta \sim \text{Normal}(\theta_o, \sigma_o^2)$. This is like a proper diffused prior, which avoids the uncertainty in inflating σ_o^2 near to

vagueness. Admittedly it double uses the data, but only 10% of the data, not all the data. Otherwise, we would need a prior for θ from an independent source. For example, if we also have the study variable from a small probability sample, we can use that as we have done here. Alternatively, we can take a small percent (*e.g.*, 5%) of the nps data, and do the same, and we can use the remaining data for the analysis. So that we do not need the ps for the construction of the prior.

The CPD of θ consists of two pieces, one from the proper prior and other from the model. It is easy to show that the contribution from the model is

$$\theta \sim \text{Normal}\left\{\frac{\sum_{g=1}^{G} r_g \mu_g}{\sum_{g=1}^{G} r_g}, \frac{\rho_1}{1-\rho_1}\sigma^2 \frac{1}{(1-\psi)\sum_{g=1}^{G} r_g}\right\},$$

where r_g are the row (column) sums of V. It is now a standard calculation to combine the two pieces.

Starting values for the Gibbs sampler can be obtained by first doing three things.

a. Find the sample averages, $\bar{y}_g = \frac{1}{n_g}\sum_{j=1}^{n_g} y_{gj}, g = 1, \ldots, G$.

b. Find clusters in these G stratum means, say 10 clusters, and this will give initial values of the d_g.

c. Form the partition matrix, P.

With this set up, we can generate starting values for all pertinent parameters. We can set $\rho_s = \frac{1}{2}, s = 1, 2, 3$, and $\delta_1 = \delta_2 = \frac{1}{2}$. Also, $t_s, s = 1, \ldots, 10$, can be obtained by averaging the appropriate y_g. Then, we can now sample the $\eta_g, g = 1, \ldots, G$. At the first iterate of the griddy Gibbs sampler, we can easily sample the d_g followed by all the parameters.

Appendix B: Unequal means in the two-component mixture model

We consider a single area (or stratum) to show how we can proceed. We assume

$$y_i \mid p, \mu, \sigma^2, \rho, \gamma \overset{ind}{\sim} (1-p)\text{Normal}_{y_i}(\mu - \gamma, \rho\sigma^2) + p\text{Normal}_{y_i}(\mu, \sigma^2), i = 1, \ldots, n,$$

where the first component has smaller mean and variance, and a priori,

$$\pi(p, \mu, \sigma^2, \rho, \gamma) \propto \frac{1}{\sigma^2}, 0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0.$$

Continuous survey data are typically skewed to the right, and so it is safe to take $\gamma > 0$. Of course, we can do regression (with covariates) in a similar manner.

The joint posterior density is

$$\pi(p, \mu, \sigma^2, \rho, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2}\prod_{i=1}^{n}\left\{\frac{1-p}{\sqrt{\rho\sigma^2}}\phi\{(y_i - \mu + \gamma)/\sqrt{\rho\sigma^2}\} + \frac{p}{\sqrt{\sigma^2}}\phi\{(y_i - \mu)/\sqrt{\sigma^2}\}\right\}, \tag{B.1}$$

where $\phi(\cdot)$ is the standard normal density and $0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0$.

We can simplify drawing samples from the joint posterior density by introducing latent variables, $z_i, i = 1, \ldots, n$, where $z_i = 0$ if an observation comes from the first component and $z_i = 1$ if an observation comes from the second component. Then, the augmented joint posterior density, starting with (B.1), is now

$$\pi(\underline{z}, p, \mu, \sigma^2, \rho, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2} \prod_{i=1}^{n} \left\{ \left[\frac{1-p}{\sqrt{\rho\sigma^2}} \phi\{(y_i - \mu + \gamma)/\sqrt{\rho\sigma^2}\} \right]^{1-z_i} \left[\frac{p}{\sqrt{\sigma^2}} \phi\{(y_i - \mu)/\sqrt{\sigma^2}\} \right]^{z_i} \right\}, \tag{B.2}$$

where $0 < p < 1/2, 0 < \rho < 1, |\mu|, \sigma^2, \gamma > 0$. It is now easy to run a Gibbs sampler to fit the joint posterior density in (B.2). It is advisable to sample the joint conditional posterior density of (μ, σ^2) (i.e., blocking).

Note that if we have only the first (or the second) component in the model (i.e., $p = 0$ or $p = 1$), then γ and ρ will not be identifiable. Therefore, it is necessary to assume there are at least two observations from each component of the mixture to avoid impropriety of the joint posterior density (i.e., $2 \leq \sum_{i=1}^{n} z_i \leq n - 2$). One way to do this is to arrange $y_1, \ldots, y_n$ in increasing order $y_{(i)}, i = 1, \ldots, n$, and take the corresponding $z_1 = z_2 = 0$ and $z_{n-1} = z_n = 1$, where $z_i, i = 3, \ldots, n-2$, are determined from the joint posterior density. Doing so will avoid specifying this difficult constraint in the joint posterior density in (B.2).

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Beaumont, J-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, **46**, 1-28.

Beaumont, J-F. and Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, **83**, 11-22, DOI: 10.17226/24893.

Battese, G. E., Harter, R., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. **83**, 28-36, DOI: 10.2307/2288915.

Chakraborty, A., Datta, G. S., and Mandal, A. (2019). A robust hierarchical Bayes small area estimation for nested error linear regression model. *International Statistical Reviews*, **87**, S1, S158-S156, DOI: 10.1111/insr.12283.

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

Chung, H. C. and Datta, G. S. (2022). Bayesian spatial models for estimating means of sampled and nonsampled small areas. *Survey Methodology*, **48**, 463-489.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, **32**, 249–264, DOI: 10.1214/16-STS598.

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

He, Z. and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, **56**, 360-367, DOI: 10.1111/j.0006-341X.2000.00360.x

Ibrahim, J. G. and Chen, M-H. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46-60, DOI: 10.1214/ss/1009212673.

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

Kim, J., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society*, Series A, **184**, 941-963.

Leon, de A. R. and Carriere, K. C. (2005). A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis*, **92**, 174-185, DOI: 10.1016/j.jmva.2003.08.006.

Lockwood, A. (2023). *Bayesian Predictive Inference for a Study Variable Without Specifying a Link to the Covariates.* PhD Dissertation, Department of Mathematical Sciences, Worcester Polytechnic Institute, pg. 1-110.

Marella, D. (2023). Adjusting for selection bias in non-probability samples by empirical likelihood approach. *Journal of Official Statistics*, **39**, 2023, 151-172, DOI: 10.2478/JOS-2023-0008.

Molina, I., Nandram, B., and Rao, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, **8**, 852-.885, DOI: 10.1214/13-AOAS702.

Nandram, B., Toto, M. C. S., and Choi, J. W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation*, **81**, 1593-1608.

Nandram, B. (2007). Bayesian predictive inference under informative sampling Via surrogate samples. In *Bayesian Statistics and Its Applications, Eds. S.K. Upadhyay, Umesh Singh and Dipak K. Dey*, Anamaya, New Delhi, Chapter 25, 356-374.

Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, **105**, 120-135, DOI: 10.1198/jasa.2009.ap08443.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 4-17, DOI: 10.5539/ijsp.v10n6p5.

Nandram, B. and Rao, J. N. K (2021). A Bayesian approach for integrating a probability sample with a nonprobability sample. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1568-1603.

Nandram, B. and Rao, J. N. K (2023). Bayesian predictive inference when integrating a nonprobability sample and a probability sample. *arXiv:2305.08997v1 [Stat.ME], 15 May 2023*, pg. 1-35.

Nandram, B. and Rao, J. N. K. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications*, **22** 345-376, ISSN 2454-7395.

Nandram, B. (2023). Overcoming challenges associated with early Bayesian state estimation of planted acres in the United States. *Special Proceedings of the Twenty-fifth Conference of the Society of Statistics, Computer and Applications*, ISBN #: 978-81-

950383-2-9, 25th Annual Conference, 15-17 February 2023; pp 51-78.

Nandram, B. (2022), A Bayesian assessment of non-ignorable selection of a non-probability Sample. *Indian Bayesians' News Letter, Invited Paper*, **14**, November 2022, 7-20.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya*, Series B, **83**, 242-272, DOI: 10.1007/s13571-020-00227-w.

Rafei, A., Elliott, M. R., and Flannagan, C. A. C. (2022). Robust and efficient Bayesian inference for non-probability samples. *arXiv:2203.14355Vi*, pp. 1-46.

Sakshaug, J. W., Wisniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, **35**, 653-681, DOI: 10.2478/jos-2019-0027.

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wisniowski, A., and Struminskaya, B. (2023). Bayesian integration of probability and non-probability samples for logistic regression. *Journal of Survey Statistics and Methodology*, 00,c1-35, DOI: 10.1093/jssam/smad041.

Scott, A. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, **64**, 830-840, DOI: 10.1080/01621459.1969.10501015.

Toto, M. C. S. and Nandram, B. (2010). A Bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference*, **140**, 2963-2979, DOI: 10.1016/j.jspi.2010.03.043.

Wisniowski, A., Sakshaug, J. W., Ruiz, D. A. P., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey Inference, *Journal of Survey Statistics and Methodology*, **8**, 120-147, DOI: 10.1093/jssam/smz051.

Wu, C. (2022). Statistical Inference with non-probability survey samples. *Survey Methodology (With Discussions)*, **48**, 283-311.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 31-50

Planes, Designs and List Designs

Navin Singhi
School of Mathematics (retd), TIFR, Colaba, Mumbai 400005

Received: 10 May 2024; Revised: 16 May 2024; Accepted: 18 May 2024

Abstract

Existence and construction of combinatorial designs, projective and affine planes, nets has been a topic, extensively studied during last 8-10 decades. Main interest arose from classical projective geometry, group theory and applications in Statistics, in designs of experiments, computer science and digital electronics *etc.* The paper gives a short survey of trends in Discrete Mathematics focused on topics of planes, nets, designs and list designs (designs with multisets as blocks). Main methods used during these decades have been algebraic methods, graph-theoretic methods, probabilistic methods and combinatorial techniques of forming bigger designs by pasting together smaller ones.

Key words: Designs; Planes; Algebraic methods; BIBD; Latin squares.

AMS Subject Classifications: 05B05, 05B25

1. Basic definitions

Let X be a finite set of v elements, $X = \{x_1, x_2, \ldots, x_v\}$. We will denote by $\mathbb{P}(X)$, the set all subsets of X and by $\mathbb{P}_k(X)$, the set of all k-subsets of X, $0 \leq k \leq v$. We will denote by $V(X)$, the set of all rational valued functions $f : \mathbb{P}(X) \to \mathbb{Q}$. Clearly $V(X)$ is a vector space over $\mathbb{Q}$, of dimension 2^v. The set $M(X) \subseteq V(X)$ of all integral valued functions, is clearly a module of rank 2^v over the ring of integers $\mathbb{Z}$. By $\mathbb{N}(X)$ we will denote the set of all nonnegative integral valued functions $f : \mathbb{P}(X) \to \mathbb{N}$. Thus $N(X) \subseteq M(X) \subseteq V(X)$.

Similarly we will denote by $V_k(X)$ the subspace of $V(X)$ of dimension $\binom{v}{k}$ of all rational valued functions $f \in V(X)$ such that $f(B) = 0$ if $|B| \neq k$. Thus when $f \in V_k(X)$, we can also think of f also as a function $f : \mathbb{P}_k(X) \to \mathbb{Q}$. We will denote by $M_k(X)$ the submodule of $M(X)$ of rank $\binom{v}{k}$ of all $f \in M(X)$ such that $f(B) = 0$ if $|B| \neq k$. Thus $f \in M_k(X)$ can also be thought of as an integral valued functions $f : \mathbb{P}_k(X) \to \mathbb{Z}$. Similarly by $N_k(X)$, we will denote the subset of $N(X)$ of all $f \in N(X)$ such that $f(B) = 0$, if $|B| \neq k$. Thus $f \in N_k(X)$, we will also think of as a nonnegative integral valued function $f : \mathbb{P}_k(X) \to \mathbb{N}$. Thus $N_k(X) \subseteq M_k(X) \subseteq V_k(X)$.

For any real valued function $f : X \to \mathbb{R}$, the subset $supp(f) \subseteq X$ defined by $supp(f) = \{x \in X \mid f(x) \neq 0\}$ is called the **support** of the function f. A **list** on X (also called a

Corresponding Author: Navin M. Singhi
Email: niknavin@gmail.com

frequency function on X) is a map $\ell : X \to \mathbb{N}$. For each $x \in X$, $\ell(x)$ is called the **multiplicity** or the **frequency** of x in the list ℓ. The subset $supp(\ell) \subseteq X$, is called the **support** of the list ℓ.

A list ℓ on X is essentially a **multiset** on X. We can also visualize the list ℓ on X as a multiset ℓ on X, where for each $x \in X$, $\ell(x)$ gives the number of times the element x occurs in the multiset ℓ.

Example 1: The multiset $\ell = [x, x, y, y, y]$ is the same as the list ℓ defined by $\ell(x) = 2$, $\ell(y) = 3$ and $\ell(z) = 0$ for $z \neq x, y$. Also, the multiset $[x, x, y, y, y]$ is the same as the multiset $[x, y, x, y, y]$ or $[y, y, y, x, x]$ *etc.* In general we can visualize a multiset or a list as an indexed family $(x_i)_{i\in I}$ or an unordered tuple $[x_i | i \in I]$. For example the multiset ℓ considered here is an unordered tuple $[x_i | 1 \leq i \leq 5]$ with $x_1 = x_2 = x$ and $x_3 = x_4 = x_5 = y$. When the indexing set $I = I_n = \{1, 2....n\}$, we may also use the notation $[x_1, x_2, \ldots, x_n]$ for the multiset $[x_i | i \in I_n]$.

We also note that the set $N(X)$ is the set of all lists (or all multisets) on the set $\mathbb{P}(X)$.

We will denote by $\boldsymbol{|\ell|}$ the sum $\sum \ell(x)$, summed over all $x \in X$ and call it the **size** of the list ℓ. If $|\ell| = k$, we will say that ℓ is a ***k*-list** or a ***k*-multiset**. A list ℓ is clearly a **subset** of X if $\ell(x) \in \{0, 1\}$ for all $x \in X$.

We will denote by $\boldsymbol{L(X)}$ the set of all lists on X and by $\boldsymbol{L_k(X)}$ the set of all k-lists on X. We note that $|L_k(X)|$ is the same as the number of ways to choose k elements from the set X with repetitions allowed and is thus given by

$$|L_k(X)| = \binom{m+k-1}{m-1} = \binom{m+k-1}{k},\ m = |X|. \tag{1}$$

Now suppose $\ell, \ell_1 \in L(X)$. We will say that ℓ_1 is a **sublist** (or a **submultiset** when we consider ℓ as a multiset) of ℓ and denote it by $\boldsymbol{\ell_1 \subseteq \ell}$, if and only if $\ell_1(x) \leq \ell(x)$ for all $x \in X$.

Unlike sets, a multiset or a list ℓ_1 can occur as a submultiset of ℓ in several ways. In fact the number of ways in which the multiset ℓ_1 occurs as a submultiset ℓ is precisely $\boldsymbol{c(\ell, \ell_1)} = \prod\limits_{x\in X} \binom{\ell(x)}{\ell_1(x)}$. We note that a product over an empty set is defined to be 1.

Example 2: Let $\ell = [x_i \mid i \in I_{10}]$ be a multiset on a set $X = \{a, b, c\}$. Suppose $x_1 = x_2 = x_3 = a, x_4 = x_5 = x_6 = b$ and $x_7 = x_8 = x_9 = x_{10} = c$. Thus $\ell = [a, a, a, b, b, b, c, c, c, c]$. Now suppose $\ell_1 = [y_j \mid j \in I_7]$ is the multiset with $y_1 = a,\ y_2 = y_3 = b,\ y_4 = y_5 = y_6 = c,\ y_7 = a$. It can be easily seen that the multiset $[x_i \mid i \in A]$ is same as the multiset ℓ_1 if and only if $A = \{j_n \mid 1 \leq n \leq 7\}$, where $j_1, j_2 \in I_3,\ j_3, j_4 \in \{4, 5, 6\}$ and $j_5, j_6, j_7 \in \{7, 8, 9, 10\}$. Thus $c(\ell, \ell_1) = \binom{3}{2}\binom{3}{2}\binom{4}{3} = 36$

A **design** is a pair $D = (X, f)$, where X is finite set and $f \in N(X)$, *i.e.*, f is nonnegative integral valued function on $\mathbb{P}(X)$. Thus f is just a list on $\mathbb{P}(X)$. The elements

of the set X are called **points** or **treatments** of the design D. When $f(B) \neq 0$, the subset$B$ of X is called a **block** of the design D. For a block B of D, $f(B)$ gives the number of times the block B is **repeated** in the design D, it is also called the **frequency** of the block B in the design D. The number $|B|$ is said to be the **size** of the block B. If all blocks of a design D have size k, then k is said to be the **block-size** of the design D. When the design has the block size k, clearly $f \in N_k(X)$.

Similarly we define a **signed design** to be a pair $D = (X, f)$, where $f \in M(X)$ and a **rational** design to be a pair $D = (X, f)$, where $f \in V(X)$. The blocks, frequency, block-size are similarly defined for these designs too. Note that the frequency of a block of a signed design is an integer thus it may be even negative and for a rational design it is a rational number. Signed designs or rational designs are useful, as a tool to study and construct designs.

When the set $X = \{x_1, x_2, \ldots, x_v\}$ of points is fixed, we may consider f, it self as the design (X, f).

Designs have been one of the main focus of studies in discrete mathematics, since 1940's at least. Specially studies of projective and affine planes, nets and t-designs have been a dominating factor in the field of discrete mathematics for last several decades. These studies have also influenced many other areas. In particular a lot of developments in the study of 2-designs, also called BIBD, was done by Statisticians. Graph and Hypergraph theory, Group theory, Computer science, Applied Algebra, Digital Electronics are some other areas which have been influenced by studies in designs and vice verse.

In the next section we will give a short survey of developments in projective planes and nets in this era. While in section 3 we will do the same in the case of more general t-designs. Note that BIBD's are particular case of t-designs. in fact they are exactly 2-designs. Also symmetric BIBD's with $\lambda = 1$ are precisely Projective planes.

We will give some references in these sections. But let us end this section with mention of three good books discussing these aspects, by Hughes and Piper on projective planes Hughes and Piper (1970), by Beth Jugnickel and Lenz on Design Theory Beth *et al.* (2000) and by Raghavarao on designs and their applications in designs of experiments Raghavarao (1971). Another classic book is Finite Geometries book by Peter Dembowski, a great reference book for both geometries and designs Dembowski (1968)

2. Projective planes and nets

In this section we will study basically finite geometries. These are special cases of designs. In a geometry, generally treatments are called points and blocks are called **lines**. We will also use usual terms from geomtry. For example if two or more points are on a line, they are also called **collinear** and similarly if three or more lines are on the same point, they are called **concurrent**.

A **partial linear space** is a design $D = (X, f)$, such that any pair of distinct points $x, y \in X$ is on at most one line of D. Such a space is called a **linear space**, if every pair of points $x, y \in X$ is on a unique line of D. Note that a linear space is essentially the same as partially balanced design (PBD) with $\lambda = 1$ (see Section 2 for definition of PBD) .

A **projective space** is a linear space $D = (X, f)$, containing four points no three of which are collinear and which satisfies the following Pasch's axiom.

Pasch's axiom: Suppose ℓ_1 and ℓ_2 are two distinct intersecting lines of D, *i.e.*, $\ell_1 \neq \ell_2$ and there is a point $x \in \ell_1 \cap \ell_2$. Also suppose ℓ_3 and ℓ_4 are two lines of D, which are **transversal** to ℓ_1 and ℓ_2, *i.e.*, both of them are not on x but each of them intersects both ℓ_1 and ℓ_2. Then ℓ_3 and ℓ_4 are also intersecting lines.

Each projective space has a unique **dimension** (see for more details Hughes and Piper (1970) or Veblen and Young (1938)). A classical theorem of projective geometry states that every projective space of dimension 3 or more is essentially coordinatized by a field.

The result is not true for a **projective plane**, *i.e.*, a projective space of dimension 2. For many years, people have believed that a finite projective plane with no proper subplane is coordinatized by a prime field (Example below describes, what generally one means by coordinatizing a plane) and that the order of a finite projective plane (order is defined below), is a power of a prime number. Axiomatizing and classifying projective planes and related structures has been a very active field. Included among a large number of mathematicians, who have made significant contributions are Pasch, Hilbert, Dickson, Albert, Hall and Bose. Some good sources for the results and theory are Albert (1961),Hall (1943) Hughes and Piper (1970) and Veblen and Young (1938).

Though the problem of classifying projective planes has been studied for more than 200 years, a spurt in the activity during last few decades was caused by Marshal Hall's via his paper Hall (1943) and by R.C. Bose via his Paper in 1939 Bose (1939). While Marshal Hall connected the problem with many algebraic structures, groups, permutation groups, fields, near fields, nonassociative rings ternary rings *etc.*, Bose was interested in looking at constructing designs, specially BIBD's from projective planes, affine planes, nets*etc.* and even from higher dimensional geometries. He used these designs for the designs of experiments, a branch of statistics, which was just evolving then.These papers made many researchers from all these areas, finite group theory, number theory, algebra, nonassociative algebras, statistics, graph theorists, computer scientists and digital electronics engineers interested in these geometric and designs problems. Perhaps more than 1000 remarkable papers may have evolved on planes nets and t-designs, as a result of these two exceptional path-breaking papers. We will describe some of the results which evolved as a result of these two papers. We will also discuss some recent work of the author (Singhi (2010), Singhi (2009).

We will restrict our discussion in this section essentially to projective and affine planes and nets. As already remarked projective planes, also affine planes are examples of BIBD's, which will be studied in the next section.

It is not too difficult to see (Hughes and Piper (1970)) from the definition of a projective space that a projective space of dimension 2, *i.e.*, a projective plane is a design $D = (X, f)$, satisfying the following conditions and conversely every such design is a projective plane.
(A). D is a linear space, *i.e.*, given any two distinct points $x, y \in X$, there is a unique line (block) ℓ of D such that $x, y \in \ell$.
(B). Any two lines of D intersect in a unique point.

(C). There is exist 4 points in X, no three of which are collinear.
(D). All lines are on the same number $n+1$ points.
(E). All points are on exactly the same number $n+1$ lines.
(F). Total number of points or lines of D are n^2+n+1.

The common number n is called **order** of the plane D.

An **affine plane** is obtained from a projective plane of order n by removing a line and all the points on it. The number n is also called the **order** of the affine plane. It can be easily seen (see Hughes and Piper (1970)) that an affine plane of order n is also a linear space, in which every line is on exactly n points and every point is on exactly $n+1$ lines. Conversely every linear space satisfying these conditions is an affine plane of order n. An affine plane of order n has exactly n^2 points and n^2+n lines.

A **parallel class** is a partial linear space $D=(X,f)$ is a set of lines of D such that every point of D is on exactly one line of this class. Thus a parallel class of partial linear space D is actually a partition of X into lines. It can be easily seen that in an affine plane of order n there are exactly $n+1$ parallel classes, which are mutually disjoint and they partition the set of lines of the affine plane.

A **net**, is a partial linear space $D=(X,f)$ on n^2 points, such that each line of D is on exactly n points and in all there are nr lines, $r\geq 2$, partitioned into r, parallel classes. The net is said to have **order** n and r parallel classes. We will also say that D is $\mathbf{Net}(n,r)$. It can be easily seen that $r\leq n+1$ for a $\mathrm{Net}(n,r)$. When $r=n+1$, one can see that the net is actually an affine plane. It is well-known that a $\mathrm{Net}(n,r)$ gives rise to $r-2$ mutually orthogonal latin squares and conversely. Nets behave as if $n+1-r$ parallel classes are removed from an affine plane of order n. Though not all nets can be completed to an affine plane. Nets were formally defined by Bruck , who studied general problem of embedding a net into an affine plane (seeBruck (1963)). Though as mutually orthogonal latin squares they were studied much earlier, (see Bose (1939)). In particular the embedding problem was solved for the case when $r=n-1$ by Marshal Hall and Connor and by Shrikhande. Bruck proved a much more general result. Bruck's paper is also well known for describing a basic technique, started by Hoffman, of using maximal claws in a graph to find large cliques. Such cliques correspond to adding more lines to the net. Bruck's paper resulted in a lot of activity in studying such problems and connected studies of designs with graphs. Almost the same time R.C. Bose generalized Bruck's ideas to define a strongly regular graph and also generalizing Bruck's nets to a much more general class of partial linear spaces. He called them partial geometries Bose (1963). Strongly regular graphs were studied earlier by statisticians as 2 class association schemes. But looking at them as graphs gave a new thrust to this area and many researchers both in mathematics and statistics started looking at such problems. Later these ideas were further generalized to multigraphs and partial geometric designs by Bose Shrikhande and Singhi and used to solve a much more general problem of embedding of a residual BIBD into a symmetric BIBD Bose *et al.* (1976). Note that projective planes are particular case of symmetric BIBD's.

Example 3: (a). Let F be a finite field of order n. Let X be the set of all ordered pairs of F, $X=\{(x,y)|x,y\in F\}$. Let $m,c\in F$. Define $\ell(m,c)=\{(x,y)\in X|y=mx+c\}$. Also define for each $d\in F, [d]=\{(x,y)\in X|x=d\}$. Let $D=(X,f)$, be the design, where f is defined as follows. For any $B\in\mathbb{P}(X)$, $f(B)=1$ if $B=\ell(m,c)$, $m,c\in F$ or $B=[d]$, $d\in F$

and $f(B) = 0$ in all other cases. It can be easily checked that D is an affine plane of order n, Hughes and Piper (1970). For the line $\ell(m, c)$, m is called the **slope** of the line and c the ***y*-intercept**. All lines of D with a given slope m are parallel and they form a parallel class. Similarly all lines $[d]$, $d \in F$, also form a parallel class, the so called lines parallel to y-axis. The line $[0]$ may be thought of as y-axis and the line $\ell(0, 0)$ is the x-axis. We say that the affine plane D is **coordinatized** by the field F.

One can also use other algebraic structures like quasi fields, near fields, nonassociative division rings *etc.* to construct an affine plane in quite similar manner.

(b). Note that in case we use real field $\mathbb{R}$ instead of a finite field, the above construction exactly gives us the usual real affine plane which we study in high school geometry.

(c). Instead of pairs, now we take a set X_1 of all triplets (x, y, z), $x, y, z \in F$. For each triplet ℓ, m, s of elements of F, define $[\ell, m, s] = \{(x, y, z) \in X_1 | \ell x + my + sz = 0\}$. Let $D_1 = (X_1, f_1)$ be a design, where f_1 is defined by $f_1(B) = 1$ if $B = [\ell, m, s]$ for some $\ell, m, s \in F$ and $f_1(B) = 0$ in all other cases. It can be easily seen that D_1 is a projective plane of order n, coordinatized by the field F.

Let $D = (X, f)$ be a projective plane (or affine plane). A projective plane (resp. affine plane) $D_1 = (Y, g)$ is said to be **subplane** of D if every line of D_1 is a subset of a line of D. A projective plane (or affine plane) is said to be a **prime plane** if it has no proper subplane.

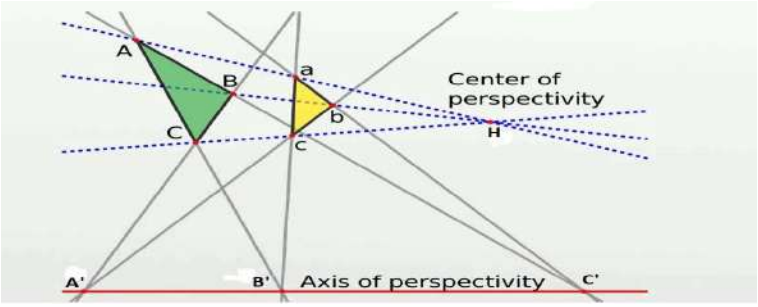
As remarked earlier, apart from the field plane, *i.e.*, the projective or affine plane coordiatized by a finite field, there are many other examples of projective planes for example coordinatized by quasi fields or near fields *etc.* But all known planes so far have order a power of prime. Also all prime planes so far known, have a prime order and are in fact the prime field plane . This gives rise to the following Conjecture.

Projective plane conjecture:
(a). Order of any projective plane is power of a prime number.
(b). A prime projective plane is coordinatized by a prime field.

A lot of research in the area of projective planes has been motivated by these conjectures and related problems. Though the problem is hard.There are very interesting examples of planes, which defy all possibilities of relationship with a field but so far no example of prime planes has been found which is not a prime field plane. Even though there are many planes, for example the so called Hughes planes, which have order q^2, where q is a power of an odd prime, which have subplanes of order 2 or 3 Caliskan and Moorhouse (2011). Such examples show difficulty in solving the problem of classifying projective planes.

As noted earlier Projective spaces of dimension more than 2 are unique and coordinatized by fields. One reason they are coordinatized by fields, is that they satisfy Desargues's Thoerem, which we describe now. Two triangles ABC and abc in a projective space are said to be **centrally perspective** if the lines Aa, Bb, Cc are concurrent, at a point say H. The point H is called the **centre of perspectivity** of the triangles ABC and abc. Similarly dually consider the three points A', B' and C' of intersection of three pairs of lines

(BC, bc), (AC, ac) and (AB, ab) respectively. Suppose the points A', B' and C' are collinear. Then The triangles ABC and abc are said to be **axially perspective** and the line $A'B'C'$ is called **axis of perspectivity**. Desargues's theorem says that if any two triangles in a projective space of dimension 3 or more are centrally perspective then they are also axially perspective.



The above figure with 10 Points A, B, C, a, b, c centre of perspectivity H, and three points A', B' and C' on the axis of perspectivity, together with the 10 lines on these points as given in the above figure is called a **Desargues's Configuration**, Hughes and Piper (1970), ChapterIV. Note that some times H may be on the axis of perspectivity too.

In a general projective plane, it is not true that two triangles which are centrally perspective are also axially perspective. But it is known that every projective plane has several pairs of centrally perspective triangles, which are also axially perspective. Desargues's configuarations play an important role in classification projective planes Hughes and Piper (1970), Chapter IV. We will discuss this later in this section.

Marshal Hall's 1943 paper was a landmark. Among many interesting ideas in the paper, one of them, idea of associating a ternary ring with a projective plane gave a completely new color to the study of these planes. If you look at Example 3(a) above, the affine plane coordinatized by a finite field F, the equation of the line $\ell(m, c)$ not parallel to the y-axis is $y = mx + c$. Marshal Hall had a bright idea that in the case of looking at projective plane, instead of looking at product and addition as operations in field or other such structures, it seems more natural to look at a ternary operation $\tau(m, x, c)$, instead of addition and product in the field, where $\tau(m, x, c) = mx + c$. He was indeed right. He defined a general such ternary ring, called it, planar ternary ring, which we will describe below and he showed that Every projective or affine plane can be coordinatized by such a ring. This insight opened up study of projective planes to a new areas. A problem which looked so far to be of geometry type, suddenly started looking equally as an algebraic problem. We now define planar ternary rings, defined by Hall in this paper.

Let S be a finite set. An ordered pair $R = (S, \tau)$ is said to be a **ternary ring**, if $\tau : S \times S \times S \to S$. When the ternary operation τ is fixed, we will call S itself, the ternary ring. Thus by a ternary ring S, we will mean, a finite set S, with a ternary operation τ on it.

A ternary ring S is said to be a **planar ternary ring** if there are two special elements $0, 1 \in S$ and the following conditions are satisfied.

(A).$\tau(x,0,c) = \tau(0,x,c) = k$ for all $x, c \in S$.
(B). $\tau(x,1,0) = \tau(1,x,0) = x$ for all $x \in S$
(C). Given $x, y, m \in S$, there is a unique $c \in S$ such that $\tau(x,m,c) = y$.
(D) Given $m, c, k, p \in S,\ m \neq k$, there is unique $x \in S$ such that $\tau(x,m,c) = \tau(x,k,p)$.

Example 4: (A). Let S be a planar ternary ring. Construct a Design $D = (X, f)$, in exactly the same manner as we did, while constructing Affine plane from a field in the previous Example. Only this time equations of lines not parallel to y-axis will be $y = \tau(x,m,c)$, instead of $y = mx + c$. It can be easily seen that the design we get is an Affine plane. A projective plane can always be obtained from Affine plane. This construction was given by Hall in his paper in 1943 (See Hughes and Piper (1970), ChapterV).

(B). Conversely given any four points in a projective plane D of order n, no three of which are collinear, using them we can construct a planar ternary ring $S, |S| = n$, such that D is coordinatized by S, as described in (A) above and further all points on the line with slope 1 are of the type $(x, x),\ x \in S$. For more details see Hughes and Piper (1970), Chapter V, Hall's method.

This example shows that studying projective planes and planar ternary rings is essentially the same thing. Thus the problem can be studied as a geometric problem or algebraic problem. We will use word **PTR** for a planar ternary ring.

Two PTR S_1 and S_2 are said to be **isotopic** if they coordinatize the same projective plane. Unlike the field case, isotopic here does not imply isomorphic.

Let us define a few more terms for a PTR to understand them better and also to see that how they behave almost like fields and yet are very different too. Let S be a planar ternary ring. Let $x, y \in S$. Define $x + y = \tau(x,1,y)$ and $xy = \tau(x,y,0)$. When S is a field, *i.e.*, $\tau(x,m,k) = xm + k$ where addition and multiplication are the field operation, clearly the above definitions of addition and multiplication in this case, are the same as the addition and multiplication in the field. Also, When S is a field, S is a group under addition, the additive group and $S^* = S/\{0\}$ is a group, the multiplicative group of the field. This is not true when S is not a field, addition or multiplication may not be associative, in a general PTR. However in every PTR S, both $(S,+)$ and $(S^*,.)$ are loops under the addition and multiplication, as defined above. We will call $(S,+)$, the **additive loop** of the planar ternary ring S and similarly $(S^*,.)$, the **multiplicative loop** of S. A PTR S is said to be **linear**, if $\tau(x,m,c) = xm + c$, for all $x, m, c \in S$. A PTR S is called a **quasifield** if the additive loop (S,+) is a group, S is linear and satisfies left distributive law, *i.e.* $a(b+c) = ab + ac$ for all $a, b, c \in S$. A quasifield satisfying right distributive law also is called **division ring**. Planes coordinatized by quasifields are very special, they are called **translation planes**. We will describe group theoretic and geometric significance of them.

Let us first see in terms of desarguesian configurations. Let H be a point of a projective plane and ℓ be a line. A projective plane is said to be (H,ℓ)**- desarguesian** if for every pair of triangles ABC and abc which are centrally perspsective with H, as the centre of perspectivity (see above figure of desargues's configuration) and ℓ as the possible axis of perspectivity, *i.e.*, any two of the points $A',\ B',\ C'$ are on ℓ, the two triangles ABC and abc are also axially perspective and the third point is also on ℓ. Thus ℓ is the axis of perspectivity.

Thus being (H, ℓ)-desarguesian, essentially says that any two centrally perspective triangles with H as centre of perspectivity, and ℓ as "possible" axis of perspectivity, are actually also axially perspective, with ℓ as axis of perspectivity.

We now describe what is meant by (H, ℓ)-transitive. For a projective plane $D = (X, f)$, we will denote by $aut(D)$, the group of all automorphisms of D, i,e., permutations of the set X which take lines into lines. For any $x \in X,\ \sigma \in aut(D)$, we will denote by $\sigma(x)$, the image of x under σ. and similarly for a line ℓ of D, we will denote by $\sigma(\ell)$ the line, which is image of ℓ under σ. If $\sigma(x) = x$, the point x is said to be **fixed** by σ. If further $\sigma(m) = m$ for all lines m of the plane which are on x, then we say that point x is **fixed line-wise** by σ. We can similarly define a line ℓ to be **fixed point-wise**, if $\sigma(\ell) = \ell$ and $\sigma(x) = x$ for all $x \in \ell$. An automorphism $\sigma \in Aut(D)$ is said to be a (x, ℓ)**-perspectivity**, if σ fixes x line-wise and line ℓ point-wise.

Now the projective plane D is said to be $(\boldsymbol{H, \ell})$**-transitive**, if for all points $y, w \in X$ such that $x \neq y,\ x \neq w,\ y \notin \ell,\ w \notin \ell$ and x, y, w are collinear, there is an (H, ℓ)-perspectivity σ such that $\sigma(y) = \sigma(w)$. Thus (H, ℓ)-transitive essentially says that (H, ℓ)-perspectivities act transitively. A line ℓ is said to be a **translation line** of the projective plane D, if D is (H, ℓ)-transitive for all points $H \in \ell$. If D has a translation line then D is called a **translation plane**. The following two theorems show how closely projective geometry and algebra are related. They show properties of algebraic structures coordinatizing the plane, transitivity properties of automorphism groups and geometric properties like deasarguessian configurations occur mutually together, Hughes and Piper (1970), Chapter IV, Dembowski (1968)

Theorem 1: A projective Plane D is (H, ℓ)-transitive if and only if D is (H, ℓ)-desarguesian.

Theorem 2: A plane is coordinatized by a quasifield, if and only if it is a translation plane.

Almost every algebraic structure which coordinatizes a projective plane can similarly be related with some similar transitivity of an automorphism group as well as occurrence of desarguesian configurations. This relationship inspired a lot of work in this area during the second half of last century. Still many interesting research papers appear regularly studying such aspects. Another idea which similarly resulted in a lot of activity was started by Lenz and Barlotti. Now it is known as Lenz-Barlotti classification Dembowski (1968) pages 123-126. They looked at non-desarguesian planes, *i.e.*, those not coordinatized by a field. It has clearly then a non-desarguesian configuration. They classified all such non-desarguesian configurations, into different classes, which could occur in a plane. Assuming such structures they classified all planes in many of these classes. Many others also completed similar work for different such classes.

Still the basic problem I described above as conjectures remains. We will end this section with another a very different style of studying projective planes via PTR. The problem to classify the planes is the same as classifying PTR. These ternary rings have many properties similar to fields, for example any subring of a PTR is a PTR. Classifying finite fields has a very direct path, one starts with the ring of integers $\mathbb{Z}$ which may be considered as a free ring generated by 1 subject to the usual rules of commutativity, associativity, distributivity, linearity (one can think of $\mathbb{Z}$ as ternary ring, satisfying linearity as we defined earlier for fields). We will write cadl rules in short for these 4 rules. One looks at maximal

ideals in this free ring $\mathbb{Z}$, to get prime fields as quotients. Other finite fields are obtained then by considering polynomial rings over these prime fields. The polynomial rings also may be thought of as free rings generated by starting variables subject to cadl laws. One difficulty with general PTR is that even though they are very similar to fields, yet many of them satisfy none of these cadl laws. In the papers Singhi (2010) and Singhi (2009), some more general structures than PTR which are more like ring of integers, instead fields were defined. Free such structures were defined and constructed, which in a way corresponded to some kind of generalized ring of "Integers" and "polynomials" which did not satisfy cadl laws. Actual ring of integers or polynomials are quotients of these general ternary rings quotiented by ideals whose elements are all cadl laws. It was shown that every such PTR is quotient of a maximal ideal in these rings. Though it is not clear that all maximal ideals give PTR. The idea is to develop a language without using cadl laws to imitate classification theory of finite (or infinite) fields. In this connection it may be interesting to observe that Albert studied division rings in some what similar manner Albert (1961). He defined a "generalized twisting" of a field to get such division rings and conjectured that all division rings are obtained from a field in this manner. The conjecture is known to be true for dimension 3 or 4.

The language developed in above two papers though very general does not still include all PTR's. For example two isotopic PTR may have very different structure. In this connection it may be interesting to look at structures more general than Hall's PTR. Grari studies such general structures (see Grari (2004)).

3. t-designs

In this section we will review some basic construction methods and results on t-designs which evolved over last few decades and describe a generalization of t-designs to t-list designs, given in a joint paper of the author with Raychaudhuri, Singhi and Raychaudhuri (2012).

Let $D = (X, f)$ be a design. Thus $f \in N(X)$. Define a function $\boldsymbol{\partial_t}$: $N(X) \to N_t(X)$ as follows. Let $\partial_t(f)(T) =$ the number of blocks containing T, $T \in \mathbb{P}_t(X)$. Thus

$$\partial_t(f)(T) = \textstyle\sum_{T \in B} f(B), \text{ where the sum is over all } B \in \mathbb{P}(X)$$

The number of blocks in $D = \partial_0(f)(\emptyset)$, will be denoted by $\boldsymbol{b}$ or $\boldsymbol{b(D)}$.

If $\partial_1(f)(\{x\}) = \partial_1(f)(\{y\})$ for $x, y \in X$, we will denote the common value of $\partial_1(f)(\{x\})$, $x \in X$ by $\boldsymbol{r}$ or $\boldsymbol{r(D)}$.

We will denote by $\boldsymbol{\partial_{t,k}}$, the restriction of ∂_t on $N_k(X)$. Thus $\partial_{t,k} : N_k(X) \to N_t(X)$, is defined by $\partial_{t,k}(f)(T) = \sum_{T \in B} f(B)$, where the sum is over all $B \in \mathbb{P}_k(X)$.

A design $D = (x, f)$ is said to be a $\boldsymbol{t\text{-}(v, k, \lambda)\text{-design}}$ if D has, v points, *i.e.* $|X| = v$, block size in D is k and $\partial_t(f)(T) = \lambda$ for every t-subset T of X.

If $D = (X, f)$ is a signed design (or rational design), all the above terms signed t-(v, k, λ)-designs and rational t-(v, k, λ)-designs *etc.* are similarly defined in those cases too. In particular $\boldsymbol{\partial_t}$ and $\boldsymbol{\partial_{t,k}}$ are similarly defined over $M(X)$ and $M_k(X)$ (or for $V(x)$ and $V_k(X)$) also.

Remark 1: It is easy to see that if $D = (x, f)$ is a t-(v, k, λ)-design (or signed design or rational design) then
(A). $\partial_t(f)(W) = \lambda\frac{\binom{v-w}{t-w}}{\binom{k-w}{t-w}}$ for every w-subset W of X, $0 \leq w \leq t$.
(B). Thus in particular $b = \lambda\frac{\binom{v}{t}}{\binom{k}{t}}$ and $r = \lambda\frac{\binom{v-1}{t-1}}{\binom{k-1}{t-1}}$
(C). Thus a t-(v, k, λ)-design is also a w-(v, k, λ_w)-design with $\lambda_w = \lambda\frac{\binom{v-w}{t-w}}{\binom{k-w}{t-w}}$, $0 \leq w \leq t$.

When $t = 2$, a 2-(v, k, λ)-design is also called a **BIBD** (Balanced Incomplete Block Design). A BIBD is also called **BIBD$(\boldsymbol{v, b, r, k, \lambda})$** or **$(\boldsymbol{v, b, r, k, \lambda})$-design**. It is difficult to construct t-designs with $t \geq 3$, specially when parameters are small. Very few such designs with small parameters are known. And yet there are interesting results which show that all such designs, with v sufficiently large, exist. We will discuss these results, how they evolved over the decades.

A t-design with $\lambda = 1$ is also called a **Steiner system**, named after Swiss Mathematician Steiner, who studied them almost 200 years back. Most of the focus is on BIBD's, specially because they are very useful in Statistics, in Designs of Experiments. Perhaps Fisher and Yates were the ones, who formalized using such designs for designs of experiments. Main current interest arose with the 1939 paper of R.C.Bose , who was perhaps the first one to methodically study them by using algebra, geometry and number theory (see Bose (1939)). He used affine and projective planes, finite fields, difference sets *etc.* to construct them. Some good reference books are Raghavarao's book on designs and their applications in designs of experiments Raghavarao (1971), Colbourn and Dinitz Handbook of Combinatorial Designs Colbourn and Dinitz (2006), Beth Jugnickel and Lenz book on Design Theory Beth *et al.* (2000) and Peter Dembowski's book on finite geometries Dembowski (1968).

Example 5: (A). Suppose $D = (x, f)$ is a projective plane of order n, then it can be easily seen that D is a 2-$(n^2+n+1, n+1, 1)$ design, *i.e.* BIBD$(n^2+n+1, n^2+n+1, n+1, n+1, 1)$. Interestingly, thus in this design, the number of blocks, $b = v$, the number of treatments. A 2-(v, k, λ)-design, in which $b = v$, is called a **SBIBD$(\boldsymbol{v, k, \lambda})$, (symmetric BIBD).** Conversely every SBIBD with $\lambda = 1$ is essentially a projective plane.

One can similarly construct an SBIBD with higher lambda using projective spaces of higher dimensions. Blocks in these designs are the hyperplanes. In an SBIBD(v, k, λ), any two blocks intersect in exactly λ treatments. SBIBD's are very challenging objects of study. There are many unsolved problems about them. We briefly mentioned them in the Section 2 also, while discussing nets. The above example from projective planes shows that there are infinitely many SBIBD's with $\lambda = 1$. SBIBD's with $\lambda = 2$ are called biplanes.

Interestingly only finitely many SBIBD with a given λ are known for any $\lambda \geq 2$. There is a

well-known conjecture formulated by Marashal Hall Jr.

Conjecture. For any given integer $m \geq 2$, there are only finitely many SBIBD with $\lambda = m$.

(B). Now suppose $D = (X, f)$ is an affine plane of order n. Again it can be easily seen that D is a 2-$(n^2, n, 1)$-design, a BIBD$(n^2, n^2 + n, n + 1, n, 1)$. As we had noted in Section 2, blocks of this design can be partitioned into parallel classes.

A design $D = (X, f)$ is said to be **resolvable**, if blocks of D can be partitioned into parallel classes. Affine planes and nets are examples of resolvable designs.

A resolvable design is called an **affine design**, if any two blocks from different parallel classes intersect in exactly the same number of treatments. In an affine plane clearly they intersect in exactly one treatment. Thus an affine plane is also an affine design. Conversely every BIBD with $\lambda = 1$, which is an affine design, actually is obtained form the affine plane in this manner.

Affine spaces of higher dimension can also be used to form similarly affine designs. The hyperplanes of affine spaces form the blocks of such designs.

Thus in affine designs intersection number of two blocks takes only two possible values, one of which is 0. BIBD's in which blocks intersect in only two possible values are called **quasi-symmetric BIBD's**. These are sort of designs next best to symmetric BIBD's. As remarked in (A), blocks in a symmetric BIBD intersect in a unique value. qausi-symmetric BIBD's have been extensively studied. It has become a subject by itself. A good source for results and theory on quasi-symmetric designs is the book by M.S. Shrikhande and S.S. Sane (Shrikhande and Sane (1991)).

(C). In High school geometry we learn that there is a unique circle through any 3 noncollinear points in a real affine plane. Suppose we take all "'circles" in an affine plane and an extra point say ∞, added to every line of the plane, it is not hard to see these new extended lines together with circles considered as blocks, give us a 3-design, in the sense that any 3 points are on a unique block.

Some what similar construction can be carried out with an affine plane over a finite field. What we get 3-$(n^2 + 1, n + 1, 1)$-design as an extension of an affine plane of order n.

Formally such a design is constructed from an ovoid in a projective space of dimension 3 over a field of order n. An **ovoid** in this 3-dimensional projective space is a set of $n^2 + 1$ points, no three of which are collinear. It can be shown that every hyperplane of this projective space (it is actually a plane since we have taken projective space of dimension 3), intersects this ovoid in 0 or $n + 1$ points.

When we take all these sets of intersections with the ovoid of size $n + 1$ as blocks, we get a 3-$(n^2 + 1, n + 1, 1)$- design.

Any 3-$(n^2 + 1, n + 1, 1)$-design is called an **inversive plane**. There are some interesting unsolved problems associated with study of inversive planes. For more details see Beth *et al.* (2000) or Dembowski (1968).

Remark 2: Necessary conditions for existence of t-designs.
From Remark 1. It is clear that a necessary conditions for existence of t-designs is that

$$\lambda\binom{v-w}{t-w} = 0 \ (mod \ \binom{k-w}{t-w}), \ 0 \leq w \leq t$$

Remark 3: Basic problem in the theory of t-(v,k,λ)-designs.

A. **Existence Problem:**. Characterize all quadruples t-(v,k,λ) satisfying the necessary conditions of Remark 2. for which there exists a t-(v,k,λ) designs.

B. **Classifying problem:**. For a given t-(v,k,λ) satisfying necessary conditions, construct all non-isomorphic t-(v,k,λ)-designs.

Main effort in the subject has been to solve the Existence Problem. This general problem is quite hard. Even for very small parameters designs are not known, nor one can prove that they do not exist. Some examples of such parameters are 2-$(22,8,4)$-design, (BIBD(22,33,12,84)), 2-(157,13,1)-design (projective plane of order 12) or 6-designs with $\lambda = 1$ for small v. Even with best computers one can not do much in such cases. May be AI and simulations could be used to study such problems properly. There are 1000's of papers on this topic, still many designs in the useful range for Statistical studies are not known.

On the other hand, Bose's 1939 paper, Bose (1939), started a spurt in research activity of studies of t-designs, specially BIBD's, which still continues. Constructing new families of BIBD's whose existence is not known or which are not isomorphic to already known designs, still creates a lot of new interest in the subject.

Remark 4: Constructing *t*-designs.
Two types of types of methods are used generally to construct BIBD's or t designs.

(A). **Direct construction methods:**.
One constructs a new design or a new family of design directly by using some algebraic objects like difference sets, transitive permutation groups *etc.* Or one constructs them from geometric objects like projective spaces, affine space or ovals *etc.* Some examples we have described in the above Example 5. Bose himself gave some examples of such constructions in his paper. Among many others, who have given very interesting such constructions, included are S.S. Shrikhande, Marshal Hall, Wilson, Ray-Chaudhuri, Hanani (see Wilson (1972a), Ray-Chaudhuri and Wilson (1971), Wilson (1973),Ray-Chaudhuri and Singhi (1988), Colbourn and Dinitz (2006), Raghavarao (1971)).

(B). **Composition Techniques:**.
Smaller Designs are used to paste together a bigger design by using a base design. We will discuss some composition methods evolved, later in this section.

Let us just note here first that these studies had led to a conjecture, the so called the existence conjecture. It stated that if v is sufficiently large compared k and λ then necessary conditions of existence for a t-(v, k, λ)-design are sufficient. Conjecture was proved by Wilson in 1975 for the $t = 2$ case, *i.e.*, BIBD's, We will describe some details of his method later. Though many of Wilson's ideas were generalized for all t- designs. But the conjecture for $t \geq 2$, remained unsolved until 2014. The conjecture was proved in the general case by Keevash in 2014 by very different methods. He used probabilistic arguments to prove the conjecture. His method may be considered as a modification of the famous Rodl's nibble method (see Rodl (1985)). Though Keevash is able to get exactness of a very different order, which was needed to construct such exact designs. Keevash calls his method randomized algebraic construction (see Keevash (2014), Keevash (2015), See also an interesting lecture by Kalai, explaining Keevash's papers, Kalai (2015)). It is a bit amazing to see that probabilistic methods can give such exact geometric objects, even though v is large for such objects. Perhaps Kim and Vu were among the first ones to show such a potential of probabilistic methods in finding such exact constructions. They showed existence of small complete arcs in projective planes with high probability (Kim and Vu (2003)).

Though Keevash's theorem implies that all t-designs with sufficiently large v exist, still the existence problem in many of the general useful practical cases remains unsolved. There is a possibility that Wilson's method and composition techniques could be modified to get existence problems solved for practical cases. In fact in three interesting papers Blanchard proved some thing similar to existence conjecture for transversal designs or orthogonal arrays, using such methods (see Blanchard (1995b), Blanchard (1995a), Blanchard (1997)). We now describe in short how one of such basic composition technique evolved and many similar composition techniques were developed. Wilson also developed some of them. Such techniques formed the main core of his proof of the existence conjecture in the BIBD case.

A design $D = (X, F)$ is called a **PBD** (pairwise balanced design) with index λ if for all $x, y \in X$, the number of blocks of D containing x, y is λ. Thus $\sum_{x,y \in B} f(B) = \lambda$ for all $x, y \in X$. A PBD is similar to a 2-(v, k, λ)-design, only now the block size is not constant. Let us define for a PBD, $D = (X, f)$, of index λ, the set K (or $K(D)$) to be the set of all block sizes of D, *i.e.*, $K = \{k \in \mathbb{N} |$ there exists $B \in \mathbb{P}(X)$ such that $f(B) \neq 0$ and $|b| = k\}$. We will say that D is a $\mathbf{PBD}(\boldsymbol{v, K, \lambda})$ or a $\boldsymbol{(v, K, \lambda)}$**-design**.

PBD's were first defined by Bose and Shrikhande. They were interested in a famous problem on nets, the so called Euler's conjecture, posed almost 200 years back. The conjecture stated that there is no Net$(n, 4)$ when ever $n = 2(mod\ 4)$. Note that a net with 4 parallel classes corresponds to two mutually orthogonal latin squares. Thus Euler's conjecture was that there are no mutually orthogonal latin squares of order n, if $n = 2(mod\ 4)$. Euler became interested in this problem because of some arrangement of army regiments and ranks, Russian Czar had asked him to arrange. It corresponded to creating 2 mutually orthogonal squares of order 6. Euler could prove that no such mutually orthogonal squares of order 6 exist. He then conjectured the same for orders $n = 2(mod\ 4)$. Bose and Shrikhande proved that the conjecture is false.

Their method was to use smaller nets or planes and paste them together by using a PBD as a base. Crucial aspect in the construction was to use these designs with unequal

block sizes to get designs with equal block sizes. Before them Parker was also trying to study the same problem. He had also come up with similar construction but he was using projective or affine planes which have blocks with the same sizes. He came up with interesting results but could not prove falsity of Euler's conjecture. Later all three of them together proved that Euler's conjecture was only true for 2 and 6, it was false in all other cases (Bose and Shrikhande (1959) and Bose *et al.* (1960)). Later, Chowla Erdos and Strauss proved using similar compositions that if n is large and $r \leq n^{1/91}$ then a net $\mathrm{N}(n, r)$ exists. Wilson later improved this bound to $n^{1/17}$. Thus largest r for which $\mathrm{N}(n, r)$ exist, does not depend on prime power decomposition of n (see Chowla *et al.* (1960) and Wilson (1974)).

The composing bigger designs from smaller designs with nonconstant block sizes became an important technique to study different type of designs and arrays. In particular it helped in construction of many specialized designs and arrays, PBIBD (partially balanced incomplete block designs), Orthogonal arrays, association schemes, resolvable designs *etc.* Wilson and Ray-Chaudhari developed several such methods to solve the famous Kirkman's School Girl Problem, posed by Kirkman, almost 200 years back (see Ray-Chaudhuri and Wilson (1971), Beth *et al.* (2000)).

Later Wilson used all this development, to unify most of such work by then. He defined a very interesting closure operation **PBD closure** on any subset of $\mathbb{N}$, in the following manner. A set $K \subseteq \mathbb{N}$ is said to be PBD-closed if the existence of a $\mathrm{PBD}(v, K, 1)$ implies that $v \in K$. Let $K \subseteq \mathbb{N}$ and let $B(K) = \{v|$ there exists a $PBD(v, K, 1)\}$. Then $B(K)$ is a PBD-closed set, called the **closure** of K. Given any set K define $\beta(K)$ to be the $\gcd\{k(k-1)|k \in K\}$. Using this closure operation, Wilson proved the following interesting result in 1972. Every closed set K is eventually periodic with period $\beta(K)$. That is, there exists a constant C such that, for every $k \in K$, $\{v|v \geq C, v = k(mod\ \beta(K))\} \subseteq K$. What this theorem implies, for example, is that if $v = k\ (mod\ k(k-1))$ and is sufficiently large, then a 2-(v, k, λ)- design exists. In fact the result implied for many such congruent classes, the existence of BIBD's for all large v (see for more details Wilson (1972b), Wilson (1972c)). Ultimately by 1975, he proved the existence conjecture for BIBD case completely, (Wilson (1975)). We will describe basic steps in his proof.

Another tool which helped in construction of BIBD's, and more generally t-(v, k, λ)-designs was studying the structure of the module which is kernel of the mapping $\partial_{t,k} : M_k(X) \to M_t(X)$. Note that if $f \in ker(\partial_{t,k})$, $\partial_{t,k}(f)(T) = 0$ for all $T \in \mathbb{P}_t(X)$. Thus we can think of such an f as a signed t-(v, k, λ) design with $\lambda = 0$. Such signed t-$(v, k, 0)$ designs are called **null *t*-designs**. Thus $\ker(\partial_{t,k})$ is a $\mathbb{Z}$-module of all null t-designs. Its rank is clearly $\binom{v}{k} - \binom{v}{t}$. Constructing a natural basis $\ker(\partial_{t,k})$ acting on $M_k(X)$ or vector-space $V_k(X)$ helps a lot in developing a proper understanding of the signed designs. Graver and Jurkat and Wilson constructed such a basis. While Graver and Jukart studied it for module $M_k(X)$, Wilson studied over the vector space $V_k(X)$. Wilson actually proved that all t-$(v, k.\lambda)$-designs exist if λ is sufficiently large (see Graver and Jurkat (1973), Wilson (1973)).

These results were used by them to show that signed t-$(v, k.\lambda)$-designs always exist. Thus interestingly rational or signed t-(v, k, λ)-designs can be directly constructed by using such algebraic methods. Could a more careful study and better base or generating set for $\ker\partial_{t,k}$ or $M_k(X)$ itself, help in direct construction of t-designs? The method was used by

Ray-Chaudhri and Singhi to construct t-(v, k, λ) designs, for large λ and v, in which no block is repeated more than 2 times (see Ray-Chaudhuri and Singhi (1988)).

We describe an interesting natural set of generators of the $\mathbb{Z}$-submodule $\ker\partial_{t,k}$ of $M_k(X)$. This interesting generating set was first described by Graham Li Li, (see Graham *et al.* (1980)).

Let $X = \{x_1, x_2, \ldots, x_v\}$. Let $A = \{y_1, y_2, \ldots y_{2t+2}, w_1, w_2 \cdots w_{k-t-1}\}$ be a $(k+t+1)$-subset of X. Define a polynomial P_A by

$$P_A = (y_1 - y_2)(y_3 - y_4)\ldots(y_{2t+1} - y_{2t+2})w_1w_2\ldots w_{k-t-1}$$

Now define a function $f_A \in M_k(X)$ as follows. For a set $B \in \mathbb{P}_k(X)$, $B = \{q_1, q_2 \ldots q_k\}$, define $f_A(B)$ to be the coefficient of the monomial $q_1q_1\ldots q_k$ in P_A. Thus $f_A(B)$ is ± 1 or 0. Using the fact that there are $t+1$ brackets in the above expression of P_A, it can be easily seen that $\partial_{t,k}(f_A)(T) = 0$ for all $T \in \mathbb{P}_t(X)$. Thus $f_A \in \ker\partial_{t,k}$ is a null t-design. Graham Li and Li showed that such singed designs f_A generate the submodule $\ker\partial_{t,k}$ of $M_k(X)$. Chahal and Singhi, using these ideas, constructed a natural basis of the module $M_k(X)$ by using lexicographic ordering. Elements of this basis they called **tags** (see for more details Chahal and Singhi (2001), Singhi (2006)). Tags can be used to study many other problems too.

Wilson's proof of existence conjecture for BIBD can be summarized in a 3-step process.

(i). Existence theorem for signed designs:. Step 1 is to show that the necessary conditions are sufficient for signed t-designs or more general similar structures. This was done, as described above, first by Graver and Jurkar and Wilson.

(ii) λ large theorem: Step 2 is to prove that given v, t and k for all sufficiently large λ the necessary conditions are sufficient. This was proved by Wilson by studying his famous $W_{t,k}$ matrices and corresponding vector spaces, which he also used for solving many other interesting problems Wilson (1973).

(iii) Block spreading: Step 3 is to replace a set X in designs constructed in Step 2 with $X \times V$ for a large set V to reduce repetitions and create 2-designs on the set $X \times V$ with much smaller λ, for example a Steiner system. This was done by Wilson by taking V to be a vector space. The method is now known as Wilson's block spreading technique (see Wilson (1980) Wilson (1975) Wilson (1990)). As we already remarked, the method was later generalized for transversal designs for any t by Blanchard.

Finally we describe the generalization of t-designs to t designs for multisets (or lists, as we remarked in Section 1, two concepts lists or multisets are the same). These generalized designs should be useful in Statistics too. Also, another possibility is that Wilson's ideas of block spreading could be applied to them too, to get construction of actual t-designs.

We first define designs on multisets. A **list Design** is an ordered pair $D = (X, f)$,

where X is finite set and f is a list on $L(X)$. Thus for each multiset ℓ of X, $f(\ell) \in \mathbb{N}$. We may also consider f as a multiset $f = [\ell_i | i \in I_{|f|}]$. Each element of f is called a block of the list design f. Thus $\ell \in L(X)$ is a block if and only if $f(\ell)) \neq 0$. Elements of X are called, as in the case of sets, points or treatments. D is said to be of **block size k**, if all blocks are multisets of size k, i.e $f(\ell) \neq 0$ implies that $\ell \in L_k(X)$. We define ∂_t and $\partial_{t,k}$, for lists in quite similar manner, as we defined them for $N(X)$ and $N_k(X)$, only now they will be defined over $L(X)$ and $L_k(X)$ respectively. Thus, for example, if f is a list on $L_k(X)$, $\partial_{t,k}(f)$ is a list on $L_t(X)$, defined by $\partial_{t,k}(f)(s) = \sum_{s \subseteq \ell} c(\ell, s) f(\ell)$, for all $s \in L_t(X)$. Thus $\partial_{t,k}(f)(s)$, essentially gives the number of ways in which s occurs as a submultiset in the blocks of f.

For a finite set X we will denote by $\boldsymbol{S(X)}$ the symmetric group of all permutations of X. We note that $S(X)$ also acts as permutation group on the set of all k-subsets $\mathbb{P}_k(X)$ as well as on the set of all k-multisets $L_k(X)$. We also note that $S(X)$ acts transitively on $\mathbb{P}_k(X)$, *i.e.*, given any two k-subsets A_1, A_2 of X, we can always find an element $\sigma \in S(X)$, such that $\sigma(A_1) = A_2$. But this is not true with k-multisets.

Example 6: Let X be finite set. $x, y \in X$, $x \neq y$. Consider two 5-multisets $A_1 = [x, x, y, y, y]$ and $A_2 = [x, x, x, y, y]$. Define a permutation $\sigma : X \to X$ by $\sigma(x) = y$, $\sigma(y) = x$ and $\sigma(z) = z$, if $z \neq x$ or y. Then clearly $\sigma(A_1) = A_2$. Now consider the 5-multiset $A_3 = [x, y, y, y, y]$. It can be easily seen that there is no $\tau \in S(X)$ such that $\tau(A_1) = A_3$. Thus in general $S(X)$ is not transitive on $L_k(X)$.

Let $\ell \in L_t(X)$ we will denote $\boldsymbol{orb_t(\ell)}$, the orbit of $\ell \in L_t(X)$ under $S(X)$. Thus $orb_t(\ell) = \{\ell_1 | \sigma(\ell) = \sigma(\ell_1)$ for some $\sigma \in S(X)\}$. Let $\boldsymbol{ORB_t(X)}$ be the set $\{orb_t(\ell) | \ell \in L_t(X)\}$ of all orbits of elements of $L_t(X)$.

Let $m \in \mathbb{N}$. a **partition** π of m is a list on the set $I_m = \{1, 2, \ldots m\}$ such that $\sum i\pi(i) = m$, where the sum is over all $i \in I_n$.

Example 7: Consider the list π on the set I_{13} defined by $\pi(1) = 3$, $\pi(2) = \pi(3) = 2$ and $\pi(g) = 0$ if $g \neq 1, 2, 3$. π corresponds to the multiset $[1, 1, 1, 2, 2, 3, 3]$. Clearly π is a partition of 13.

Now suppose $\ell \in L(X)$. Define a partition of $\boldsymbol{\pi(\ell)}$ of integer $|\ell|$ by $\pi(\ell)(i) = |\{x \in supp(\ell) | \ell(x) = i\}|$. Thus $\pi(\ell) = [\ell(x)\ x \in supp(\ell)]$.

Remark 5: Suppose $\ell, \ell_1 \in L_t(X)$. Then, it can be easily seen that $orb_t(\ell) = orb_t(\ell_1)$, if and only if $\pi(\ell) = \pi(\ell_1)$.

We can now define t-list designs. Let $0 \leq t \leq k$, $|X| = v$. A ***t*-list design** on X with block size k is a list design $D = (X, f)$, with block size k such that for all $s_1, s_2 \in L_t(X)$ with $\pi(s_1) = \pi(s_2)$, $(\partial_{t,k}(f))(s_1) = (\partial_{t,k}(f))(s_2)$. Thus t-list design with block size k is a list design with block size k on the set X such that if any two t-lists s_1, s_2 on X are in the same orbit under the action of $S(X)$, then they occur the same number of times in blocks of D. We define t-list designs also in terms of parameters, only note that now λ is not a constant, it is a function on $ORB_t(X)$.

Let $\lambda : ORB_t(X) \to \mathbb{N}$ be a list on $ORB_t(X)$. A list design $D = (X, f)$ on a set X of

size v and block size k is said to be a t-(v, k, λ)-list design if for all $s \in L_t(X)$, $\partial_{t,k}(f)(s) = \lambda(orb_t(s))$.

In the paper Singhi and Raychaudhuri (2012), list designs, signed list designs, rational list designs are studied, the concept of tags is extended to list designs. Signed list designs are constructed for all parameters. And similarly second step in Wilson's three step process described above, of creating list designs when λ is large for all orbits, is carried out. some ideas of block spreading are also discussed.

Acknowledgements

The paper is based on Prof Aditya Shastri Memorial lecture delivered by the author during the 26th Annual Conference of the Society of Statistics, Computer and Applications held at Banasthali Vidyapith during February 26-28, 2024. The author expresses his gratitude to Bansathali Vidyapith and the organizers of the conference for inviting him to deliver this special lecture.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Albert, A. A. (1961). Generalised twisted fields. *Pacific Journal of Mathematics*, **11**, 1–8.

Beth, T., Jugnickel, D., and Lenz, H. (2000). *Design Theory:Volume 2*. Cambridge University Press, Cambridge.

Blanchard, J. L. (1995a). A construction for orthogonal arrays with strength $t \geq 3$. *Discrete Mathematics*, **137**, 1–3.

Blanchard, J. L. (1995b). A construction for steiner 3-designs. *Journal of Combinatorial Theory, Series A*, **71**, 60–67.

Blanchard, J. L. (1997). *The existence of orthogonal arrays of any strength with large order.* unpublished manuscript.

Bose, R. C. (1939). On the construction of balanced incomplete block designs. *Annals of Eugenics*, **9**, 315–414.

Bose, R. C. (1963). Strongly regular graphs, partial geometries and partially balanced designs. *Pacific Journal of Mathematics*, **13**, 39–419.

Bose, R. C. and Shrikhande, S. S. (1959). On the falsity of Euler's conjecture about the non existence of two orthogonal Latin squares of order $4t + 2$. *Proceedings of National Academy Sciences, U.S.A*, **45**, 734–737.

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

Bose, R. C., Shrikhande, S. S., and Singhi, N. M. (1976). Edge regular multigraphs and partial geometric designs. *Proceedings Of International Colloquium In Combinatorial Theory 1973, Rome, Academia National Die Lincei*, , 49–81.

Bruck, R. H. (1963). Finite nets II - uniqueness and imbedding. *Pacific Journal of Mathematics*, **13**, 421–457.

Caliskan, C. and Moorhouse, E. G. (2011). Subplanes of order 3 in Hughes planes. *The Electronic Journal of Combinatorics*, **18**, 1–8.

Chahal, J. S. and Singhi, N. M. (2001). Tags on k-subsets and t-designs. *Journal Combinatorics, Information and System Sciences*, **36**, 33–51.

Chowla, P., Erdos, P., and Strauss, E. G. (1960). On the maximal number of pairwise orthogonal latin squares of a given order. *Canadian Journal Mathematics*, **12**, 204–208.

Colbourn, C. J. and Dinitz, J. H. (2006). *Handbook of Combinatorial Designs, 2nd ed.* Chapman- Hall.

Dembowski, P. (1968). *Finite Geometries*. Springer Verlag, NY, Berlin.

Graham, R. L., Li, S. Y. R., and Li, W. W. (1980). On the structure of the t-designs. *SIAM Journal on Algebraic and Discrete Methods*, **1**, 8–14.

Grari, A. (2004). A necessary and sufficient condition so that two planar ternary rings induce isomorphic projective planes. *Archiv der Mathematik*, **83**, 183–192.

Graver, J. and Jurkat, W. B. (1973). The module structure of integral designs. *Journal of Combinatorial Theory, Series A*, **15**, 75–90.

Hall, M. (1943). Projective planes. *Transactions of American Mathematical Society*, **54**, 229–277.

Hughes, D. R. and Piper, F. C. (1970). *Projective Planes*. Springer Verlag, NY, Berlin.

Kalai, G. (2015). *Designs exist-Peter Keevash.* https://www.bourbaki.fr/TEXTES/1100.pdf5.

Keevash, P. (2014). *The Existence of Designs*. arXiv:1401.3665.

Keevash, P. (2015). *Counting Designs*. arXiv:1504.02909.

Kim, J. H. and Vu, V. H. (2003). Small complete arcs in projective planes. *Combinatorica*, **23**, 311–363.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments*. John Wiley, NY.

Ray-Chaudhuri, D. K. and Singhi, N. M. (1988). The existence of t-designs with large v and λ. *SIAM Journal Discrete Mathematics*, **1**, 98–104.

Ray-Chaudhuri, D. K. and Wilson, R. M. (1971). Solution of Kirkman's schoolgirl problem. *Proceedings of the Symposia in Pure Mathematics, Combinatorics*, **19**, 187–203.

Rodl, V. (1985). On a packing and covering problem. *European Journal of Combinatorics*, **5**, 69–78.

Shrikhande, M. S. and Sane, S. S. (1991). *Quasi-symmetric Designs, (London Mathematical Society Lecture Note Series 164)*. Cambridge University Press.

Singhi, N. M. (2006). Tags on subsets. *Discrete Mathematics*, **306**, 1610–1623.

Singhi, N. M. (2009). Twisted representations of semiadditive rings. *Journal of Combinatorics. Information and System Sciences*, **34**, 241–254.

Singhi, N. M. (2010). Projective planes-I. *European Journal of Combinatorics*, **31**, 622–643.

Singhi, N. M. and Raychaudhuri, D. K. (2012). Studying designs via multisets. *Designs Codes and Cryptography*, **65**, 365–31.

Veblen, O. and Young, jr, W. (1938). *Projective Geometry Volume I and II.* Ginn and Company, Boston, NY, London, 1938.

Wilson, R. M. (1972a). Cyclotomy and difference families in elementary abelian groups. *Journal of Number Theory*, **4**, 17–47.

Wilson, R. M. (1972b). An existence theory for pairwise balanced designs I, composition theorems and morphism. *Journal of Combinatorial Theory, Series A*, **13**, 220–245.

Wilson, R. M. (1972c). An existence theory for pairwise balanced designs II, the structure of pbd-closed sets and the existence conjectures. *Journal Combinatorial Theory, Series A*, **13**, 246–273.

Wilson, R. M. (1973). The necessary conditions are sufficient for something. *Utilitas Mathematica*, **4**, 207–215.

Wilson, R. M. (1974). Concerning the number of mutually orthogonal Latin squares. *Discrete Mathematics*, **9**, 181–198.

Wilson, R. M. (1975). An existence theory for pairwise balanced designs III, proof of existence conjecture. *Journal Combinatorial Theory, Series A*, **18**, 71–79.

Wilson, R. M. (1980). *Thoughts on Spreading Blocks.* unpublished manuscript.

Wilson, R. M. (1990). A diagonal form for incidence matrices of t-subsets vs. k-subsets. *European Journal of Combinatorics*, **11**, 601–615.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 51-60

Forecasting Models for the Production of Walnut in Jammu and Kashmir - A Comparative Study

Nishant Jasrotia[1], Manish Sharma[1], Anil Bhat[2], Sunali Mahajan[1] and Shavi Gupta[1]

[1]*Division of Statistics and Computer Science, SKUAST-Jammu*
[2]*Division of Agril. Economics and ABM, SKUAST-Jammu*

Received: 30 April 2024; Revised: 03 June 2024; Accepted: 06 June 2024

Abstract

Horticulture is an important sector that contributes towards the economic growth of our country. The UT of Jammu and Kashmir is the largest producer of walnut in India and provides important source of livelihood for many people. The study aims to forecast the production of walnut for Jammu and Kashmir using Time series models. Therefore, Holt linear exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) model have been applied and it shows that ARIMA(1,2,1) is appropriate model for forecasting on the basis of minimum value of information criterion and maximum value of coefficient of determination as compared to other models. Based on the forecast provided by the proposed model, there is a projected 56.69 percent increase in walnut production for the year 2035 with respect to 2022. This increase is contingent upon policymakers implementing policies aimed at boosting production.

Key words: ARIMA; Walnut production; Holt's linear exponential smoothing model.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Jammu and Kashmir is northern most state of India. The Jammu and Kashmir holds important position in horticultural production. The contribution of horticulture in gross state domestic product (GSDP) is more than 9 percent as per Sharma *et al.* (2023). Horticulture is an important sector which provides employment to people. Growing dependence on service sector for employment has added a lot of burden on union territory so; horticulture development makes a strong case. Walnut productions are light demanding species and are drought tolerant. Drought tolerating nature of walnuts makes a special case for their cultivation in Jammu & Kashmir. Walnuts (genus Juglans) are plants in the family Juglandaceae. Walnut is believed to have originated in Iran and its surrounding areas and brought to Europe by Alexander the Great and from Europe it was brought to China. In India walnut was earlier confined to Jammu and Kashmir and in late century it was brought to Himachal Pradesh, Uttarakhand hills and expanding up to Darjeeling and Sikkim. Walnut is called by

Corresponding Author: Manish Sharma
Email: manshstat@gmail.com

different names in different parts of India. The most commonly used name is akhroot, while in Kashmir it is called dun. Walnuts became the viable horticulture industry in India since 1980s particularly in the valley of Kashmir (Pandey and Shukla, 2007). Jammu and Kashmir occupy almost 90 per cent share of walnut industry in India. According to provisional data from the National Horticulture Board, Jammu and Kashmir recorded a production of 206.43 thousand metric tons of walnuts, cultivated across 69.24 thousand hectares in the 2016-17 period. In contrast, the rest of India produced 21.8 thousand metric tons of walnuts, covering an area of 22.85 thousand hectares during the same period (Horticulture Statistics at a Glance 2017). Jammu and Kashmir have been declared as an Agri-Export Zone for walnuts as discussed by Shah *et al.* (2021). The major walnut production areas of Jammu and Kashmir are Anantnag, Kupwara, Kulgam, Budgam, Doda, Poonch, Kishtwar, Rajouri and Kathua. The demand of Kashmiri walnut is increasing rapidly which needs to bring more land under it and require a regular attention to this industry so that it can better flourish in the times to come. Forecasting is an important problem that spans many fields including business and industry, government, economics, environmental sciences, medicine, social science, politics, agriculture and finance. Forecasting problems are often classified as short-term and long-term. Short-term forecasting problems involve predicting events only a few time periods (days, weeks, and months) into the future and long-term forecasting problems can extend beyond that by many years. Short term forecasts required for activities that range from operations management to budgeting and selecting new research and development projects. Long-term forecasts impact issues such as strategic planning. Most forecasting problems involve the use of time series data which is a time-oriented or chronological sequence of observations on a variable of interest. It is a sequential set of data points, measured typically over successive times. The measurements taken during an event in a time series are arranged in a proper chronological order. The time series in general supposed to be affected by four main components, which can be separated from the observed data. These components are: Trend, Cyclical, Seasonal and Irregular components. A time series model is linear or non-linear whether the variable of interest is forecasted using a linear or non-linear combination of past value of the variable. The linear time series models are designed to model the auto-covariance structure in the time series. Some of the forecasting models like Exponential smoothing, Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA). The two popular sub groups of Linear time series models are the Autoregressive (AR) and Moving Average (MA) models. AR models combined with moving average (MA) model to form a general and useful class of time series models called the Autoregressive Moving Average (ARMA) models. Sharma *et al.* (2018) applied Box-Jenkins methodology to build Autoregressive Integrated Moving Average (ARIMA) model for monthly arrival of Rohu fish in Jammu region of J&K state among many models the best model obtained was ARMA (2, 2) on the basis of significance of model and parameters. Mahajan *et al.* (2020) applied ARIMA model for the production of Rice crop in India. The best model would be ARIMA(0,2,2) on the basis of minimum AIC and SBIC. Kumari *et al.* (2022) applied Exponential smoothing and ARIMA model for the area, production and productivity of total fruit crops in Gujarat.

2. Material and methods

This study used annual time series data of walnut production in MT from 1973 to 2022 (Directorate of Horticulture Kashmir). We have to find Instability Index to check if

the time series exhibit any trend. Cuddy Della Valle Index (CDVI) method has been used as proposed Cuddy and Della Valle (1978) for measuring the instability in time series data. It is measured through CDVI = CV $*\sqrt{1-\bar{R}^2}$ where, CV is the coefficient of variation in percent, and $\bar{R}^2$ is the adjusted coefficient of determination. An appropriate modeling technique has been used for the forecasting of walnut production. Box and Jenkins methodology or ARIMA modeling has been introduced by Box and Jenkins (1976) is commonly used for forecasting purpose. It combines the Autoregressive Process (AR) and Moving Average Process (MA). The structure of ARIMA model is; ARIMA (p, d, q), where p and q are the order of the autoregressive and moving average process respectively while d is the order of differencing. The mathematical form of ARIMA (p, d, q) model is:

$$Y_t = C + (\Phi_1 Y_{t-1} + \ldots + \Phi_{t-p} Y_{t-p}) + \epsilon_t(-\theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q})$$

where, C is the constant, Y_t is the data on which the ARIMA model is to be applied, $\Phi_1, \ldots \Phi t - p$ are AR coefficients, $\theta_1, \ldots, \theta_q$ are MA coefficients and e_t is the random error. However, the AR model of order p is $Y_t = C + (\Phi_1 Y_{t-1} + \ldots + \Phi_{t-p} Y_{t-p}) + e_t$. Similarly, the MA structure of order q is $Y_t\theta = C + \epsilon_t - (\theta_1 \epsilon_(t-1) + \ldots + \theta_q \epsilon_(t-q))$.

The three main stages of Box-Jenkins forecasting model are used, first is identification of the model or specification of the model second is estimating the parameters and third is diagnostic checking of the residuals and forecasting. Another most commonly used univariate time series forecasting technique is the exponential smoothing (ES). In this technique, forecasts are weighted averages of past observations, with the weights decaying exponentially as the observations get older. In other words, recent observations are given relatively more weight in forecasting than the older observations. Exponential smoothing method classified according to the type of component presented in the time series data. The current study exclusively employs a single exponential smoothing method, namely Holts linear trend exponential smoothing technique, utilizing time series data. Holt (2004) introduced an extension of simple exponential smoothing tailored to forecast data exhibiting a trend. This method entails a forecast equation and two smoothing equations.

Forecast equation $\hat{Y}_t = l_t + hb_t$

Level Equation $l_t = \alpha Y_t + (1-\alpha)(l_{t-1} + b_{t-1})$.

Trend Equation $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$.

where, Y_t, $\hat{Y}_t$ is observed and predicted value of series at time t, l_t and b_t are the estimate of level and trend of the series at time t. The α, β are the smoothing parameters for the level and the trend, $0 \leq \alpha, \beta \leq 1$.

Model selection is done on the basis of following measures:

- Akaike's information criterion (AIC) proposed by Akaike (1979) is a useful statistic for model identification and evaluation. It defined as $-2logL + 2n$, where, L is the likelihood function and n is the number of hyper parameters estimated from the model.
- Bayesian Information Criterion (BIC) also as known as Schwarz Criterion. Schwarz (1978) proposed the criterion from Bayesian likelihood maximization. And is defined as $SBIC = -2LogL + nLogT$ where, T is total number of observations.

- Coefficient of determination (R^2) Wright (1921) and calculated as $R^2 = 1 - RSS/TSS$. where, RSS is residual sum of square and TSS total sum of square Range of R^2 is 0 to +1.

- Mean Absolute Percentage Error: The mean absolute percentage error (MAPE) is one of the most popular measures of the forecast accuracy. It was used as the primary measure in the M-competition Makridakis *et al.* (1982). MAPE is defined as $MAPE = \frac{1}{N}\sum_{t=1}^{N}|\frac{A_t - F_t}{A_t}| * 100$ where, A_t is the actual value, F_t is the forecast value, N is the number of data points.

3. Results and discussion

The annual time series data of the walnut production in Jammu & Kashmir for the last fifty years are used in the forecasting model. The data has been segmented into two 25-year phases: 1972-1997 and 1998-2022, aiming to analyze trends and stability. Notably, the average production during the latter phase surpassed that of the former (refer to Table 1). A similar trend is evident in the standard deviation. Consequently, the instability index was higher during the initial phase (33.30 percent) compared to the second phase (12.49 percent), possibly attributed to government subsidies, support programs, and the use of high-quality seeds to promote walnut cultivation. The overall instability index of the walnut production was 32.62 percent clearly indicating the production instability. The increasing market demand for walnuts incentivized farmers in Jammu and Kashmir to ramp up production for greater profits. The Durbin-Watson test yielded a value of 0.257, indicating positive autocorrelation among the residuals, necessitating the utilization of time series models.

Table 1: Descriptive statistics of walnut production from 1973-1997 and 1998-2022

Period	Mean (MT)	Std. Dev. (MT)	Max.Value (MT)	Min.Value (MT)	Instability Index (%)	Durbin Watson test (Overall)
Phase I (1973-1997)	27954.20	18620.05	68880.00	10212.00	33.03	0.257
Phase II (1998-2022)	169281.00	75563.61	279422.00	74906.00	12.49	
Phase III (1973-2022)	98617.60	89786.98	279422.00	10212.00	32.62	

Both graphical and empirical methods have been employed for this investigation. The line chart presented in Figure (1) illustrates an upward trend in walnut production. Additionally, long-term patterns suggest that the data is non-stationary, with a mean production of 98,617.60 metric tons (MT) and a standard deviation of 88,884.50 MT.

Table 2: ADF test value of actual Series and second differenced series

Test Statistic	Actual Series		Second Differenced Series	
	Value	P-Value	Value	P-Value
ADF	1.05	0.99(NS)	-10.12	0.000**

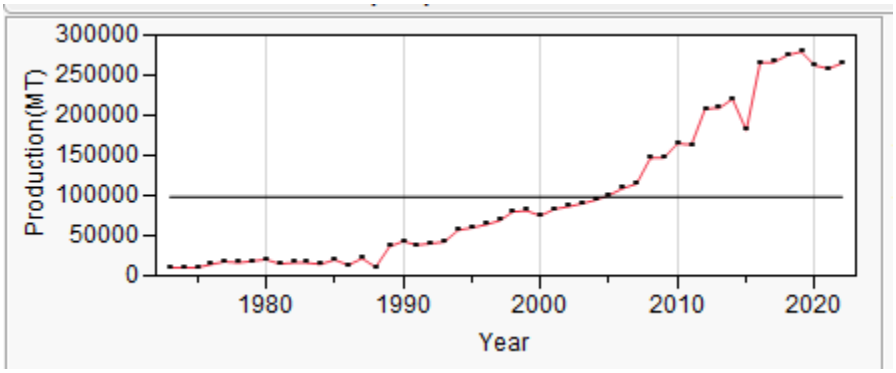


Figure 1: Trend of the annual production of walnut in Jammu and Kashmir

The Augmented Dickey-fuller (ADF) test value (see Table 2) is 1.05 and non-significant. It depicts non-stationarity of the data and p-values of Ljung-Box Q values are significant which means the residuals are dependent.

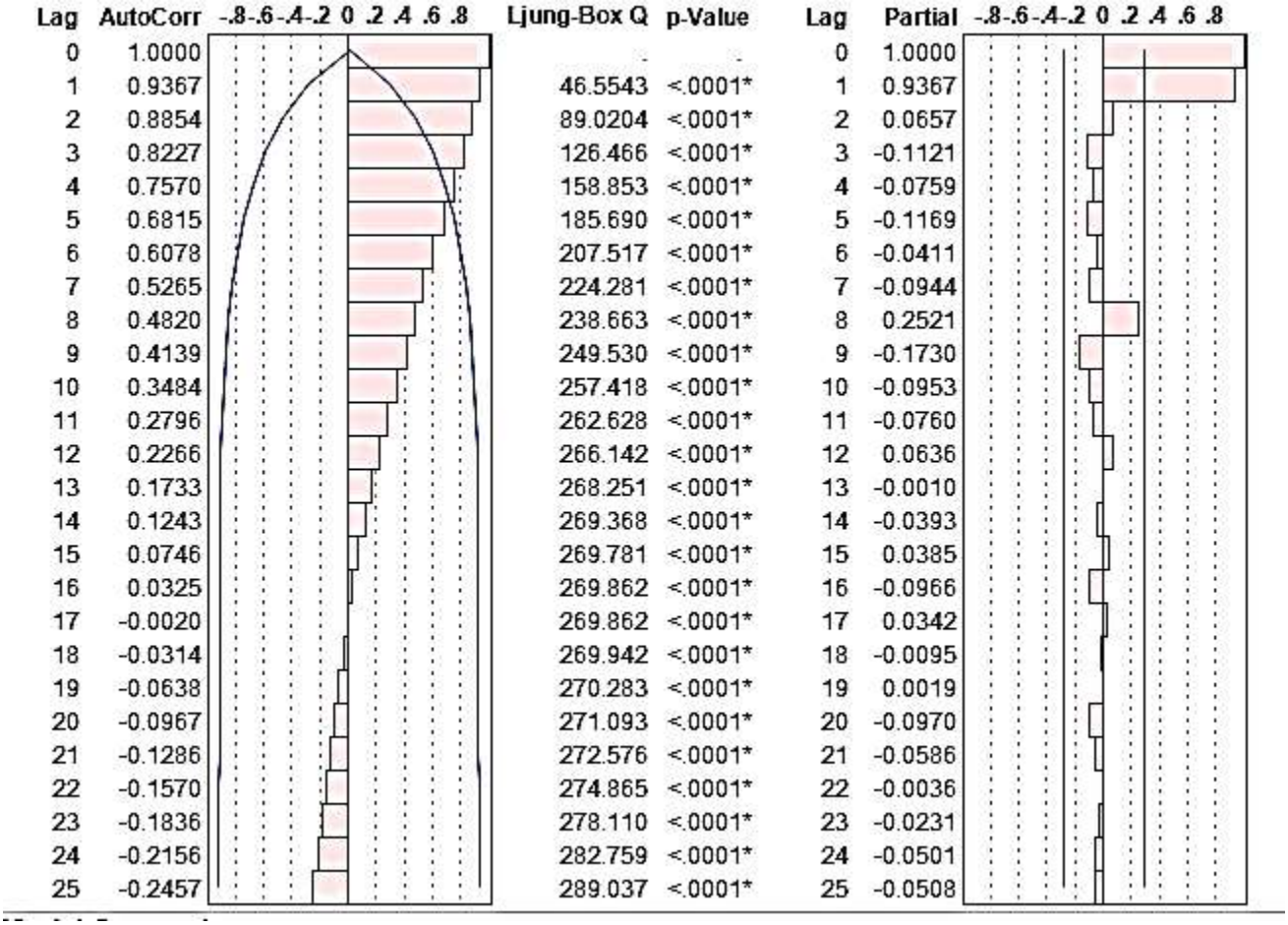


Figure 2: ACF and PACF plots of production of walnut in Jammu and Kashmir

The ACF and PACF plots in figure (2) show that the spikes are outside from the insignificant zone and fail to follow the assumption of randomness. Therefore, in order to met the stationary of the data first is to apply differencing method Mahajan *et al.* (2020).

The line chart, correlogram and ADF test have been used again after taking first and second order differencing until stationarity is achieved.

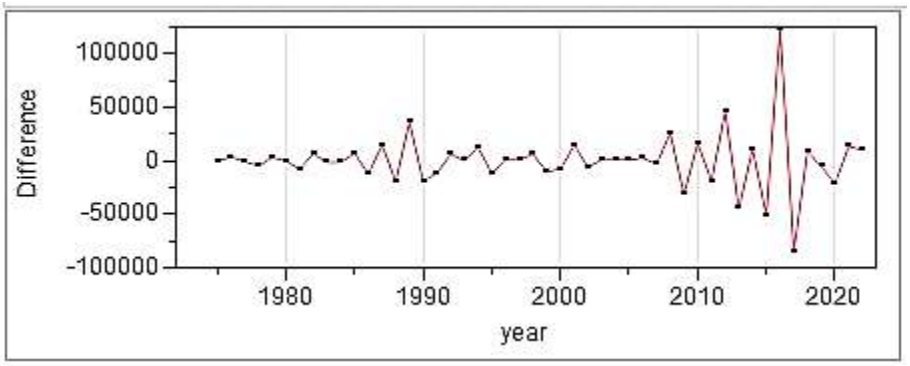


Figure 3: Trend of the data after taking differencing of order 2

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.7092	25.6865	<.0001*	1	-0.7092
2	0.3587	32.4010	<.0001*	2	-0.2903
3	-0.3187	37.8181	<.0001*	3	-0.4314
4	0.2804	42.1058	<.0001*	4	-0.3007
5	-0.1771	43.8559	<.0001*	5	-0.2354
6	0.1430	45.0240	<.0001*	6	-0.1323
7	-0.1375	46.1301	<.0001*	7	-0.1092
8	0.0920	46.6382	<.0001*	8	-0.1083
9	-0.0382	46.7281	<.0001*	9	-0.0290
10	0.0047	46.7295	<.0001*	10	-0.0744
11	0.0212	46.7585	<.0001*	11	-0.0093
12	-0.0263	46.8045	<.0001*	12	0.0148
13	-0.0108	46.8126	<.0001*	13	-0.0745
14	0.0087	46.8180	<.0001*	14	-0.1309
15	0.0390	46.9285	<.0001*	15	-0.0785
16	-0.0258	46.9782	<.0001*	16	-0.0174
17	-0.0549	47.2118	0.0001*	17	-0.1683
18	0.0698	47.6012	0.0002*	18	-0.1675
19	-0.0014	47.6013	0.0003*	19	-0.0647
20	-0.0171	47.6263	0.0005*	20	-0.0691
21	-0.0173	47.6529	0.0008*	21	-0.0736
22	0.0048	47.6551	0.0012*	22	-0.1024
23	0.0547	47.9419	0.0017*	23	0.0174
24	-0.0533	48.2260	0.0024*	24	0.0535
25	0.0446	48.4331	0.0033*	25	0.1658

Figure 4: Trend and correlogram of the data after taking differencing of order 2

The line chart in Figure (3) after taking the second order differencing shows that mean has no change and having constant variation. Thus, stationarity has been achieved through differencing. Moreover, ADF test value after second order differencing found to be significant which indicate that data are stationary. The order of (p) and (q), as depicted in Figure (4)

through the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), is determined to be one. In the ACF, only one spike falls outside the range, while in the PACF, two spikes extend beyond the range. This suggests that the data has achieved stationarity. Different ARIMA models have been developed on the basis of Box-Jenkins methodology. Among them five best models have been proposed on the basis of minimum AIC (Akaike Information Criterion), SBIC (Schwartz Bayesian Information Criterion) and R2. The model ARIMA(1,2,1) have minimum AIC(1068.32), and SBIC(1073.94) values with R2 (0.97) with significant parameters found to be the best. The estimates of the parameters are shown in Table (3) having AR(1) as -0.40. MA(1) 1.00 are found to be significant respectively.

Table 3: Different models for annual production of rice in India

Models	R^2	AIC	SBIC	Significance of Parameters/models
ARIMA(1,2,2)	0.95	1094.00	1101.54	Non-Significant
ARIMA(1,2,1)	**0.97**	**1068.32**	**1173.94**	**Significant**
ARIMA(2,2,1)	0.97	1092.63	1100.19	Non-Significant
ARIMA(2,2,2)	0.97	1072.00	1082.01	Non-Significant
ARIMA(1,2,3)	0.94	1093.62	1100.01	Significant

Table 4: Parameters estimates of ARIMA(1,2,1) for production of walnut

Terms with orders	Estimates	Standard Error	t-Ratio	P-Value
AR(1)	-0.40	0.12	-3.10	0.00**
MA(1)	1.00	0.06	15.85	< 00**
Intercept	202.30	108.03	1.87	0.06

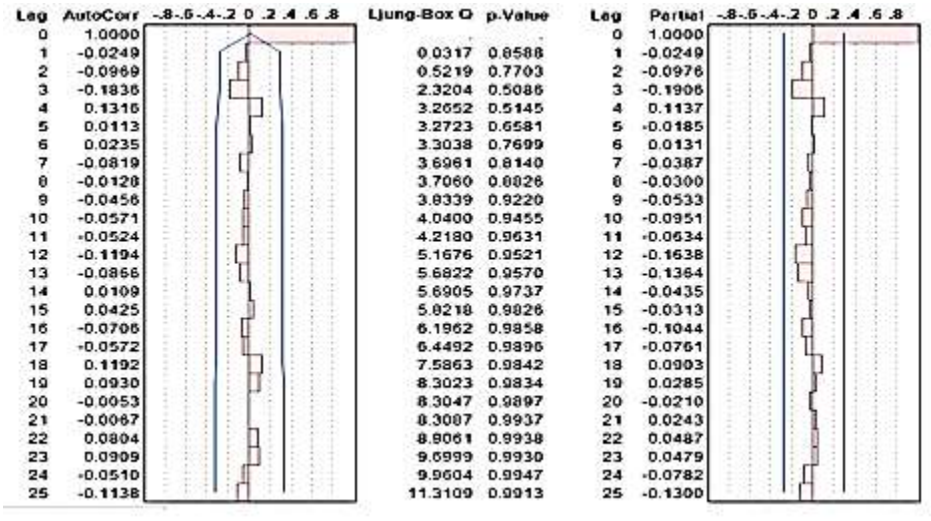


Figure 5: ACF and PACF plots of ARIMA(1,2,1)

The ACF and PACF plots in Figure (5) of the residuals indicate a good fit of the model, with p-values of the Ljung-Box Q test exceeding the significance level of 0.05. This

implies that the residuals are independent, satisfying the assumption of randomness. Meanwhile, the (Holt) linear exponential smoothing model is employed to estimate walnut production. It exhibits an AIC of 1068.71 and SBIC of 1074.12, with an R-squared value of 0.96. Table (5) displays the parameter estimates of the Holt model. The Level smoothing weight is calculated to be 0.59 and the Trend Smoothing Weight is 0.14. The Level smoothing weight is found to be significant, while the Trend smoothing weight is not significant so this model is not considered.

Table 5: Parameter estimates of the linear (Holt) ES model for production of walnut

Term	Estimate	Std Error	t-Ratio	Prob $>\|t\|$
Level Smoothing Weight	0.59	0.13	4.29	$<.0001^{**}$
Trend Smoothing Weight	0.14	0.13	1.05	0.29

Comparison of performance of model done on the basis of AIC,SBIC, R2, and MAPE. Table (6) shows that the ARIMA(1,2,1) have maximum R2, with minimum Mean Absolute Percentage Error as compared to Linear Holt Exponential Smoothing model. So, on the basis of that ARIMA(1,2,1) model is best model for forecasting of walnut production of Jammu and Kashmir.

Table 6: Comparison of performance of different fitted models

Model	Model Selection Measures						
	AIC	SBIC	R^2	MAPE %	Forecasted Value (2022)	Actual value (2022)	Difference
ARIMA	1068.32	1073.94	0.97	13.10	274397.68	265423	8974.68
Holt Linear Smoothing Model	1068.71	1074.12	0.96	13.97	276750.55	265423	11327.55

Thus, the proposed model for estimating walnut production is ARIMA(1,2,1), specified as: $\bar{Y} = 202.30 + (-0.40)[Y_t + Y_{t-1} + Y_{t-2}] + 1.000[\epsilon_t - \epsilon_{t-1} - \epsilon_{t-2}]$. For validation, data are splitted into training(80 percent) and testing (20 percent) and as per the result of Table (7) the testing has minimum RMSE as compared to the training. Thus the model is best fitted.

Table 7: RMSE of training and testing for ARIMA (1,2,1)

Model	RMSE	
	Training	Testing
ARIMA (1,2,1)	115859.98	96998.59

The graphical representation of forecasting of ARIMA(1,2,1) model in Figure (6), shows there is upward trend in the production of walnut in Jammu and Kashmir.

Table 8: Forecasting values for walnut production in Jammu and Kashmir

year	Forecasted values (MT)	U-95	L-95	Percentage increase in production on the basis of 2022
2025	297927.24	339403.54	256450.95	12.24
2030	354356.67	417331.60	291381.74	33.50
2035	415908.24	494827.67	336988.81	56.69

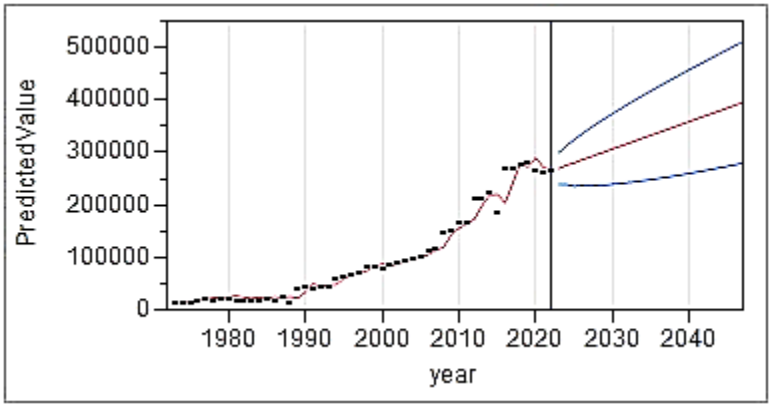


Figure 6: Forecasting graph of ARIMA (1,2,1) for annual production of Walnut

After conducting a forecast, the results indicate minimal variance between the actual and predicted values for the year 2022, affirming the accuracy of the model. The model's effectiveness is further validated by evaluating the lower and upper bounds of the forecasted values. Table (8) presents the projected walnut production for the years 2025, 2030, and 2035 based on this model. It illustrates a consistent upward trend in future walnut production. Specifically, the percentage increase in walnut production for 2025, 2030, and 2035 is 12.24%, 33.50%, and 56.69% respectively.

4. Conclusion

The current research aimed to forecast walnut production in Jammu and Kashmir, employing several time series models including linear Holt exponential smoothing and autoregressive integrated moving average (ARIMA). The findings suggest that the ARIMA model is the most appropriate for predicting walnut production. According to this model, the projected increase in walnut production for the year 2035 is 56.69%. This ARIMA model will be utilized for future walnut production forecasts, integrating up-to-date data.

Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am also thankful to the organizers of the Conference for giving me an opportunity to present my work.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

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

Box, G. E. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, volume 1.

Cuddy, J. D. and Della Valle, P. (1978). Measuring the instability of time series data. *Oxford Bulletin of Economics & Statistics*, **40**, 79–85.

Holt, C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages (onr memorandum no. 52). *International Journal of Forecasting*, **20**, 5–10.

Kumari, P., Parmar, D., Sathish Kumar, M., Mahera, A., and Lad, Y. (2022). Prediction of area, production and productivity of total fruit crops in Gujarat. *The Pharma Innovation*, **11**, 750–754.

Mahajan, S., Sharma, M., and Gupta, A. (2020). ARIMA modelling for forecasting of rice production: A case study of India. *Agricultural Science Digest-A Research Journal*, **40**, 404–407.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, **1**, 111–153.

Pandey, G. and Shukla, S. (2007). The walnut industry in India - current status and future prospects. *International Journal of Fruit Science*, **6**, 67–75.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **1**, 461–464.

Shah, R. A., Bakshi, P., Sharma, N., Jasrotia, A., Itoo, H., Gupta, R., and Singh, A. (2021). Diversity assessment and selection of superior persian walnut (juglans regia l.) trees of seedling origin from north-western himalayan region. *Resources, Environment and Sustainability*, **3**, 1–14.

Sharma, M., Jasrotia, N., Kumar, B., Bhat, A., and Mahajan, S. (2018). Modeling of monthly arrival of rohu fish using ARIMA in Jammu region of J&K state. *Journal of Animal Research*, **8**, 259–262.

Sharma, M., Singh, I. J., and Gupta, S. (2023). Horticulture in Kashmir valley: opportunities and challenges. *Current Agriculture Research Journal*, **11**, 1057–1067.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26 - 28 February 2024; pp 61-96

Modeling Multivariate Data Using Copula Theory: Analysis of an Environmental Dataset

Rahul Chatterjee[1] **and Nabendu Pal**[1,2]
[1]*Department of Mathematics,*
University of Louisiana at Lafayette, Lafayette, Louisiana, USA
[2]*Faculty of Mathematics and Statistics,*
Ton Duc Thang University, Ho Chi Minh City, Vietnam

Received: 13 May 2024; Revised: 02 June 2024; Accepted: 04 June 2024

Abstract

The normal distribution, though widely popular and heavily used in modelling datasets, has its own limitations, especially dealing with engineering and environmental data. In the univariate case, when the variable of interest is positively skewed, one can use a host of other distributions such as Gamma, Weibull, Lognormal *etc.*, just to name a few. However, in a multivariate set-up, the multivariate normal distribution appears to be the default choice, either by omission or by commission. The multivariate normal model has a host of advantages as its inferential problems are well studied, and the sampling distributions of its key statistics are relatively convenient to deal with. To be precise, the sample average follows a multivariate normal, and the sample cross-product matrix follows a Wishart distribution, and these two statistics are independent. Further, conditional expectation of any component given the remaining components is a linear function (of those remaining components) which is the foundation of the linear regression analysis.

But what happens if our multivariate data, which we commonly see in many applied problems, do not follow normal? The first casualty is the aforementioned mutual independence between the two commonly used statistics, let alone them being the minimal sufficient. Secondly, the linear regression model may not hold, thereby complicating the further conditional inferences. Also, multivariate normality forces one to assume marginally univariate normal distributions which may not seem reasonable as seen from the marginal empirical relative frequency histograms. One possible way out of this difficult situation is to transform the individual components to achieve multivariate normality, but this faces two big hurdles – (a) it would be an ad hoc approach to begin with; and (b) such ad hoc transformations may distort the natural association(s) among the components as well as the units being used, thus rendering the subsequent analyses questionable. On this backdrop, the copula theory comes handy in modelling multivariate data. Multiple individual components, apparently following skewed distributions, can be adequately combined by a suitable copula (also known as a link function) in order to model the given multivariate data. As opposed to the multivariate normal distribution's 'top-down' approach, the copula theory provides a 'ground-up'

Corresponding Author: Nabendu Pal
Email: nabendupal@gmail.com

approach where diversely distributed marginals can be combined into a suitable multivariate distribution for further inferences, including regressions.

Our study of the copula theory was motivated by an environmental dataset from the Mekong Delta Region (MDR) of Vietnam. In a bivariate set-up we have used a special copula, known as the Farlie-Gumbel-Morgenstern Copula (FGMC) to analyze the data. But this has also opened up a host of other research problems, such as estimation of the copula parameter, hypothesis testing, goodness of fit tests of FGMC, *etc.* Further, FGMC is just one of many, - possibly three dozen copulas, and thus this is a very rich emerging research field which has received relatively less attention, but has tremendous implications in 'Big Data' or 'Data Analytics'. We will also discuss some of the major challenges in copula theory which are related to heavy yet efficient computations in a reasonable amount of time. Thus this is a rich research area where experts in efficient algorithms and/or numerical analysis are very much welcome.

Key words: Copula; Jeffrey's prior; Parametric bootstrap method; Prediction mean absolute error; Prediction root mean squared error; Kolmogorov-Smirnov statistic.

AMS Subject Classifications: 62F10, 62F15, 62C05

1. Introduction

1.1. Why copula?

The normal distribution, though widely popular and heavily used in modelling datasets, has its own limitations. In the univariate case, when the variable of interest is positively skewed, one can use a host of non-normal distributions such as Gamma, Weibull, Lognormal *etc.*, just to name a few. However, in a multivariate set-up, the multivariate normal distribution appears to be the default choice, either by omission or by commission. The multivariate normal model has a host of advantages as its inferential problems are well studied, and the sampling distributions of its key statistics are relatively convenient to deal with. To be precise, let $\mathbf{X} = (X_1, X_2, ..., X_p)'$ be a p-variate random vector whose distribution is assumed to be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} = ((\sigma_{ij})) > 0$ (p.d.). Based on a random sample $\mathbf{X}_i$, $1 \leq i \leq n$, (*i.e.*, n copies of $\mathbf{X}$), assuming $n > p$, the maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively $\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$, and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/n$, where $\boldsymbol{S} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})'$. Further, under the above normality of $\mathbf{X} = (X_1, ..., X_p)'$, it is well known that $E(X_1|X_2, ..., X_p) = \beta_1 + \sum_{k=2}^p \beta_k X_k$, for suitable value of $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ which depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and this is the motivation behind the usual multiple linear regression where X_1 is intended to be explained as a linear function of $(X_2, ..., X_p)$ subject to some variation. But what happens if $\mathbf{X}$ does **not** follow $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?
(a): Can we have the aforementioned $\overline{\mathbf{X}}$ and $\boldsymbol{S}/n$ as the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$?
- Possibly not.
(b): Can we have the independence of $\overline{\mathbf{X}}$ and $\boldsymbol{S}$ (which is the foundation of most of the normality based inferential results)?
- Most likely not.
(c): Does regressing X_1 on $(X_2, ..., X_p)$ through a linear function make sense?
- Doesn't seem so, since $E(X_1|X_2, ..., X_p)$ may not be linear at all if the distribution of $\mathbf{X}$ is non-spherically symmetric and/or does not follow homoscedasticity.

Also, if $\boldsymbol{X} = (X_1, ..., X_p) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then it yields $X_k \sim N(\mu_k, \sigma_{kk} = \sigma_k^2), 1 \leq k \leq p$. Thus, in dealing with a multivariate data set, if one assumes the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ model then inadvertently univariate normality is assumed for individual components, and this can be problematic. Yet, in multivariate modelling, ranging from psychology to anthropology, from agriculture to environmental science, especially in a 'Big Data' setting, multivariate normal distribution is being used hastily without paying closer attention to whether such model fitting is appropriate or not.

If the multivariate normal is found to be inappropriate for the data $\mathbf{X}_i, 1 \leq i \leq n$, then one may transform the variable(s) suitably hoping that the transformed data would follow normal. But there are two major issues with such transformations. There is no magic formula to tell us what transformation would be suitable for normality. Secondly, often such transformed variables are hard to interpret, and they lose significance to the original problem which gave rise to the dataset to begin with.

This study has been motivated by several datasets where component-wise histograms indicate that marginals are heavily skewed, and therefore the joint distribution of the marginals ought to be something other than a multivariate normal distribution (not even elliptically symmetric one). In such a situation, it makes sense to follow a 'ground-up' approach to build a multivariate model starting with marginals, rather than the 'top-down' approach of starting with a (questionable) multivariate model and then live with its consequences at the marginal level.

Copula theory is a convenient 'ground-up' approach where one theorizes a multivariate distribution for the random vector $\mathbf{X} = (X_1, X_2, ..., X_p)'$ based on the marginal of each $X_k, 1 \leq k \leq p$. This is based on the understanding that the desired joint distribution ought to obey a particular structure involving the marginals which we have much control over. The following subsection gives a brief introduction of the copula theory. The focus of this work is on the bivariate set-up; however, we may present some general multivariate results occasionally.

1.2. General copula framework

The path breaking theorem in Sklar (1959) plays the most important role in the Copula theory. In the simplest case of a bivariate distribution, it tells us that given a random vector (X_1, X_2) with absolutely continuous marginal cumulative distribution functions (cdfs), F_1 and F_2, with corresponding probability density functions (pdfs) f_1 and f_2 respectively, and its joint cdf denoted by F, with joint pdf f, there exist unique copula C (a functional), such that

$$\begin{aligned} F(x_1, x_2) &= C(F_1(x_1), F_2(x_2)), \\ i.e., f(x_1, x_2) &= \partial^2 F(x_1, x_2)/\partial x_1 \partial x_2 \\ &= C^{(x_1, x_2)}(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2), \end{aligned} \tag{1}$$

where $C^{(u,v)}(u, v) := \partial^2 C(u, v)/\partial u \partial v$.

In general, given a continuous random vector in p-dimension, *i.e.*, $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$, with marginal cdfs $F_k, k = 1, 2, ..., p$, if we use the transformations such that $U_k := F_k(X_k), k = 1, 2, .., p$, then we have $U_k \sim \text{Uniform}(0, 1), k = 1, 2, ..., p$. The copula function $C : [0, 1]^p \to$

$[0,1]$ is a joint multivariate cdf of $\boldsymbol{U} := (U_1, U_2, ..., U_p)'$, *i.e.*,

$$C(u_1, u_2, ..., u_p) = P(U_1 \leq u_1, ..., U_p \leq u_p). \tag{2}$$

The joint cdf of $\mathbf{X}$, denoted by $F(x_1, ..., x_p)$, can be given in terms of $C(u_1, ..., u_p)$. By Sklar's theorem there exists a unique copula C such that

$$F(x_1, x_2, ..., x_p) = C(F_1(x_1), F_2(x_2), ..., F_p(x_p). \tag{3}$$

Simply put, the copula C is viewed as a dependence structure among the marginal cdfs.

Since the inception of the copula idea, one can find several copulas in the literature such as Gaussian copula, Exponential copula, Clayton copula, Frank Copula *etc.*, just to name a few. Out of the many available copula structures we focus on the Farlie-Gumbel-Morgenstern copula (FGMC) (see Morgenstern (1956)). The following section gives a brief introduction about the joint distribution based on FGMC, henceforth referred to as Farlie - Gumbel - Morgernstern Distribution (FGMD). The main reason behind our choice of FGMC (and subsequently that of FGMD) is its simplicity. Moreover, the nature of our investigation is completely new, and to the best of our knowledge the type of our investigation has not been carried out for FGMC. Therefore, this work of ours can be used as a template of future research for all other copulas as needed.

1.3. Farlie - Gumbel - Morgenstern Copula (FGMC) and the resultant model

As mentioned in the earlier section, a host of Copula structures have been discussed in the existing literature and one can find an overview of the available copula structures in Nelsen (2007). Several bivariate and multivariate non-normal probability distributions based on copula structures can be found in Kotz *et al.* (2004).

Morgenstern (1956) first introduced the following bivariate probability distribution on the square $[-1,1] \times [-1,1]$ of the form

$$f(x_1, x_2) = \frac{1}{4}(1 + \lambda x_1 x_2), \tag{4}$$

where $|\lambda| \leq 1$ and $-1 \leq x_1, x_2 \leq 1$. Farlie (1960) further studied various standard correlation coefficients between X_1 and X_2 for the bivariate distribution in (4). The limitations that a bivariate normal distribution brings to a dataset were first pointed out by Gumbel (1960) while he constructed a bivariate distribution with exponential marginals using Morgenstern's underlying copula in (4).

The pdf of the bivariate Farlie-Gumbel-Morgenstern distribution (FGMD) with general marginals based on the FGMC is given by

$$f(x_1, x_2) = f_1(x_1) f_2(x_2)[1 + \lambda(2F_1(x_1) - 1)(2F_2(x_2) - 1)], \tag{5}$$

where $|\lambda| \leq 1$ is the *association* parameter, f_1, f_2 are the marginal pdfs of the components X_1 and X_2, with corresponding marginal *cdfs* F_1, F_2 respectively. The range of λ happens to be $[-1,1]$, similar to many common correlation coefficients.

As a special case of (5), D'este (1981) considered a special biavriate Gamma distribution with gamma marginals and studied the structures of the covariance, conditional

expectations as well as other distributional properties. Since the inception of the FGMC, it has undergone several modifications over the years leading to some wider family of FGMC by different researchers. All these modifications were done with the goal of capturing a wider range of dependence among the components through common dependence measures such as Pearson's Correlation Coefficient (ρ), Spearman's Correlation coefficient (ρ_s), Kendall's Tau (ρ_K), *etc*. In their modified FGMD Huang and Kotz (1999) showed that with a polynomial type single parameter extension of the FGMC with uniform marginals the maximal attainable range of ρ is $[-0.39, 0.333...]$. Bairamov and Kotz (2002) proposed a new generalization of FGMD by introducing new association parameters and were able to attain a maximal positive (Pearson's) correlation of $\rho = 0.5021$ for some specific values of the model parameters. All these generalizations were made to accommodate a larger spectrum of the Pearson's correlation coefficient values. However the Pearson's correlation coefficient measures the strength of linear relationship between the components; therefore, paying attention only to this aspect of dependency, at the cost of adding more parameters to the model, is a rather narrow approach. Amblard and Girard (2009) gave a new family of copulas by generalizing the FGMC and highlighted the main feature of the proposed family as to permit modelling of data with high positive dependency, in particular over the range of $\rho_s \in [-0.75, 1]$. Another new generalization of the $FGMC$ was put forward by Bekrizadeh *et al.* (2012) and they were able to show the usefulness of the proposed generalized model in data with high negative dependence value by showing the (Spearman's rank correlation) values of $\rho_s \in [-0.5, 0.43]$. All these generalizations were made by introducing new parameters which only adds to the complexity of the statistical inferences of the FGMD model.

1.4. A motivational example with a real life dataset

This work has been motivated by an excellent investigation carried out by Merola *et al.* (2015) where the researchers have presented, among other things, a useful dataset on arsenic (As) concentration as well as a few other apparently benign elements from a survey carried out in Dong Thap province within the Mekong Delta Region (MDR) of Southern Vietnam. The complete dataset is given in Appendix A.1.

Vietnam is one of the worst affected countries where arsenic contamination in groundwater is particularly worrisome in two areas, - The Red River Delta (RDR) in the northern part, and the Mekong Delta Region (MDR) in the southern part. The MDR is the most economically vibrant region of the country which comprises twelve southern provinces and one major city (Can Tho) municipality. The provinces adjacent to Mekong river and its distributaries have been witnessing a very high concentration of arsenic in groundwater which is caused by both natural as well as man-made factors as discussed below.

As mentioned at the beginning of this section, Merola *et al.* (2015) collected data on arsenic concentration in groundwater in two subregions within Dong Thap province of MDR. Dong Thap, along with An Giang and Long An, is one of the provinces bordering Cambodia that has a high level of arsenic and poses a public health hazard. Thus, measuring arsenic in groundwater and issuing guidelines if and when needed is of paramount importance for the local administrations to mitigate arsenic poisoning. However, measuring arsenic level frequently and accurately is a time consuming and/or expensive exercise. Therefore it would be of great help to all the stakeholders if the level (or concentration) of arsenic could be predicted from the other benign elements when it is established, based on some existing

survey data, that in certain region there is an association between arsenic and one or more benign element(s) which can be measured easily (and cheaply), often through user friendly devices.

1.5. Scope of this research

The initial exploratory analysis points towards the fact that the MDR dataset consists of components which, firstly, have distinct distributions over the two sub-regions as mentioned above and secondly, have mostly skewed marginal distributions. We further delve into the exploration of the nature of pairwise association present among the variables in this dataset. We employ the FGMD model for the purpose.

The flexibility of the copula structure lies in allowing the freedom of choice of the desired marginal distributions. Hence, the association parameter λ of the copula structure (5) becomes a pivotal parameter in conserving the dependency between the components. As a result, inferences on the association parameter λ in (5) is of paramount interest. The basis of this current work has been the FGMD given in (5) with the goal of studying the inferential aspects of the association parameter λ comprehensively, with known marginals.

The inferential aspect includes parameter estimation where we have discussed a host of estimators and recommend the most suitable ones. Secondly, we have studied the existence of association among the variates through hypothesis testing under the FGMD model. We proposed a family of parametric bootstrap (PB) tests which addresses the problem of $\lambda = 0$ vs $\lambda \neq 0$. Along with the regular asympototic tests, we have studied the proposed PB tests and have shown that they tend to attain the nominal level very accurately.

While various correlation measures reveal some interesting patterns in terms of association between Arsenic (As) and Chlorine (Cl), between As and Hydrogen Potential (pH), and between As and Redox Potential or level (Eh), they do not address the objective of this work, *i.e.*, predicting the value of As when a suitable covariate, which is known to be significantly associated with As, is known. For example, in the southern region, where As and Cl are apparently strongly associated, can we predict the value (or, do we know the expected value) of As when Cl is equal to, say, 10 ppm? The prediction problem which has been posed above can be answered only by fitting an appropriate bivariate probability distribution to the given data on two relevant variates.

Let us denote the variate As by Y for the time being, and its suitable covariate by X (where X can be either Cl, or pH, or Eh). (For convenience in notation, these three covariates can be denoted by X_1, X_2 and X_3, respectively.) We addressed the suitable distribution of (X, Y) through FGMD which fits the given data. Once that suitable distribution of (X, Y) fits the data, then we use the conditional distribution of $(Y|X)$ to draw inferences on Y when X is given. As noted earlier, the joint probability distribution of (X, Y) has to be a non-normal one because the univariate normality tests reject such a notion most of the time (six out of eight cases - four variables in two subregions).

Finally, one can raise the question of 'goodness of fit' (GoF) of FGMD. It is worthy of noting that there is no "one stop solution" for the goodness of fit problem for the host of available copula in the literature and it remains an open problem. Several goodness of fit tests are available across the literature but to the best of our knowledge there doesn't exist

one for FGMC which considers the parametric nature of the distribution. We have proposed and developed a novel data driven goodness of fit test for FGMD, which does not assume any known distribution of the test statistic under the null hypothesis. A detailed study of this goodness of fit test including the test procedures, as well as its performance is discussed which validates the application of FGMD model to the MDR dataset.

2. Point estimation of the association parameter

If one looks at the existing applications of the copula theory with real-life data sets then it becomes abundantly clear that the preferred estimator of the association parameter has always been the maximum likelihood estimator (MLE). But how good is the MLE? From an asymptotic point of view the MLE has nice tractable limiting distributional properties. But, for small to moderate sample sizes the performance of the MLE of the FGMD association parameter λ is totally unknown. Worse, the existing literature is completely silent on other possible estimators, especially the Bayes ones under noninformative priors. In a parametric set up, one should study various estimators of all the model parameters simultaneously which include the association parameter λ as well as other parameters of the marginal distributions. (For example, if one assumes a two parameter gamma model for each of the two marginals, then one ends up with a total five parameters.) It has been noted that estimating just the association parameter with known marginals itself is a research problem as it entails several point estimators with corresponding sampling distributions, followed by hypothesis testing which allows us to verify, under the FGMD assumption, whether the components are independent or not. The computational challenges that one faces with Bayes estimators in this simplistic scenario (*i.e.*, just for the association parameter) can be quite overwhelming. However, the simplistic model that we are using in this work can be applied in a totally non-parametric marginal set up where one can use the empirical marginal cdf of each component to replace the aforementioned known marginal, and then can proceed with the subsequent inferences. With that above objective in mind, the following subsections present a brief review of parametric estimation of the association parameter λ as available in the existing literature. Also, the following lemma will be useful in deriving Bayes estimators under noninformative priors.

Lemma 1: Based on the iid observations $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ from (5) with marginals f_1 and f_2 completely known, the Fisher information $I(\lambda)$ is given as $I(\lambda) = nI^0(\lambda)$, where $I^0(\lambda)$ is the Fisher information per observation $(FIPO)$, and

$$I^0(\lambda) = (1/4) \int_{-1}^{+1} \int_{-1}^{+1} u_1^2 u_2^2 (1 + \lambda u_1 u_2)^{-1} du_1 du_2. \tag{6}$$

Note that the $FIPO$ expression is free from f_1 and f_2. A further simplification yields

$$I^0(\lambda) = \sum_{m=0}^{\infty} \lambda^{2m}/(2m+3). \tag{7}$$

Note that the infinite sum in the above expression is convergent. Using that expression of the infinite sum, the final form of the FIPO, is given by

$$I^0(\lambda) = \{-\lambda + tanh^{-1}(\lambda)\}/\lambda^3, \tag{8}$$

where $tanh^{-1}(\lambda) = (0.5)log((1+\lambda)/(1-\lambda))$. See Chatterjee (2022) for the proof.

Remark 1: It is not at all surprising to see that the expression of I^0 in (7) or (8) is free from f_k's $(k = 1, 2)$. Since the marginals are assumed to be completely known, without any loss of information one can look at $Y_{ik} = F_k(X_{ik})$, $k = 1, 2$, $1 \leq i \leq n$. Note that Y_{ik}'s are iid Uniform$(0, 1)$. Each $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})'$ then follows the $FGMD(\lambda)$ with joint pdf, say $g(\boldsymbol{y}) = [1+\lambda(2y_1-1)(2y_2-2)]$ on the unit square $[0, 1] \times [0, 1]$. The transformation $\boldsymbol{X}_i \to \boldsymbol{Y}_i$ does not change the problem as far as inference on λ is concerned, and yields the $FIPO$ expression as stated above.

In the following subsections, we propose a wide variety of estimators of the association parameter λ based on n iid observations from (5) with known marginals f_1 and f_2.

2.1. Method of moment estimation

Method of moment estimator is attained essentially by equating the sample raw moment with the population moment. For the joint population moment, using the simple calculation of the expectation of the distribution in (5) and some further simplification lead us to the following form

$$E(X_1X_2) = E(X_1)E(X_2) + \lambda I_1 I_2, \tag{9}$$

where $I_k = \int_{-1}^{1}(u/2)F_k^{-1}((1+u)/2)\partial u$, $k = 1, 2$. For convenience define $\mu_k = E(X_k)$, $k = 1, 2$, *i.e.*, the means of the known marginals. Therefore from (9) it can be easily established that $Cov(X_1, X_2) = \lambda I_1 I_2$. For the method of moment estimator $\hat{\lambda}_{MM}$ we equate $\lambda I_1 I_2$ with the sample equivalent of $Cov(X_1, X_2)$ which is $(1/n)\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$ where $\overline{X}_k = (1/n)\sum_{i=1}^{n} X_{ki}$, $k = 1, 2$. Therefore,

$$\hat{\lambda}_{MM} = (nI_1I_2)^{-1}\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2). \tag{10}$$

2.2. Maximum likelihood estimation

For the brevity in derivation, let us denote $2F_k(x_{ik}) - 1 = G_k(x_{ik}), k = 1, 2$. The log-likelihood function of the data denoted by $l(\lambda)$ is as follows

$$l(\lambda) = C + \sum_{i=1}^{n} ln(1 + \lambda G_1(x_{i1})G_2(x_{i2})), \tag{11}$$

where C is a constant, free of λ. It is tempting to take derivative of $l(\lambda)$ and equating it with zero, *i.e.*,

$$\sum_{i=1}^{n} G_1(x_{i1})G_2(x_{i2})/(1 + \lambda G_1(x_{i1})G_2(x_{i2})) = 0, \tag{12}$$

to find the MLE of λ. But this can lead to a computational error as the solution may lie outside the parameter space which may go unnoticed in simulation studies. (We suspect that this issue may arise for other copula - based joint distributions as well, and may have gone unnoticed in applications.)

Theorem 1: The MLE of λ *i.e.*, $\hat{\lambda}_{ML}$ as it is called here, which maximizes $l(\lambda)$ in (11), exists, and it is unique.

Proof. See Appendix of Chatterjee (2022).

Remark 2: Define $a_i = G_1(x_{i1})G_2(x_{i2})$, $1 \leq i \leq n$, and $G_k(x_{ik}) = 2F_k(x_{ik}) - 1$, $k = 1, 2$. Let $h(\lambda) = \sum_{i=1}^{n} a_i/(1 + \lambda a_i)$, $\lambda \in [-1, +1]$. As seen from the details of the proof of the above theorem, $\hat{\lambda}_{ML}$ takes the following form

$$\hat{\lambda}_{ML} = \begin{cases} -1 & \text{if } h(-1) < 0 \\ \text{solution of (12)} & \text{if } h(-1) > 0 \text{ and } h(+1) < 0 \\ +1 & \text{if } h(+1) > 0, \end{cases} \tag{13}$$

Remark 3: It is further seen that if all the a_i's are > 0, which happens with probability $(0.5)^n$, then $l(\lambda)$ is monotonically increasing in λ. Hence $\hat{\lambda}_{ML}$ is $+1$. Thus, $\{(a_1, ..., a_n)|\, a_i > 0\, \forall\, i\} \subseteq \{(a_1, ..., a_n)|\, h(+1) = \sum_{i=1}^{n} a_i/(1 + a_i) > 0\}$. Similarly, if all the a_i's are < 0, which again happens with probability $(0.5)^n$, then $l(\lambda)$ is monotonically decreasing in λ. Hence $\hat{\lambda}_{ML}$ is -1. Thus, $\{(a_1, ..., a_n)|\, a_i < 0\, \forall\, i\} \subseteq \{(a_1, ..., a_n)|\, h(-1) = \sum_{i=1}^{n} a_i/(1 - a_i) < 0\}$. We will see later in our simulation study that $\hat{\lambda}_{ML}$ can take ± 1 with substantially high probabilities depending on the sample size as well as λ.

2.3. Bayes' estimators

For any suitable prior $\pi(\lambda)$ over the parameter space $[-1, +1]$, the posterior distribution of $(\lambda|data)$, denoted by $g(\lambda|data)$, is

$$g(\lambda|data) = \frac{\prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)}{\int_{-1}^{1}\prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)\partial\lambda}. \tag{14}$$

A natural choice of the prior for the association parameter is a modification of the beta distribution which is originally defined over the space $(0, 1)$. The beta-type prior density function defined over the parameter space $[-1, +1]$ is

$$\pi(\lambda) = 1/(2B(a, b))((1 + \lambda)/2)^{a-1}((1 - \lambda)/2)^{b-1}, \tag{15}$$

where a, b are the hyper-parameters.

The most common loss function for estimating a parameter is the usual squared error loss. However, when a parameter is restricted to a finite range, as we have here for the association parameter λ, a weighted quadratic loss is more meaningful which can assign a heavy penalty near the boundary. Hence, we consider a general structure of the loss function of the form

$$L(\hat{\lambda}, \lambda) = w(\lambda)(\hat{\lambda} - \lambda)^2, \tag{16}$$

where $w(\lambda)$ is a suitable weight function. In this work we are going to consider weight function $w(\lambda)$ of the form

$$w_\delta(\lambda) = (1 - \lambda^2)^{-\delta},\ \delta \geq 0. \tag{17}$$

Note that $\delta = 0$ leads to the usual squared error loss. For any $\delta > 0$, the loss (16) goes to ∞ as λ approaches ± 1 and $|\hat{\lambda} - \lambda| > 0$. In other words, a small deviation of $\hat{\lambda}$ from λ near the boundary can be very costly.

Under the general weighted quadratic loss (16), the general structure of the Bayes' rule is given as

$$\begin{aligned}\hat{\lambda}_B &= \frac{E(\lambda w(\lambda)|\lambda \sim g(\lambda|data))}{E(w(\lambda)|\lambda \sim g(\lambda|data))} \\ &= \frac{\int_{-1}^{1} \lambda w(\lambda) \prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)d\lambda}{\int_{-1}^{1} w(\lambda) \prod_{i=1}^{n}[1 + \lambda G_1(X_{i1})G_2(X_{i2})]\pi(\lambda)d\lambda}.\end{aligned} \tag{18}$$

With the special structure of $w(\lambda) = w_\delta(\lambda) = (1 - \lambda^2)^{-\delta}$, we are now ready to derive the Bayes' rule, denoted by $\hat{\lambda}_{B\delta}$ as follows.

In order to attain a tractable structure of the Bayes' rule, we resort to a simple algebraic manipulation within Equation (18). Let us focus on the term $\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2}))$ in the Equation (18). Recalling from Remark 2 that $a_i = G_1(X_{i1})G_2(X_{i2})$, the following product term can be rewritten as

$$\begin{aligned}\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2})) &= (1 + \lambda a_1)(1 + \lambda a_2)....(1 + \lambda a_n) \\ &= 1 + \lambda \sum_{i_1=1}^{n} a_{i_1} + \lambda^2 \sum_{(1 \le i_1 < i_2 \le n)} a_{i_1}a_{i_2} + ... \\ &... + \lambda^k \sum_{(1 \le i_1 < i_2 < ... < i_k \le n)} a_{i_1}a_{i_2}...a_{i_k} + ... + \lambda^n a_{i_1}a_{i_2}...a_{i_n}.\end{aligned} \tag{19}$$

Call $\sum_{(1 \le i_1 < i_2 < ... < i_k \le n)} a_{i_1}a_{i_2}...a_{i_k} = D_k$, $1 \le k \le n$, and define $D_0 = 1$. Therefore $\prod_{i=1}^{n}(1 + \lambda G_1(X_{i1})G_2(X_{i2})) = \sum_{k=0}^{n} \lambda^k D_k$. Hence, the Bayes' rule in (18) can be simplified as -

$$\hat{\lambda}_B = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1}(1 - \lambda^2)^{-\delta}\pi(\lambda)d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^k(1 - \lambda^2)^{-\delta}\pi(\lambda)d\lambda}. \tag{20}$$

Further, we will consider the special case of $a = b = d$, which implies a symmetric prior about 0. We are going to introduce the notation β as $\beta = d - \delta$ and the estimator (18) with the prior in (17) will be denoted as $\hat{\lambda}_{B\beta}$, *i.e.*,

$$\hat{\lambda}_{B\beta} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1}(1 - \lambda^2)^{\beta-1}d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^k(1 - \lambda^2)^{\beta-1}d\lambda}. \tag{21}$$

2.3.1. Special case of $\beta = 1$ (Bayes estimator under flat prior or BFP)

A particular case of interest is $\beta = 1$ which can happen if $\delta = 0$ and $d = 1$ or $\delta = 1$ and $d = 2$ *etc.* Since $\beta = 1$ (due to $\delta = 0$ and $d = 1$) also implies the Bayes' estimator under

the flat prior (FP) using the ordinary squared error loss function, we denote $\hat{\lambda}_{B1}$ as $\hat{\lambda}_{BFP}$ and is given by

$$\hat{\lambda}_{BFP} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1} d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k} d\lambda} = \frac{\sum_{k=0}^{n} (D_k/(k+2))\{1-(-1)^k\}}{\sum_{k=0}^{n} (D_k/(k+1))\{1-(-1)^{k+1}\}}. \tag{22}$$

2.3.2. Bayes' estimator under Jeffrey's prior (BJP)

Let us step back to the initial form of the Bayes' estimator as mentioned in equation (20). A natural non-informative prior is the Jeffrey's prior, denoted by $\pi_{JP}(\lambda)$, which is

$$\pi_{JP}(\lambda) \propto (I(\lambda))^{1/2},$$

where $I(\lambda)$ = Fisher Information of λ from a sample of size n. Hence, from (7), we have

$$\pi_{JP}(\lambda) \propto \sum_{m=0}^{\infty} \lambda^{2m}/(2m+3)$$

Therefore, the Bayes' estimator under Jeffrey's prior using $\delta = 0$ in the weight function in (18) and, denoting $\hat{\lambda}_{BJP}$, is given by

$$\hat{\lambda}_{BJP} = \frac{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k+1} (\sum_{m=0}^{\infty} \lambda^{2m}/(2m+3))^{1/2} d\lambda}{\sum_{k=0}^{n} D_k \int_{-1}^{1} \lambda^{k} (\sum_{m=0}^{\infty} \lambda^{2m}/(2m+3))^{1/2} d\lambda}. \tag{23}$$

2.3.3. Bayes' estimator under an approximate Jeffrey's prior (BAJP)

Note that in either of (20) or (18) the Bayes' estimator involves an infinite series. For the ease of simplification and being able to study the performance of a suitable Bayes' estimator analytically, we propose a simplistic approximation of the Jeffrey's prior which is given by $\sum_{m=0}^{\infty} |\lambda|^m/(2m+3)^{1/2}$. Also, note that this infinite series is convergent and has a finite value. In fact, the above series converges to $\sqrt{2}\Phi(|\lambda|, 1/2, 3/2)/2$, where $\Phi(x, y, z)$ is called the confluent hypergeometric function of the first kind (Abramowitz and Stegun (1964)), which is a function of x when y, z are held constants. Due to this fact, we can use this approximation as a new prior distribution. We call this as the approximate Jeffrey's prior and is given by

$$\pi_{AJP}(\lambda) \propto \sum_{m=0}^{\infty} |\lambda|^m/(2m+3)^{1/2}. \tag{24}$$

Hence, the Bayes' estimator with respect to (24), denoted by $\hat{\lambda}_{BAJP}$, is

$$\hat{\lambda}_{BAJP} = \frac{\sum_{k=0}^{n} \sum_{m=0}^{\infty} D_k (2m+3)^{-1/2} (1+(-1)^{k+1})(m+k+2)^{-1}}{\sum_{k=0}^{n} \sum_{m=0}^{\infty} D_k (2m+3)^{-1/2} (1+(-1)^{k})(m+k+1)^{-1}}. \tag{25}$$

The derivation of (25) is available in Chatterjee (2022).

2.4. Sampling distributions of various point estimators

One has to be extremely careful about obtaining the MLE by maximizing the log-likelihood function by differentiation which yields (11). However, the solution of this equation exhibit a tendency to go outside of the parameter space $[-1, +1]$, especially when true λ is near the boundary values, with a high probability. Therefore, $\hat{\lambda}_{ML}$ needs to be truncated at ± 1 which shows a high probability concentration (*i.e.*, high relative frequency in the simulation study) at the boundaries. This feature hasn't been discussed by other researchers earlier. As the sample size increases, this behavior of $\hat{\lambda}_{ML}$ diminishes considerably, especially for $n \geq 50$ (see Chatterjee (2022)).

On the other hand, the Bayesian estimators are always strictly within the parameter space compared to the traditional estimator MLE showing a bimodal trend while estimating λ close to the center of the parameter space, (see Figure 2.1 in Chatterjee (2022)). Figure 1 illustrates the simulated sampling distributions of the 4 estimators described earlier for a sample of size $n = 25$ based on 10^4 replications.

As a demonstration, we apply the FGMD model to the MDR dataset. This gives us an opportunity to estimate the pairwise association among the variates in our dataset.

Table 1: Estimates of the FGMD association parameter in two MDR subregions

Pair of Elements	North				South			
	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{BFP}$	$\hat{\lambda}_{BJP}$	$\hat{\lambda}_{BAJP}$	$\hat{\lambda}_{ML}$	$\hat{\lambda}_{BFP}$	$\hat{\lambda}_{BJP}$	$\hat{\lambda}_{BAJP}$
As vs Cl	0.085	0.053	0.064	0.108	−0.982	−0.621	−0.674	−0.81
As vs Eh	−1	−0.587	−0.646	−0.803	−1	−0.872	−0.892	−0.941
As vs pH	0.746	0.423	0.475	0.666	0.611	0.431	0.469	0.648

Remark 4: (a) With the application of FGMD(λ) we were able to estimate the underlying association among the pairwise variables using the four estimators. According to the estimates in Table 1 there exists a strong negative association between Eh and As in the northern region. $\hat{\lambda}_{ML}$ estimates the strongest negative association among the variables, followed by $\hat{\lambda}_{BAJP}$, $\hat{\lambda}_{BJP}$ and $\hat{\lambda}_{BFP}$. The highly negative association between As and Eh in the northern subregion, which was partially captured by the Spearman's and Kendall's, is ratified by the estimates of the association parameter of FGMD (see the details of the standard estimated correlation measures in Table 7 within Section 5).

(b) In the instance of As vs pH it is crucial to note that in the northern sub-region, Spearman's rho and Kendall's Tau contradicted pearson's correlation coefficient which showed a strong linear association. This is in agreement with our FGMD model.

(c) In the southern subregion, the standard correlation coefficients estimate a considerable negative linear association in (As, Eh). Although there is visible evidence of association present in (As, Eh) and to some extent in (As, Cl) but labeling it as a linear association will be an over simplification and inaccurate. The estimates in Table 1 of the association parameter λ shows a strong negative association in (As, Eh) and in (As, Cl). The MLE registers the strongest association among the variables (As, Eh) followed by $\hat{\lambda}_{BAJP}$, $\hat{\lambda}_{BJP}$

and $\hat{\lambda}_{BFP}$. The same holds for (As, Cl) as well. There is a positive association among the variables (As, pH) as estimated by all the standard correlation measures, reiterating the same phenomenon by FGMD.

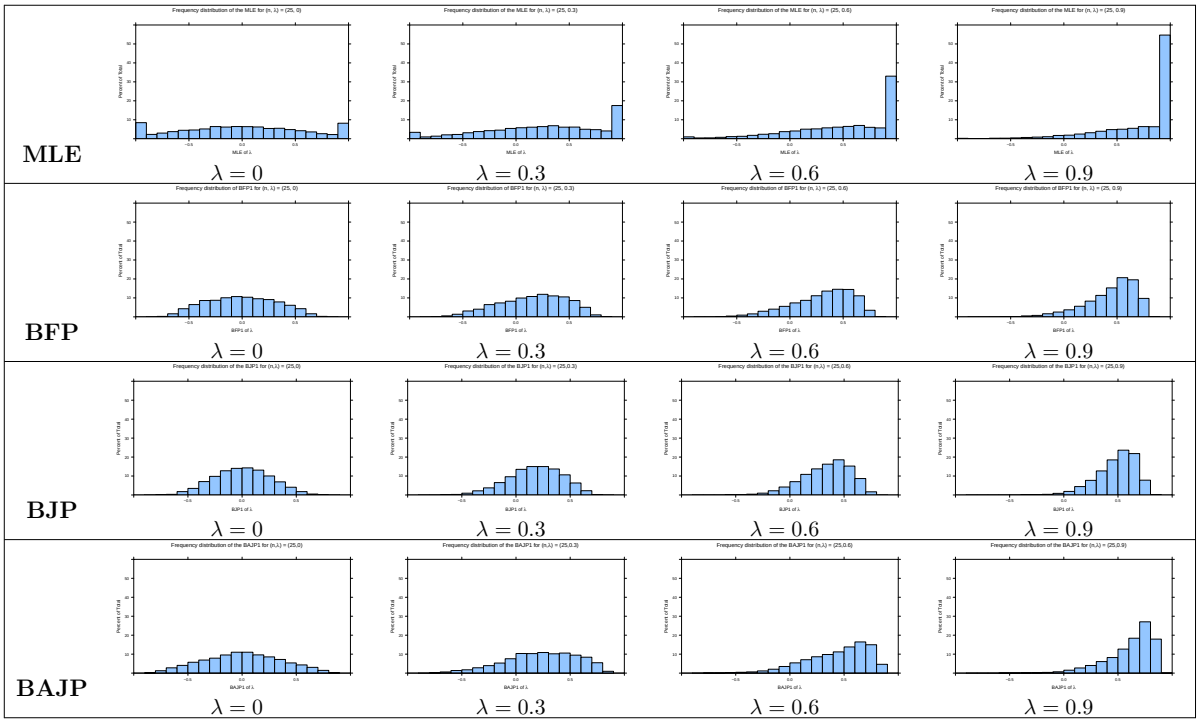


Figure 1: Simulated relative frequency histograms of four estimators of λ, $n = 25$

3. Hypothesis testing on the association parameter

3.1. The rationale behind hypothesis testing

In most of the applied cases one would be interested in knowing whether the components are associated or not. The copula based distributions such as FGMD which preserves the information of association through a single parameter λ (in the bivariate case), if proven to be suitable, can provide an answer to this problem. In this Section, we study the performance of different types of hypothesis testing procedures to test the hypotheses $H_0 : \lambda = \lambda_0$ vs $H_A : \lambda \neq \lambda_0$. Hence to examine whether the association indeed exists or not, one particular value of λ_0 is of interest, that is $\lambda_0 = 0$. The following tests have been proposed and studied through size and power for the aforementioned hypotheses.

1. **Asymptotic tests**:

   (a) Asymptotic Normal Test. (ANT)

   (b) Asymptotic Likelihood Ratio Test $(ALRT)$

2. **Parametric bootstrap tests based on the LRT statistic** $(PBLRT)$

3.2. Asymptotic normal test (ANT)

While testing the values of λ, this is probably the simplest approach of developing a hypothesis test utilizing the asymptotic property of the MLE of λ. Earlier we have already seen that for an iid sample of size n from FGMD(λ) the MLE exists and it is unique. It is a well known result that as $n \to \infty$, $\hat{\lambda}_{ML} \xrightarrow{d} N(\lambda,\ AV(\lambda))$, where $AV(\lambda)$ is the asymptotic variance of the MLE and is given by the inverse of the fisher information of λ, *i.e.* $I^{-1}(\lambda)$, assuming that marginals are fully known. Therefore, if we assume that the null hypothesis is true, then $\hat{\lambda}_{ML} \xrightarrow{d} N(\lambda_0, I^{-1}(\lambda_0))$ as $n \to \infty$. Therefore we reject the null hypothesis if

$$|\sqrt{nI^0(\lambda_0)}(\hat{\lambda}_{ML} - \lambda_0)| > z_{(1-\alpha/2)},$$

where $z_{(1-\alpha/2)}$ is the right tail $(\alpha/2)$ - probability cutoff point of the standard normal distribution and $I^0(\lambda_0)$ is the $FIPO$ in the relation $I(\lambda_0) = nI^0(\lambda_0)$.

3.3. Asymptotic likelihood ratio test ($ALRT$)

Based on the iid observations derive the likelihood ratio statistic Λ as

$$\Lambda = \frac{\underset{H_0}{Sup}\, L(\lambda|data)}{\underset{H_0 U H_A}{Sup}\ L(\lambda|data)} = \frac{L(\lambda_0|data)}{L(\hat{\lambda}_{ML}|data)}.$$

Define $\Lambda_* = -2ln(\Lambda)$. Asymtotically, as $n \to \infty$, $\Lambda_* \xrightarrow{d} \chi^2_1$ under H_0. So we reject the null hypothesis at level α if

$$\Lambda_* > \chi^2_{1;(1-\alpha)},$$

where $\chi^2_{1;(1-\alpha)}$ is the right tail (α) - probability cut off point of Chi squared distribution with 1 degree of freedom. The following Table 2 shows the simulated size values of the two asymptotic tests based on the MLE $\hat{\lambda}_{ML}$.

Table 2: Simulated size values of the two asymptotic tests for $\lambda_0 = 0, \alpha = 0.05$

Test	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 75$	$n = 100$
ANT	0.000	0.297	0.284	0.274	0.270	0.265	0.262
$ALRT$	0.002	0.017	0.026	0.028	0.028	0.027	0.026

Remark 5: Both the asymptotic tests are far from satisfactory as far as size is concerned. For $n = 10$, both of them are hopelessly conservative. For $n \geq 20$, ANT is overall a very liberal test and $ALRT$ on the other hand is a very conservative test. It is clearly visible that ANT has a monotonically decreasing (albeit very slowly) size property with the increase in sample size, whereas $ALRT$'s size values indicate a conservative behavior. Even for sample of size 100, which are generally considered to be 'large', these tests are still unable to achieve the level condition satisfactorily.

3.4. Parametric bootstrap (PB) tests

As seen in the earlier section, the asymptotic tests do not perform well for small to moderately large sample sizes. Therefore, in this subsection we propose a class of four tests based on the idea of LRT with the added parametric bootstrap (PB) concept.

The traditional LRT calls for using $\Lambda_* = -2\ ln(\Lambda)$ which, under H_0, follows χ^2_1 asymptotically. However, in this present $FGMD$ case, the null distribution of Λ_* has been found to be way off from the asymptotic distribution χ^2_1. Therefore, the cut-off point $\chi^2_{1;(1-\alpha)}$ is not applicable for the statistic Λ_* in order to test H_0. A situation like this calls for coming up with different cut-off points for Λ_* depending on sample size n as well as the data $\mathbf{X}$ through a PB method. Note that the expression Λ has $\hat{\lambda}_{ML}$ in the denominator as an estimator of λ while the numerator uses the null value λ_0 of λ. As a result, the value of Λ is always between 0 and 1, and a value of Λ closer to 1 implies a probable validity of H_0.

We extend the above traditional LRT concept a bit further by incorporating the other three estimates of λ which have shown considerable improvement over $\hat{\lambda}_{ML}$, especially in the mid region of the parameter space. In this regard we are going to consider $\hat{\lambda}_{BFP}$, $\hat{\lambda}_{BJP}$ and $\hat{\lambda}_{BAJP}$ (along with $\hat{\lambda}_{ML}$) in the LRT structure. In its generic form, the structure of Λ_* is going to be redefined as $\Lambda_*(\hat{\lambda}) = -2\ ln(\Lambda(\hat{\lambda}))$, where $\Lambda(\hat{\lambda}) = [\ L(\lambda_0|data)/L(\hat{\lambda}|data)\ ]$, where $\hat{\lambda}$ can be any one of the four aforementioned estimators of λ.

One difficulty with the above $\Lambda(\hat{\lambda})$ is that the denominator is not guaranteed to be greater or equal to the numerator unless $\hat{\lambda} = \hat{\lambda}_{ML}$. In other words, $\Lambda_*(\hat{\lambda}) = -2\ ln\Lambda(\hat{\lambda})$ is not guaranteed to be non-negative unless $\hat{\lambda} = \hat{\lambda}_{ML}$. However, a value of $\Lambda_*(\hat{\lambda})$ closer to 0 still conforms the validity of H_0. Therefore, to find suitable cut-off points for the statistic Λ_*, we consider

$$\Lambda_{**}(\hat{\lambda}) = |\Lambda_*(\hat{\lambda})|, \tag{26}$$

which is always nonnegative. The four versions of Λ_{**} using four aforementioned estimators will be referred to as

$$\begin{aligned}
\Lambda_{**1}\ (or PBLRT\ 1) &= \Lambda_{**}(\hat{\lambda}_{ML}) \\
\Lambda_{**2}\ (or PBLRT\ 2) &= \Lambda_{**}(\hat{\lambda}_{BFP}) \\
\Lambda_{**3}\ (or PBLRT\ 3) &= \Lambda_{**}(\hat{\lambda}_{BJP}) \\
\Lambda_{**4}\ (or PBLRT\ 4) &= \Lambda_{**}(\hat{\lambda}_{BAJP})
\end{aligned} \tag{27}$$

Algorithmic steps to implement $\Lambda_{**}(\hat{\lambda})$ as a test:

Step - 1: For the given data $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ from $FGMD$, compute $\hat{\lambda}$ (which is one of the above 4 estimators as mentioned earlier). Obtain the corresponding $\Lambda_{**}(\hat{\lambda})$.
Step - 2: Assume that $H_0 : \lambda = \lambda_0$ is true. Generate a bootstrap sample of size n (say, $\mathbf{X}^*_1, \mathbf{X}^*_2, ..., \mathbf{X}^*_n$) from $FGMD(\lambda_0)$. Once this bootstrap data is generated, pretend that λ is unknown, estimate λ using the bootstrap data by $\hat{\lambda}$, and call it $\hat{\lambda}^*$, which in turn produces the value of $\Lambda_{**}(\hat{\lambda}^*)$. (See Chatterjee (2022) about generating data from FGMD(λ).)
Step - 3: Repeat the above Step - 2 a large number of times (say, B times). This produces B copies of $\Lambda_{**}(\hat{\lambda}^*)$, and call them as $\Lambda^{(b)}_{**}(\hat{\lambda}) = \Lambda_{**}(\hat{\lambda}^{*(b)})$, $1 \le b \le B$, where $\hat{\lambda}^{*(b)}$ is the b^{th}

copy of $\hat{\lambda}^*$ as mentioned in Step - 2. These $\Lambda_{**}^{(b)}(\hat{\lambda})$ values are supposed to approximate the null distribution of $\Lambda_{**}(\hat{\lambda})$.

Step - 4: Order $\Lambda_{**}^{(b)}(\hat{\lambda})$, $1 \leq b \leq B$, and let $\Lambda_{**}(\hat{\lambda}|\alpha)$ be the $100(1-\alpha)^{th}$ percentile value of $\Lambda_{**}^{(b)}(\hat{\lambda})$, $1 \leq b \leq B$. This $\Lambda_{**}(\hat{\lambda}|\alpha)$ is the critical value for $\Lambda_{**}(\hat{\lambda})$ in Step - 1.

Step - 5: Reject H_0 if $\Lambda_{**}(\hat{\lambda})$ (from Step -1) $> \Lambda_{**}(\hat{\lambda}|\alpha)$, and retain H_0 if otherwise.

A complete comparison of the four PB test in terms of size and power, for different sample size is detailed in Chatterjee (2022)

3.5. Application to MDR dataset

Using the estimates of λ from the MDR dataset in Table 1, we proceed to perform $PBLRT1$, $PBLRT2$, $PBLRT3$ and $PBLRT4$ to test the hypothesis $H_0 : \lambda = 0$ vs $H_A : \lambda \neq 0$. The following Table 3 gives the $PBLRT$ test statistic values along with the simulated P-values.

Table 3: $PBLRT$ test statistic and their p-values for the MDR data

Test	North			South		
	As vs Cl	*As vs Eh*	*As vs pH*	*As vs Cl*	*As vs Eh*	*As vs pH*
$PBLRT1$	0.017	3.959*	1.366	4.447*	17.099*	1.364
	(0.899)	(0.047**)	(0.263)	(0.039**)	(0.000***)	(0.251)
$PBLRT2$	0.013	2.792	1.108	3.841	15.266*	1.262
	(0.208)	(0.101*)	(0.485)	(0.076*)	(0.002***)	(0.503)
$PBLRT3$	0.015	2.986	1.186	4.006	15.577*	1.305
	(0.216)	(0.101*)	(0.491)	(0.076*)	(0.000***)	(0.506)
$PBLRT4$	0.016	3.475	1.350	4.309	16.280*	1.353
	(0.239)	(0.098*)	(0.503)	(0.076*)	(0.000***)	(0.495)

Remark 6: The results of Table 3 show that As is not associated with pH in both the regions based on the FGMD model. However, As is significantly associated with both Cl and Eh in the southern subregion, and with Eh in the northern subregion thereby opening up the possibility of further prediction (see more in Section 5).

4. Goodness of fit tests for FGMD

4.1. The rationale behind goodness of fit (GoF) tests

Since the pathbreaking work of Sklar (1959) about three dozen copulas have been proposed by various researchers for different applications. A particular copula presents a particular family of multivariate distributions of the random vector $\mathbf{X}$ which combines p suitably hypothesized univariate marginal distributions of the components. Therefore, before adopting a particular copula for a specific dataset one must come up with a suitable GoF test for that copula, and this is where there appears to be an ample room for further research.

The problem of finding an optimal GoF for a given copula is an open problem. It appears that there does not exist a robust test which can identify the most appropriate

copula for a given dataset. Therefore, one can take each copula, from a handful of copulas and see their applicability by running a GoF for a given dataset. Most of the available GoF tests in the literature are developed with either a specific copula or a specific family of copula based probability distributions in mind. A brief review of the available GoF tests and their inadequacy is discussed in the following section.

4.2. Inadequacy in the existing literature

There are several works on GoF tests involving copulas, such as Fermanian (2005), Genest *et al.* (2006), Genest and Favre (2007), Genest *et al.* (2009) (which is primarily a review of the existing methods with a limited power study), Genest *et al.* (2011) (which is a goodness of fit test for the bivariate extreme value copulas). However Genest *et al.* (2006) appears to encompass the overall GoF test methods for copulas.

Genest *et al.* (2006) provided two test statistics that have been developed to test the GoF of a given copula. These two test statistics, say S_n and T_n, which are essentially Cramer-Von Mises and Kolmogorov-Smirnov statistics respectively, can be computed for FGMD through the following steps.

(i) Given the bivariate data X_{ik}, $i = 1, 2, ..., n$ and $k = 1, 2$, define the pseudo observations $V_1, V_2, ..., V_n$ as $V_i = (1/n)\sum_{l=1}^{n} I(X_{l1} \leq X_{i1}, X_{l2} \leq X_{i2})$, $1 \leq i \leq n$.

(ii) Define $K_n(t) = (1/n)\sum_{i=1}^{n} I(V_i \leq t) =$ (Number of V_i's $\leq t)/n$.

(iii) Define $K(t|\lambda)$ as

$$K(t|\lambda) = \int_0^t \int_s^1 h(x, s|\lambda)dxds,$$

where

$$h(x, s|\lambda) = \frac{1}{(1-x)r(x, s|\lambda)} + \frac{1}{x} - \frac{1}{(1-x)},$$

with

$$r(x, s|\lambda) = [\{1 - \lambda(1-x)\}^2 + 4\lambda(1-x)(1-s/x)]^{1/2}.$$

Note that while implementation of the GoF tests, λ in the above expression is to be replaced by a suitable estimate $\hat{\lambda}$.

(iv) Both $K_n(t)$ and $K(t|\hat{\lambda})$ are to be evaluated at (j/n) as well as $((j+i)/n)$ with $j = 0, 1, 2, ..., (n-1)$ and $i = 0, 1$, such that the test statistics S_n and T_n have the desired expressions as follow (Genest *et al.* (2006))

$$S_n = \frac{n}{3} + n\sum_{j=1}^{n-1} K_n^2(\frac{j}{n})\{K(\frac{j+1}{n}|\hat{\lambda}) - K(\frac{j}{n}|\hat{\lambda})\}$$
$$-n\sum_{j=1}^{n-1} K_n(\frac{j}{n})\{K^2(\frac{j+1}{n}|\hat{\lambda}) - K^2(\frac{j}{n}|\hat{\lambda})\}$$

$$T_n = \sqrt{n}\max_{i=0,1;0\leq j\leq n-1}\{|K_n(\frac{j}{n}) - K(\frac{j+i}{n}|\hat{\lambda})|\}.$$

Remark 7: The above tests are shown to be applicable on samples coming from a few chosen copulas but the applicability of these tests has not been demonstrated when the data follow FGMD. In our work we have tried to implement the tests proposed by Genest *et al.* (2006) for a sample coming from FGMD. Following are the observations made while implementing these tests and why they do not work.

(a) In order to apply the above GoF tests, $K(t|\lambda = \hat{\lambda})$ needs to be evaluated at $t = 1$ which is encountered when $j = (n-1)$. Note that we have used the extension of the Gaussian quadrature in two dimension to evaluate the definite integral. This numerical integration can be implemented by the 'quad2d' function in the 'pracma' package in R software.

(b) The function $K(t|\hat{\lambda})$ does not take a finite value, meaning the double integration is not convergent at the boundary for a given value of $\hat{\lambda}$. This issue is particularly evident at $t = 1$, but it can also occur at other values of t depending on the value of $\hat{\lambda}$. This phenomenon should be taken into account when implementing the GoF tests.

(c) The double integration for $K(t = 1|\hat{\lambda})$ yields an "NaN" error in R since the integration fails to converge to a finite value. We provided a plot of this phenomenon in the following Figure 2 where $K(t|\lambda)$ has been plotted over t, $0 \leq t \leq 1$, for five different values of λ, $\lambda = -1, -0.5, 0, 0.5, 1$. It is evident from Figure 2 that the double integration fails to converge at the boundary value of $t = 1$ for all λ values. Also, note that for $\lambda = 0.5$ and 1 the integration fails to converge not only at $t = 1$ but also at other values of t between (0,1), thereby making the test statistic S_n or T_n questionable.

(d) The behavior of the function $K(t|\hat{\lambda})$ can be studied for any arbitrary $\hat{\lambda} \in [-1, 1]$. For example, at $\hat{\lambda} = 0$, this function fails to converge to a finite value. The function $K(t|\hat{\lambda})$ fails to attain a finite value when the upper limit of the definite integral is 1, and this is evident from Figure 3, where the function $K(t = 1|\hat{\lambda})$ has been redefined as $K(1 - 10^{-L}|\hat{\lambda})$ has been plotted against L such that the value of $t = 1$ is dependent on L through the relation $t = 1 - 10^{-L}$, *i.e.* as $L \to \infty$, $t \to 1$. This plot gives us an idea on how close to 1 we can achieve a finite value for the integration which defines the function $K(t|\hat{\lambda})$. The integration yields finite value approximately upto $L = 13$ *i.e.* the double integration would converge only upto $t = 1 - 10^{-13}$, not beyond that.

(e) Interestingly, $K(t|\hat{\lambda})$ is a distribution function of the iid pseudo observations V_i's, and by definition it is supposed to demonstrate the non-decreasing property. However, going by the definition of Genest *et al.* (2006) as applied for the bivariate FGMD, $K(t|\hat{\lambda})$ as a function of t for any given λ fails to show the non-decreasing property as seen in Figure 2.

(f) The two test statistics S_n and T_n seem to work well for some non-FGMDs with large sample sizes of 100, 250 or 1000 in case of simulated data, or sample sizes of 1500 and 655 in case of application to real data Genest *et al.* (2006). It is crucial to note that FGMD, although briefly mentioned in Section 3.4 of Genest *et al.* (2006), it has neither been applied to any simulation exercise nor in the real data example. Hence, a GoF test for samples from FGMD became imperative, which has been developed and discussed in the following subsections.

(g) Finally, there is no evidence that the above mentioned tests based on S_n and T_n would work well in case of small and moderate sample sizes for any copula in terms of size as well as power. Table 5 and Table 6 of Genest *et al.* (2006) presents the size and power values only for large samples whereas in this study of ours, the datasets have sample sizes of 23 and 44.

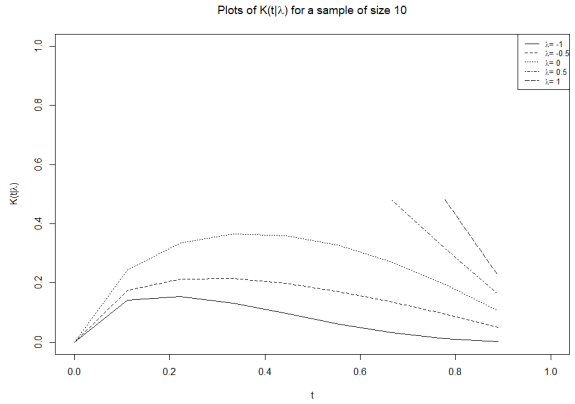


Figure 2: Plot of $\mathbf{K}(t|\lambda)$ for different values of λ.

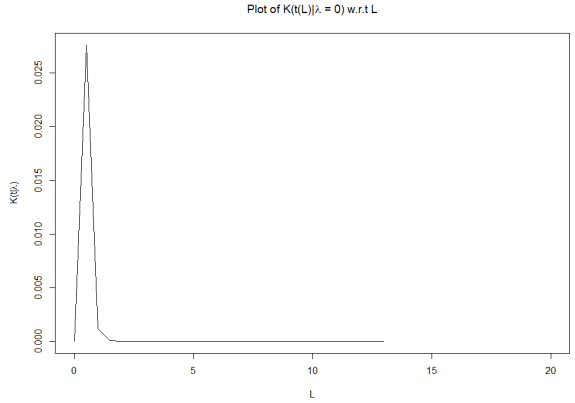


Figure 3: Plot of $\mathbf{K}(1-10^{-L}|\lambda)$ as a function of L for $\lambda = 0$.

4.3. A Bootstrap approach to GoF test for $FGMD$

4.3.1. Developing the test statistic (bivariate case)

Suppose $\mathbf{X} = (X_1, X_2)'$ follows a bivariate distribution with *pdf* $f(\mathbf{x})$ and the corresponding *cdf* $F(\mathbf{x})$. How do we test that $f(\mathbf{x})$ is the $FGMD$ *pdf* given earlier?

For convenience let us denote the marginal *pdf* and *cdf* of X_k by $f_k(.)$ and $F_k(.)$, respectively, $k = 1, 2$. If $\mathbf{X}$ follows $FGMD$, then the above joint *pdf* $f(\mathbf{x})$ and the corresponding bivariate *cdf* will be denoted by say $F_{FGMD}(\mathbf{x}|\lambda)$ as well. Our objective is to test

$$H_0 : F = F_{FGMD}(\mathbf{x}|\lambda), \text{for some } \lambda \quad vs \quad H_A : F \neq F_{FGMD}(\mathbf{x}|\lambda), \text{for any } \lambda, \tag{28}$$

where $\lambda \in [-1, +1]$.

Note that if we use the transformed data $\mathbf{Y}_i$, $1 \leq i \leq n$, where $\mathbf{Y}_i = (Y_{1i}, Y_{2i})'$, $i = 1, 2, ..., n$ and $Y_{ki} = F_k(x_{ki})$, $k = 1, 2$, $i = 1, 2, .., n$, then $\mathbf{Y}_i$'s are n copies of the bivariate random vector $\mathbf{Y} = (Y_1, Y_2)'$ where marginally Y_1 and Y_2 are uniformly distributed over (0,1), and have the joint pdf, say $g(\mathbf{y})$ with the corresponding cdf, say $G(\mathbf{y})$ over the unite square $(0, 1) \otimes (0, 1)$.

If $\mathbf{X}$ has the specified distribution in (5), then equivalently $\mathbf{Y}$ has the distribution with *pdf*, say $g_{FGMD}(\mathbf{y}|\lambda)$, where

$$g_{FGMD}(\mathbf{y}|\lambda) = (1 + \lambda(2y_1 - 1)(2y_2 - 1)) \tag{29}$$

over the unit square $(0, 1) \otimes (0, 1)$. Testing (28) then boils down to testing

$$\overline{H}_0 : G = G_{FGMD}(\mathbf{y}|\lambda) \text{ for some } \lambda \;\; vs \;\; \overline{H}_A : G \neq G_{FGMD}(\mathbf{y}|\lambda) \text{ for any } \lambda, \tag{30}$$

based on the data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$, where G_{FGMD} is the *cdf* corresponding to the pdf g_{FGMD} given in (29).

Remark 8: The broad idea of our testing mechanism will rely on finding a suitable distance between $G(\mathbf{Y})$ and $G_{FGMD}(\mathbf{Y}|\lambda)$. But since F_i's $(i = 1, 2)$ are unknown, we are going to replace them by the corresponding marginal empirical cdfs, *i.e.*, we are going to work with

$$\begin{aligned} \hat{Y}_{ij} &= \hat{F}_{ij}(X_{ij}),\ i = 1, 2 \ and \ j = 1, 2, .., n. \\ &= (1/n)((Number\ of\ X_{ik}\ values \leq X_{ij})),\ 1 \leq k \leq n. \end{aligned} \tag{31}$$

Theoretically, $\mathbf{Y}$ is supposed to have a joint distribution with approximate *pdf* $g(\mathbf{y})$ and approximate *cdf* $G(\mathbf{y})$ whose marginals are uniform. The joint *cdf* $G(\mathbf{y})$ can be approximated by the observed empirical *cdf* $\hat{G}(\mathbf{y})$ as

$$\hat{G}(\mathbf{y}) = (1/n)\{Number\ of\ (\hat{Y}_{1s}, \hat{Y}_{2t})\ values\ \ni\ \hat{Y}_{1s} \leq y_1\ and\ \hat{Y}_{2t} \leq y_2\}, \tag{32}$$

$1 \leq s, t \leq n$. Notice that $\hat{G}_(\mathbf{y})$ is a bivariate step function which can be visualized on the grid points $(\hat{Y}_{1s}, \hat{Y}_{2t})$, $1 \leq s, t \leq n$.

At the same time, if we assume that $\mathbf{X} \sim f_{FGMD}(\mathbf{x}|\lambda)$ (*i.e.*, equivalently, $\mathbf{Y} \sim g_{FGMD}(\mathbf{y}|\lambda)$), then the *cdf* of $\mathbf{Y}$ under $\overline{H}_0$ can be approximated by

$$G_{FGMD}(\mathbf{y}|\hat{\lambda}) = \int_0^{y_1} \int_0^{y_2} g_{FGMD}(\mathbf{u}|\hat{\lambda})\ du_2\ du_1, \tag{33}$$

where $\hat{\lambda}$ is the estimated value of λ, and $g_{FGMD}(.|\hat{\lambda})$ expression is given in (29). It is easy to see that

$$G_{FGMD}(\mathbf{y}|\lambda) = y_1 y_2\{1 + \hat{\lambda}(y_1 - 1)(y_2 - 1)\}, \tag{34}$$

where $\mathbf{y} \in (0, 1) \otimes (0, 1)$. Similar to Kolmogorov - Smirnov test statistic, the distance between $\hat{G}(\mathbf{y})$ in (32) and $G_{FGMD}(\mathbf{y}|\hat{\lambda})$ can be measured by the statistic $\Delta(\mathbf{Y})$ for the given (transformed) data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$; as

$$\Delta(\mathbf{Y}|\hat{\lambda}) = \underset{\mathbf{y} \in (0,1) \otimes (0,1)}{Sup} |\hat{G}(\mathbf{y}) - G_{FGMD}(\mathbf{y}|\hat{\lambda})|. \tag{35}$$

Numerical computation of $\Delta(\mathbf{Y}|\hat{\lambda})$ can be done easily by taking a very fine mesh over the unit square $(0,1) \otimes (0,1)$. Intuitively, one should reject $\overline{H}_0$ if $\Delta(\mathbf{Y}|\hat{\lambda})$ is too "large", and retain $\overline{H}_0$ otherwise. In the following we present a bootstrap method to find a data dependent cut-off value and the p-value.

4.3.2. Leveraging $\Delta(\mathbf{Y}|\hat{\lambda})$ to draw an inference via bootstrap

Step - 1: For the given data $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n)$, compute $\Delta(\mathbf{Y}|\hat{\lambda})$, where $\hat{\lambda}$ is the estimated value of λ (from one of the four estimators as discussed in Section 2).
Step - 2: Assume that $\overline{H}_0$ holds. Using $\hat{\lambda}$ computed in Step - 1, generate a bootstrap sample $\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_n^*$ iid from $g_{FGMD}(\mathbf{y}|\hat{\lambda})$. [This is equivalent to generating $\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_n^*$ iid from $f_{FGMD}(\mathbf{x}|\hat{\lambda})$, and then transforming them back to $\mathbf{Y}_j^*$'s.]
Step - 3: Using the bootstrap data $\mathbf{Y}^* = (\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_n^*)$, recalculate $\hat{\lambda}$ as in Step-1 (pretending that it were unknown). Call this estimate of λ as $\hat{\lambda}^*$ (*i.e.*, $\hat{\lambda}$ based on $\mathbf{Y}^*$). Then compute $\Delta(\mathbf{Y}|\hat{\lambda})$ using this bootstrap data $\mathbf{Y}^*$, and $\hat{\lambda}^*$. Call this $\Delta^* = \Delta(\mathbf{Y}^*|\hat{\lambda}^*)$.
Step - 4: Repeat the above Step - 2 and Step - 3 a large number of times (say, B times). This yields B copies of Δ^*, say Δ^{*b}, $1 \leq b \leq B$, which are then ordered as $\Delta^{*(1)} \leq \Delta^{*(2)} \leq ... \leq \Delta^{*(B)}$.
Step - 5: Our α-level bootstrap cut-off point is found as $\Delta_\alpha^B = \Delta^{*((1-\alpha)B)}$. If Δ (from Step - 1) $> \Delta_\alpha^B$, then reject $\overline{H}_0$, and retain it otherwise.

Alternatively, one can obtain the bootstrap p-value of the GoF test as

$$p_B = \{Number\ of\ \Delta^{*b}\ values\ >\ \Delta(\mathbf{Y}|\hat{\lambda})\}/B,$$

and compare it with α. Essentially with the four proposed estimators of λ we can have four different GoF tests which are named as follows - (i) GoF_1 when $\lambda = \hat{\lambda}_{ML}$, (ii) GoF_2 when $\lambda = \hat{\lambda}_{BFP}$, (iii) GoF_3 when $\lambda = \hat{\lambda}_{BJP}$ and (iv) GoF_4 when $\lambda = \hat{\lambda}_{BAJP}$.

4.4. Results of goodness of fit tests on MDR dataset

In the following table we present the bootstrap p-values of $FGMD$ goodness of fit test through the statistic $\Delta(\mathbf{Y}|\hat{\lambda})$ for the MDR data as discussed earlier. The results have been obtained for both north and south regions using all the four estimators of λ (*i.e.*, $\hat{\lambda} = \hat{\lambda}_{ML}, \hat{\lambda}_{BFP}, \hat{\lambda}_{BJP}\ and\ \hat{\lambda}_{BAJP}$).

Table 4: Goodness of fit p-values through bootstrap for testing $FGMD$

Variable Combination	North				South			
	MLE	BFP	BJP	$BAJP$	MLE	BFP	BJP	$BAJP$
As and Cl	0.499	0.498	0.496	0.496	0.647	0.661	0.665	0.660
As and Eh	0.571	0.558	0.557	0.562	0.838	0.826	0.829	0.837
As and pH	0.926	0.926	0.927	0.927	0.509	0.500	0.508	0.519

Table 4 clearly shows that bivariate FGMD is definitely an acceptable joint distribution to model As along with each of the three other variables, *i.e.*, Cl, Eh and pH.

Obviously a natural extension of this observation is that of studying the conditional distribution of As given a benign element, and then making a suitable prediction of As. This aspect of prediction will be reported in a separate comprehensive work.

Remark 9: (a) In this section, we have made an attempt to address the elusive query of goodness of fit of a copula, specifically the FGMC. We have proposed a class of parametric bootstrap (PB) tests based on the Kolmogorov - Smirnov (KS) distance between the two $cdfs$, - the hypothesized FGMD and the empirical one (see (32)). To the best of our knowledge, this is first time that GoF of FGMD has been addressed in a comprehensive manner for small to moderate sample sizes.

(b) We have been able to show that our proposed parametric bootstrap tests not only adhere to the size criterion (see Chatterjee (2022)) but also there is no need to know the null distributions of the test statistics, either for a fixed sample or asymptotically. The performance of the tests in terms of power indicate that they are almost identical, and hence any one of the four can be used in applications.

(c) Though this section deals with GoF of a bivariate FGMD to model As with another element, one should look at a possible extension in a multidimensional set up ($i.e.$, going beyond the dimension 2) so that As can be modeled along with (Cl, Eh, pH) for a more meaningful analysis of the data. This is a future research problem which will be taken up later. Another potential avenue for further GoF study is to use a different distance measure (other than the KS one) and study the resultant implications.

5. Predictions under FGMD

5.1. The rationale behind prediction

Our in-depth analysis of the given data shows that each variate (As, Cl, pH and Eh) individually has a vastly different probability distribution in each of the two subregions. The following Figure 4 shows the sample relative frequency histograms of the four variables in two subregions.

We have used two well-known and widely accepted formal test methods, namely - Anderson-Darling Test (ADT) and Shapiro - Wilk Test (SWT), to check if the sample histograms in Figure 4 conform to normality. Unfortunately six out of eight sample histograms rejected normality. Only the variate pH, and that too for the southern subregion, accepted normality (by both ADT and SWT) comfortably with very high p-values. In the north, Eh appears to follow normality with moderately large p-values. Usually one should feel comfortable with the assumption of normality if both ADT and SWT show substantially large p-values. The following Table 5 shows the p-values for all the four variates in the two subregions when both the tests are applied.

Further rigorous investigation showed that not only the six out of eight subdatasets (four variables in two subregions) are non-normal, each variable's probability distributions in the two subregions are vastly different. In this regard we show the p-values of the well known Kolmogorov-Smirnov Test (KST) to test the equality of two distributions in the following Table 6.

Table 5: *ADT* and *SWT* p-values to test the normality in two subregions

Test	Subregion	*As*	*Cl*	*Eh*	*pH*
North	ADT	< 0.0001	0.0414	0.1404	0.0029
	SWT	< 0.0001	0.0286	0.0914	0.0012
South	ADT	< 0.0001	< 0.0001	< 0.0001	0.4363
	SWT	0.0002	< 0.0001	< 0.0001	0.6933

Table 6: KST p-values to check equality of distributions in two subregions

	As	*Cl*	*Eh*	*pH*
p-value	< 0.0001	0.0004	< 0.0001	< 0.0001

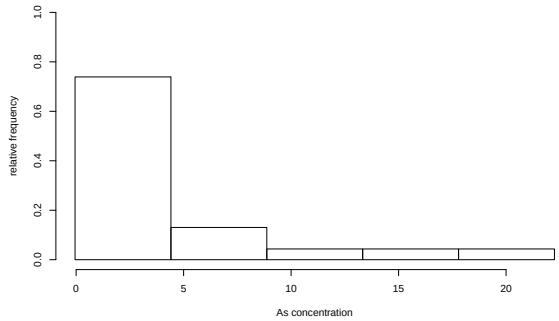
The above observations about the distributional properties of four variables now set the ground for bivariate scatter plots between *As* and each of the other three variables. Figure 5 comprehensively shows the six scatterplots in the two subregions.

Notice that out of six bivariate scatterplots, four do not show any linear trend (and these are (a), (b), (d) and (f)). Plots (c) and (e) are somewhat linear, but the variations (or dispersion) of *As* against *Eh* (in (c)), and *pH* (in (e)) do not look uniform (*i.e.*, the conditional probability distribution of *As* given another variable appears to be heteroscedastic). As a result, the standard Pearson's correlation coefficient ρ_P is not going to be an adequate measure to assess the association between *As* and other variables in a bivariate framework.

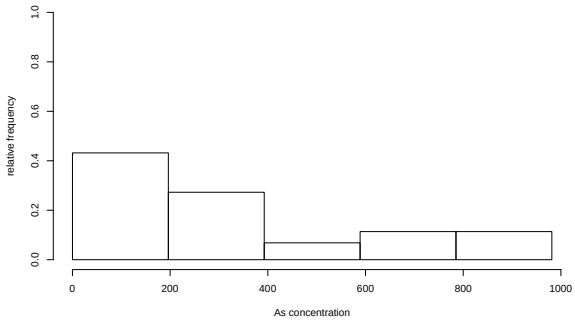
Yet, for the sake of argument, one can compute the three standard correlation estimates, including ρ_P, while the other two being the Spearman's rank correlation (denoted by ρ_S) and Kendall's 'Tau' (or, Kendall's rank correlation), denoted by ρ_K, to get an overall sense of these associations. While ρ_P measures the strength of linear association, ρ_S and ρ_K are much more robust, and indicate the strength of monotonic association between the two variables of interest. The following Table 7 provides the three sample correlation measures for three pairs of variables in the two subregions. The value in parentheses under each entry is the p-value for testing the null hypothesis (H_0) which states that the true (or population) correlation measure is zero, against the alternative hypothesis (H_A), which negates the null. Note, in Table 7: '***' implies p - value ≤ 0.01; '**' for ≤ 0.05; and '*' for ≤ 0.10.

Table 7: Estimated standard correlation coefficients with corresponding p-values

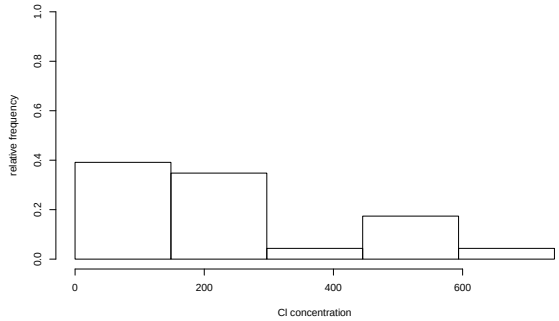
	North ($n_N = 23$)			South ($n_S = 43$)		
	As vs Cl	*As vs Eh*	*As vs pH*	*As vs Cl*	*As vs Eh*	*As vs pH*
$\hat{\rho}_P$	0.182	0.525	0.754	-0.325	-0.668	0.119
	(0.405)	(0.010*)	(0.000***)	(0.031**)	(0.000***)	(0.442)
$\hat{\rho}_S$	0.018	-0.414	0.260	-0.320	-0.753	0.156
	(0.936)	(0.050**)	(0.231)	(0.035**)	(0.000***)	(0.314)
$\hat{\rho}_K$	0.012	-0.323	0.188	-0.230	-0.577	0.101
	(0.937)	(0.032**)	(0.213)	(0.028**)	(0.000***)	(0.336)



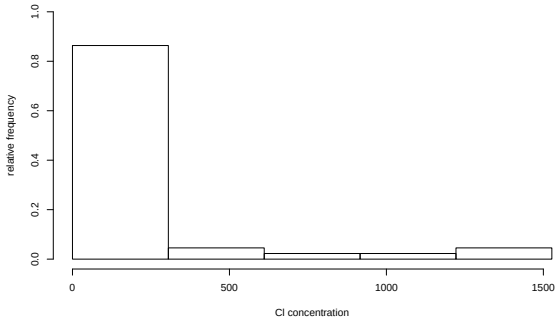
(a) *As* (North)



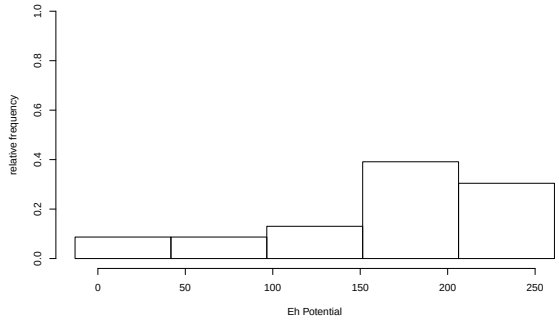
(b) *As* (South)



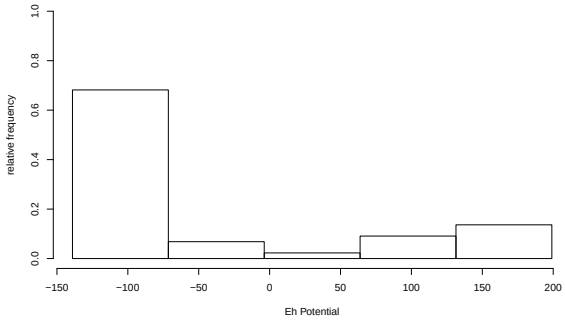
(c) *Cl* (North)



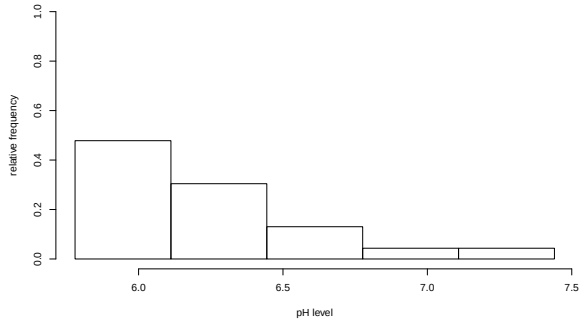
(d) *Cl* (South)



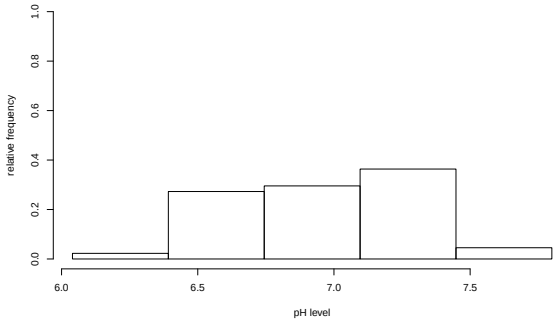
(e) *Eh* (North)



(f) *Eh* (South)

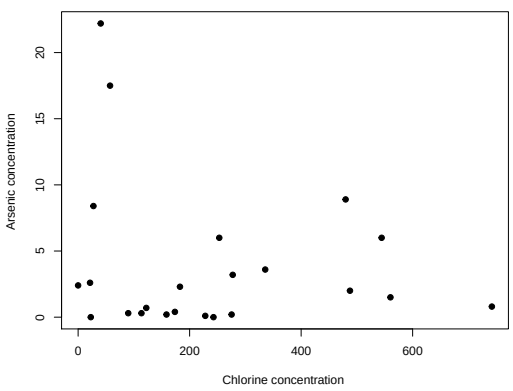


(g) *pH* (North)

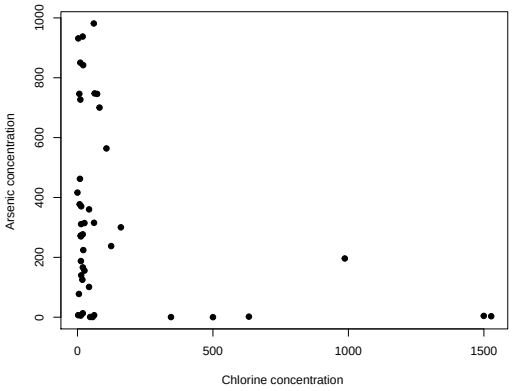


(h) *pH* (South)

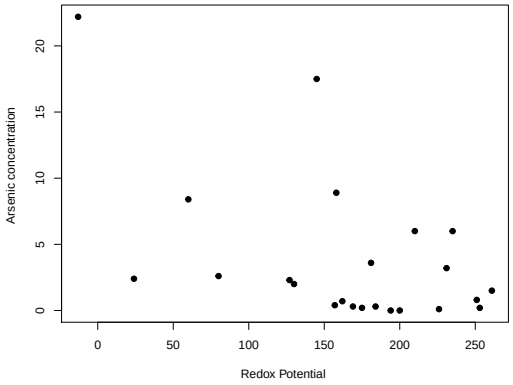
Figure 4: Relative frequency histograms of the four variables in two subregions



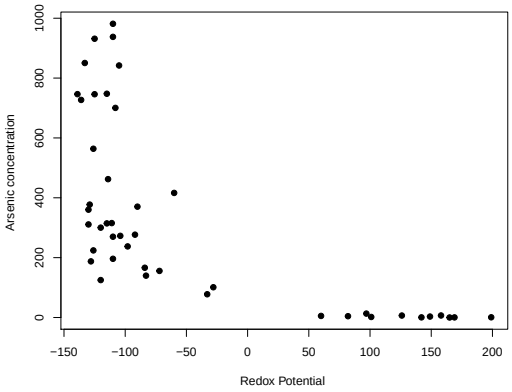
(a) As vs Cl (North)



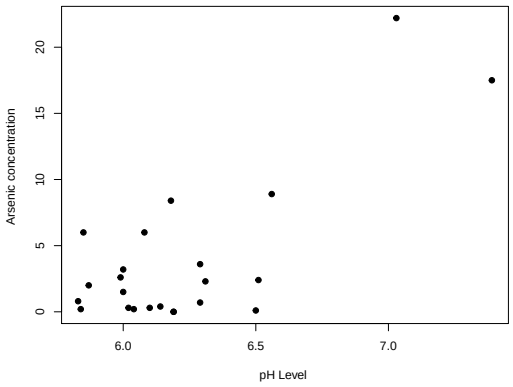
(b) As vs Cl (South)



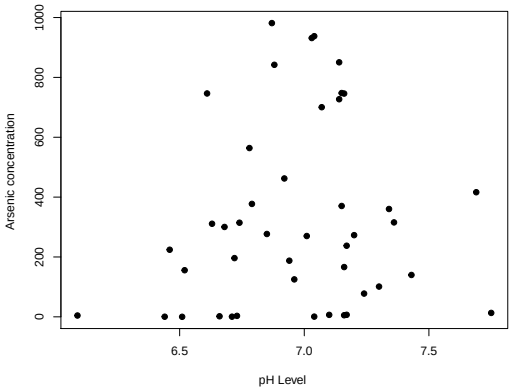
(c) As vs Eh (North)



(d) As vs Eh (South)



(e) As vs pH (North)



(f) As vs pH (South)

Figure 5: Scatter plots of As against each of the other three benign variables

Remark 10: The three estimated standard correlation measures portray some interesting scenarios as summarized below.

(a) Between As and Cl, in the North, all three correlations indicate that there is no association. However, in the South, they all indicate a significant negative association.

(b) Between As and Eh, there appears to be a significant negative association in both the subregions mostly. Only a conflicting picture is provided by ρ_P in the north which shows a significant positive linear association.

(c) Between As and pH, there appears to have no significant association mostly as all but one p-values are quite high (more than 20%). However, only ρ_P shows a strong linear positive association in the north.

The usual correlation measures show interesting associations between As and Cl, As and pH, and As and Eh, but they do not help predict As values based on these variables. For instance, in regions where As and Cl are strongly associated, predicting As when Cl is $10ppm$ or Eh is $100mv$ requires more than just linear regression, which assumes normality and homoscedasticity. The problem is better addressed by fitting a suitable bivariate probability distribution to the data. Specifically, we need a non-normal bivariate distribution of (X, Y), where Y is As and X can be Cl, pH, or Eh, using copula theory. This approach allows us to use the conditional distribution of Y given X to make predictions. The goal is to explore the association between As and the other variables beyond linear correlation and to exploit this association for prediction, recognizing that simplistic linear models might lead to incorrect conclusions.

5.2. Prediction of Y using a covariate under FGMD

The parameter λ, by its appearance, has some similarities with the three standard correlation measures discussed earlier. If $\lambda = 0$, then X and Y are independent; if $\lambda < 0$ (> 0), then they are negatively (positively) associated.

Note that, marginally the pdfs f_X and f_Y of X and Y are unknown, and so are the cdfs F_X and F_Y. It is possible to adopt a suitable parametric model for f_X and f_Y, but that is not the main focus of our study. We want to bypass this aspect of unknown marginals by replacing F_X (and F_Y) by $\hat{F}_X$ (and $\hat{F}_Y$) where $\hat{F}_X$ (and $\hat{F}_Y$) represents the empirical distribution function defined as (where the subscript has been dropped for generalization)

$$\hat{F}(t) = \{\text{number of observed data points} \leq t\}/n, \tag{36}$$

n being the sample size. Therefore, we pretend that the marginals (F_X and F_Y) are known, and they can be replaced by their estimates whenever needed.

The conditional distribution of Y given X is

$$f_{Y|X}(y|x) = f_Y(y)[1 + \lambda(2F_X(x) - 1)(2F_Y(y) - 1)], \tag{37}$$

and this will be used in predicting the value of Y when $X = x$ is given. The main challenge here is the estimation of λ. Some noteworthy works along this line, *i.e.*, regressing Y based

on X from copula based predictive models is discussed as follows. We now proceed with the conditional distribution given in (37) to predict the value of Y when $X = x$ is given. We have discussed the structures of the predictors of Y based on the three standard measures of center of a conditional probability distribution as follows.

5.3. Conditional mean as a predictor

The first predictor is the conditional mean, denoted by $\hat{Y}_{mean}(x)$, given as

$$\hat{Y}_{mean}(x) = E(Y|X = x) = \int y f_{Y|X}(y|x)dy, \tag{38}$$

where the integration is over the appropriate range of Y given $X = x$. In the current context where Y represents the As level, this range of Y is $(0, \infty)$. It can be shown that (see Chatterjee (2022)) the expression (38) simplifies to

$$\hat{Y}_{mean}(x) = \mu_Y + \lambda G_X(x) I_Y(F_Y), \tag{39}$$

where μ_Y is the unconditional mean of Y, $G_X(x) = (2F_X(x) - 1)$ and

$$I_Y(F_Y) = \int_{-1}^{1} (t/2) F_Y^{-1}((1 + t)/2)dt. \tag{40}$$

In applications, μ_Y will be replaced by $\hat{\mu}_Y = \overline{Y}$ = sample mean of Y observations, and $G_X(x)$ will be replaced by $\hat{G}_X(x) = 2\hat{F}_X(x) - 1$ with $I_Y(F_Y)$ being replaced by $I_Y(\hat{F}_Y)$. Also, λ will be replaced by a suitable estimator as mentioned earlier.

5.4. Conditional median as a predictor

Another simple predictor is the median calculated from the conditional distribution of $Y|X$. When the conditional distribution is skewed, which is expected in real life applications, the conditional median tends to be a robust predictor than the mean. We get the conditional median $\hat{Y}_{median}(x)$ of $Y|X$ by solving the following Equation (41) in terms of M, the desired median of the conditional distribution.

$$0.5 = \int_{-\infty}^{M} f_Y(y)(1 + \lambda(2F_X(x) - 1)(2F_Y(y) - 1)\partial y. \tag{41}$$

For brevity, we use the following notation $\lambda G_X(x) = A$ and $F_Y(M) = w$. It can be shown (see Chatterjee (2022)) that solving Equation (41) boils down to solving

$$2Aw^2 + 2(1 - A)w - 1 = 0. \tag{42}$$

The feasible solution from the above quadratic equation and inverting the cdf for Y gives us our estimate of the Y for a given x based on the median of the conditional distribution. Hence the predictor is as follows

$$\hat{Y}_{median}(x) = F_Y^{-1}[\frac{\{\lambda(2F_X(x) - 1) - 1\} + \sqrt{1 + \lambda^2(2F_X(x) - 1)^2}}{2\lambda(2F_X(x) - 1)}]. \tag{43}$$

In applications, λ will be replaced by $\hat{\lambda}$, the marginal $F_X(x)$ is replaced by $\hat{F}_X(x)$, and F_Y^{-1} should be replaced by $\hat{F}_Y^{-1}$.

5.5. Conditional mode as a predictor

One can use a predictor of the third kind, *i.e.*, the conditional mode, which can be found simply by deriving the mode of the conditional distribution of $Y|X$, which is found by differentiating the conditional pdf and equating it with zero, *i.e.*,

$$(\partial/\partial y)(f_{Y|X}(y|x)) = 0, \tag{44}$$

assuming that the $f_{Y|X}$ is absolutely continuous. Using the usual notation of $f_Y(y) = f_Y$ and $\lambda(2F_X(x) - 1) = \lambda G_X(x) = A$, the equation (44) above then can be written as $f_Y' + A(2(f_Y)^2 + (2F_Y(y) - 1)f_Y') = 0$. Substituting for $2F_Y(y) - 1 = u(y) = u$(say), we have $2f_Y = u'$ and $2f_Y' = u''$. The above equation (44) thus yields a second-order ordinary differential equation as $(u''/2) + 2A(u'/2) + Auu''/2 = 0$, *i.e.*, $(u')^2 + (B + u)u'' = 0$ where $B = 1/A$. Let $(B + u) = v$, *i.e.*, $u' = v'$, *i.e.*, $u'' = v''$. Then the above equation boils down to the differential equation $(v')^2 + vv'' = 0$, *i.e.*, $\partial(vv') = 0$, *i.e.*, $vv' = c$, for some constant c, *i.e.*, $v\partial v/ = c\partial y$, *i.e.*, $v^2/2 = cy + d$, *i.e.*,

$$(B + u)^2 = c_1 y + d_1, \tag{45}$$

where $c,\ d,\ c_1,\ d_1$ are suitable constants. The above final expression (45) gives the general solution of the differential equation (44). In order to find the values of the constants in the solution, specific boundary values were chosen, say $y = y^*$ and $y = y^{**}$, where y^* and y^{**} are two suitable small and large extreme values of the variable y over its support. For the purpose of computational convenience we have taken $y^* = y_{(1)}$ and $y^{**} = y_{(n)}$, the smallest and the largest observed values of the variable Y respectively. Then, $u(y^*) \approx -1$ and $u(y^{**}) \approx 1$ respectively. Plugging-in these choices of $y = y^*$ and $y = y^{**}$ as boundary values in our general solution (45), we get the following solution with $c_1 = c_1^* = 4B/y^{**}$ and $d_1 = d_1^* = (B - 1)^2$ as

$$4(\hat{F}_Y(y))^2 + 2(B - 1)(\hat{F}_Y(y)) = c_1^* y. \tag{46}$$

Remark 11: The solution of (46) in terms of Y gives an approximate mode of the conditional distribution of $Y|X$. Further, note that this conditional mode depends on $X = x$, through the term $B = 1/A$, which involves x. Thus the solution of the Equation (46) will be the intersection of the plots of the left hand side (LHS) and the right hand side (RHS) of the said equation within the range of Y. But the plot of the LHS depends on the sign of B. The sign of B in turn is dictated by the sign of λ and the sign of the term $(2F_X(x) - 1)$. For example, in the southern subregion, the data on *As* and *Eh*, all the estimates of λ are negative. This phenomenon is true for the estimate calculated for the entire data and all the estimates calculated by the "Leave-One-Out Bootstrap" (LOOB) computation (elaborated in Section 5.6). This means, the sign of B is determined on the basis of three distinct scenarios eventually giving rise to three distinct cases: (*i*) $x >$ median (X); (*ii*) $x <$ median (X); (*iii*) $x \equiv$ median (X).

Eventually, the mode predictor of the conditional distribution, denoted by $\hat{Y}_{mode}(x)$,

is defined in the following way

$$\hat{Y}_{mode}(x) = \begin{cases} \text{(possible) solution(s) of (46)} & \text{if } x \neq \text{median}(X) \\ \text{unconditional mode of } (Y) & \text{if } x \equiv \text{median}(X). \end{cases} \tag{47}$$

Remark 12: Regarding the above predictor expression in (47) note the following -

(a) The sign of B (negative or positive) is determined by whether $x > \text{median}(X)$ or $x < \text{median}(X)$ from the definition of B as noted earlier.

(b) For example, in As vs Eh in the southern subregion, we observe that $\text{median}(X) = -106.5\ mV$. When $x < \text{median}(X)$, B is positive and we have noted the plots of both RHS and LHS are monotonically increasing. The opposite happens when $x > \text{median}(X)$. Representative plots for both the cases have been considered in Figure 6 and Figure 7 *i.e.*, *As* vs *Eh* data from the southern subregion in MDR for $x = -126\ mV$ and $x = 126\ mV$ which are less than and greater than the $median(X)$ respectively.

(c) Analytically, there is a possibility of having multiple solutions of (46) due to multiple intersections between LHS and RHS. The intuition behind finding the mode of the conditional density function is as follows. One can collect all the multiple solutions and check which one yields the maximum value of the conditional density given in (37). Under the assumed parameter free model of the marginals, the next step is to estimate the marginal density in the expression (37), *i.e.* how to get $f_Y(y)$. This can be achieved in a multiple ways but we have presented a simple and straightforward way in our *LOOB* calculations. Assume one obtains multiple solutions as y_i^*, $i = 1, 2, ..., k$ in (47) and $k < n$ within the range of Y, then the unconditional density function at y_i^*, by definition, is the rate of change of the cumulative distribution function at y_i^*. In light of this definition, one can approximate $f_Y(y_i^*)$ as $f_Y(y_i^*) \approx (1/nh) \sum_{i=1}^{n} K((y_i^* - y_i)/h)$, where $K(.)$ is a suitable kernel - a non-negative function, y_i, $i = 1, 2, ..., n$ are the sample observations and $h > 0$ is a smoothing parameter called the bandwidth.

(d) An in-built R-package has been used for the above density estimation which uses the Gaussian kernel function. While choosing the bandwidth in Kernel, which is still an open topic of research, the default choice of bandwidth rule selection in the R - package is by Silverman's 'rule of thumb' (Silverman (1986, page 48, eq.(3.31)). This choice is more appropriate if the original distribution (*i.e.* the true marginals) is bell-shaped and symmetric in nature. In contrast, none of our marginals are symmetric and bell-shaped. Therefore, in our case Sheather and Jones method (Sheather and Jones (1991)) is more applicable which is a more robust and data dependant approach. Moreover, in theory, a finer kernel bandwidth reveals more intricacies in the true distribution. But there is a risk of under-smoothing by choosing a too small 'h'. On the other hand, a risk of over-smoothing exists if 'h' is too large. We have examined several bandwidths under the Sheather and Jones rule and have chosen $h = 10^{10}$.

(e) Finally, when $x \equiv \text{median}(X)$, it is straightforward to note from the conditional distribution in (37) that the predictor would be the unconditional mode of Y, say Y_{mode}.

However, the exact sampling distributions of these three predictors are intractable theoretically. Therefore, the performance of these three predictors have been evaluated

through the 'Leave-One-Out-Bootstrap' ($LOOB$) method, which has been discussed and applied on the groundwater data from MDR in the following subsection. Extending on the existing idea and incorporating ten estimators of λ the association parameter of FGMD (one traditional estimator - MLE, and nine Bayes' estimators - under each of the three types of priors - Flat, Jeffrey's and approximate Jeffrey's prior, three types of central tendency measure (expectation, median and mode) of the three resulting posterior distribution) and three predictors, we have achieved a collection of thirty predicted values of Y for a given value of X.

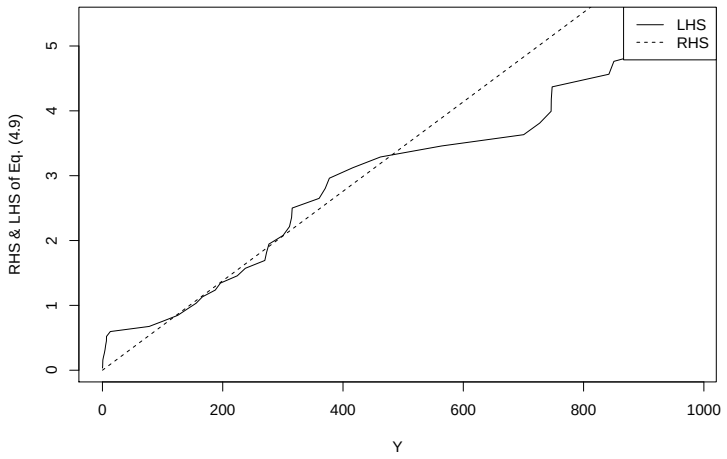


Figure 6: Plots of LHS and RHS of equation (46) when $x = -126\ mV$ ($<$ median(X)), where $X = Eh$ in the southern subregion.

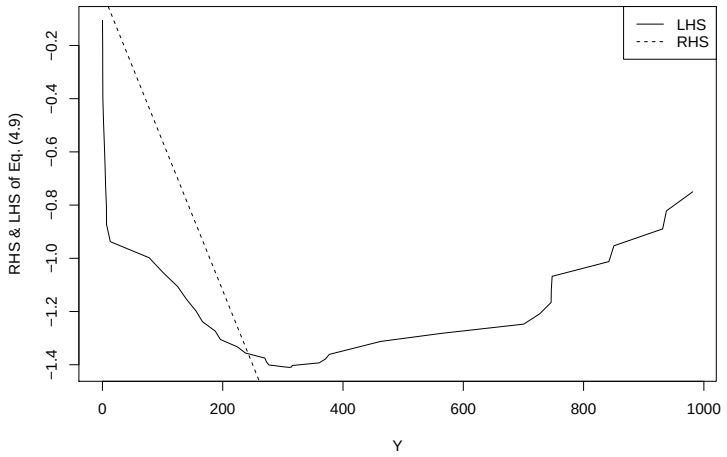


Figure 7: Plots of LHS and RHS of equation (46) when $x = 126\ mV$ ($>$ median(X)), where $X = Eh$ in the southern subregion.

5.6. Applications: Predicting arsenic from Cl

In this section we are going to demonstrate how to predict As from the three covariates but focusing mostly on Cl in the southern subregion just as an example. Before delving into the performances of the different predictors in the As prediction study, let us look at the $FGMD$ association parameter estimates from the two subregions in Table (8). Note that, for a pair of variables that includes As, we can employ a total of ten estimates of λ. Selection of an estimator of λ will depend on the overall $LOOB$ performance to be explained below.

Table 8: Estimates of the FGMD association parameter in the MDR subregions

(a) Northern subregions

Pair of Elements	MLE	Posterior								
		BFP			BJP			$BAJP$		
		Mean	Median	Mode	Mean	Median	Mode	Mean	Median	Mode
As vs Cl	0.089	0.051	0.065	0.061	0.064	0.075	0.134	0.108	0.195	0.560
As vs Eh	-1	-0.585	-0.665	-0.254	-0.646	-0.745	-0.341	-0.803	-0.905	-0.114
As vs pH	0.749	0.418	0.475	0.371	0.475	0.555	0.722	0.666	0.815	0.311

(b) Southern subregion

Pair of Elements	MLE	Posterior								
		BFP			BJP			$BAJP$		
		Mean	Median	Mode	Mean	Median	Mode	Mean	Median	Mode
As vs Cl	-0.979	-0.624	-0.675	-0.176	-0.674	-0.745	-0.242	-0.810	-0.895	-0.083
As vs Eh	-1	-0.869	-0.905	-0.001	-0.892	-0.925	-0.001	-0.941	-0.975	-0.001
As vs pH	0.593	0.425	0.465	0.302	0.469	0.525	0.754	0.648	0.765	0.344

We present the findings on the performance of the predictors using As as Y and Cl as X. In a generic sense, when we apply FGMD to a data set, we should first estimate the association parameter λ in the model, and call it $\hat{\lambda}$. We now apply the three predictors as discussed in Section 3, to predict As from Cl. We implement a Leave-One-Out-Bootstrap ($LOOB$) method to evaluate the performance of the three predictors. The scheme of $LOOB$ is simple where we drop one observation (a pair of As and the corresponding Cl observation) from the dataset and then we fit the FGMD model onto that reduced dataset with $(n-1)$ observations. We can estimate the association parameter λ by making use of any suitable estimator as mentioned in Section 2. Finally, with the estimated model, we use Cl (the independent variable or X in our study) of the dropped off observation to estimate the corresponding As (the dependent variable or Y in our study). This $LOOB$ mechanism is applied to all the n observations which in turn helps us to see how a predictor fared against all the true observations.

The performance of a predictor in conjunction with an estimator of λ is then evaluated by the Prediction Mean Absolute Error ($PMAE$) and Prediction Root Mean Squared Error ($PRMSE$). The following Table 9 presents $LOOB-PMAE$ and $LOOB-PRMSE$ of the three predictors each with one of the ten estimators of λ. Let Y_i be the i^{th} observation of Y (= As), and $\hat{Y}_i^{(-i)}$ is the predicted value of Y_i based on the remaining $(n-1)$observations (after fitting the $FGMD$) and using X_i (= the i^{th} value of Cl), then $PMAE = \sum_{i=1}^{n} |Y_i - \hat{Y}_i^{(-i)}|/n$, and $PRMSE = [\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(-i)})^2/n]^{1/2}$.

Table 9: Performance of the predictors of As from Cl in the southern subregion

		λ Estimate Used									
	Predictor	MLE	Posterior Mean			Posterior Median			Posterior Mode		
			BFP	BJP	$BAJP$	BFP	BJP	$BAJP$	BFP	BJP	$BAJP$
PMAE	$Mean$	248.12	247.15	246.90	245.86	247.37	247.15	245.73	248.98	245.12	251.70
	$Median$	242.06	243.81	244.16	242.64	244.54	244.14	241.19	244.54	244.14	241.19
	$Mode$	**230.76**[2]	241.98	249.63	253.55	248.69	247.41	**237.87**[3]	241.42	251.36	**204.33**[1]
PRMSE	$Mean$	306.47	305.55	305.26	304.46	305.60	305.41	304.34	305.52	**302.03**[3]	307.69
	$Median$	314.65	313.65	314.22	314.63	314.55	315.18	313.50	314.55	315.18	313.5
	$Mode$	**298.86**[2]	319.35	325.18	334.94	324.57	328.74	309.14	319.5	330.34	**286.28**[1]

Remark 13: (a) Overall, the mean predictor and the mode predictor based on the conditional distribution show better performance than the mode predictor in terms of both $PRMSE$ as well as $PMAE$. Out of the ten estimators of the association parameter, MLE of λ is consistently the top performer followed by $BAJP2$ in the second place and $BJP3$ in the third place. This $LOOB$ based work is highly data dependent, *i.e.*, for another dataset the performance evaluation measures can vary drastically and hence one must apply all the three predictors and all the ten estimators of the association parameter to see which predictor (along with which $\hat{\lambda}$) has the best overall performance.

(b) If the parabolic shape of the scatterplot between As and one of the covariates is ignored and the usual simple linear regression model is force-fitted, then it can cause several theoretical as well as practical complexities, such as: (i) the normality assumption of the errors which is implicit in linear regression, is violated; (ii) the homoscedasticity of the error variance is not tenable; and (iii) the predicted value of As may result in negative values as seen for several of our data points.

(c) Forcing a linear regression upon ignoring the previously stated concerns, and truncating the As value at 0 for negative predicted values (unrealistic in nature though) may still result in poor performance.

(d) One has to keep in mind that these aforementioned computational results are based on using a single predictor (Eh, Cl or pH). Things will definitely improve if we use two predictors (say, Eh and Cl) or all the three predictors (Eh, Cl and pH). This requires upgrading our bivariate $FGMD$ to a trivariate or a quadravariate $FGMD$ and this is currently under investigation. While a bivariate $FGMD$ has a single association parameter λ ($= \lambda_{12}$) between the components 1 and 2, a trivariate $FGMD$ has four association parameters λ_{12}, λ_{13}, λ_{23} and λ_{123}. Expanding it further, a quadravariate $FGMD$ has a total eleven association parameters. Ota and Kimura (2021) considered the three variate $FGMC$ and the resultant $FGMD$ mainly from an asymptotic point of view. More specifically, they considered the special case of $\lambda_{12} = \lambda_{23} = \lambda_{13} = \lambda_{123} = \lambda$ (say), and considered estimation of the common association parameter λ. However, more work needs to be done to investigate the exact sampling distribution of the MLE either for all the four parameters or the single common parameter in dimension three. How the behavior of high probability concentration of MLE near the boundary, as we have seen in the bivariate case and discussed in Section 2, permeates to 3 or higher dimensions, needs to studied extensively especially for small to moderate sample sizes.

Further, the Bayesian estimation of the association parameter vector in dimension greater than 2 may lead to interesting results.

(e) The main challenge in dealing with a p-dimensional ($p > 2$) $FGMD$ is to carry out a very complex set of computations within a reasonable amount of time which requires sophisticated computational codes. We are currently studying the trivariate $FGMD$ and how it can be used for the arsenic prediction study. This will be reported in near future as we sort out the computational complexities. The case of $p = 2$ is the springboard for the higher dimensional generalizations. Even for $p = 3$, in order to find the maximum likelihood estimates of λ_{12}, λ_{13}, λ_{23} and λ_{123} is a computational nightmare as the optimization is to be done in a 4-dimensional space over a feasible region subject to 8 linear inequalities (*i.e.*, the feasible region has a 'diamond cut' shape).

6. Conclusion

With the onset of copula theory which brought about an influx of several copula based joint distributions and its growing application across several disciplines, it is of paramount interest to investigate the copula models more closely. The flexibility of the copula model lies in producing a unique link function (in the continuous random variate case) which essentially joins the marginals. This copula function preserves the entire information about the mutual dependence between two marginals through a single association parameter.

In our work, we have provided a template of a comprehensive inferential investigation of the association parameter of $FGMD$. In our application, we have taken up the bivariate case *i.e.*, we have studied the pairwise components of the groundwater data of MDR. There is, in fact, an array of future directions that are in the works for this research stream -

(a) The generalization of the copula model to a p - dimensional ($p > 2$) set up. Investigating the sampling distribution of the different estimators in the general case and construction of confidence bands.

(b) We have seen the superiority in performance of the Bayes' estimators but the computational challenge was stifling at times. It is intuitive that this challenge will only grow as p increases. Tackling this computational challenge in itself will be an interesting data science research problem.

(c) Development of higher dimensional predictors and subsequent GoF test will be another research problem. Our GoF tests which show adherence to the size criteria in the bivariate case, need to be studied in higher dimensional cases.

(d) The nature of our study for bivariate $FGMD$ has been comprehensive and covers several inferential aspects. This template of investigation can be extended to other commonly used Archimedean and non-Archimedean copulas.

(e) Even though our comprehensive study on $FGMD$ was motivated by an environmental dataset, one might be interested to study how the copula based models can reveal some hidden information for other datasets especially sparse gene expression datasets.

Acknowledgements

A version of this paper was presented by the second author as an invited talk at the 26th Annual Conference of The Society of Statistics, Computer & Applications (SSCA), hosted and co-organized by the Department of Mathematics and Statistics & Center for Artificial Intelligence, Banasthali Vidyapith, Rajasthan, India (February 26 - 28, 2024). The authors are deeply indebted to the organizers for their hospitality and generous support.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government printing office.

Amblard, C. and Girard, S. (2009). A new extension of bivariate fgm copulas. *Metrika*, **70**, 1–17.

Bairamov, I. and Kotz, S. (2002). Dependence structure and symmetry of huang-kotz fgm distributions and their extensions. *Metrika*, **56**, 55–72.

Bekrizadeh, H., Parham, G. A., and Zadkarmi, M. R. (2012). The new generalization of farlie–gumbel–morgenstern copulas. *Applied Mathematical Sciences*, **6**, 3527–3533.

Chatterjee, R. (2022). *Inferences for the Bivariate Probability Distribution Using Farlie - Gumbel -Morgenstern Copula*. PhD thesis. Available at https://www.proquest.com/dissertations-theses/inferences-bivariate-probability-distribution/docview/2882152890/se-2; Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-10-3.

D'este, G. (1981). A morgenstern-type bivariate gamma distribution. *Biometrika*, **68**, 339–340.

Farlie, D. J. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, **47**, 307–323.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, **95**, 119–152.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12**, 347–368.

Genest, C., Kojadinovic, I., Nešlehová, J., and Yan, J. (2011). A goodness-of-fit test for bivariate extreme-value copulas. *Bernoulli*, **17**, 253–275.

Genest, C., Quessy, J.-F., and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics*, **33**, 337–366.

Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, **44**, 199–213.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, **55**, 698–707.

Huang, J. S. and Kotz, S. (1999). Modifications of the farlie-gumbel-morgenstern distributions. a tough hill to climb. *Metrika*, **49**, 135–145.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2004). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley & Sons.

Merola, R., Hien, T., Quyen, D., and Vengosh, A. (2015). Arsenic exposure to drinking water in the mekong delta. *Science of the Total Environment*, **511**, 544–552.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, **8**, 234–235.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Ota, S. and Kimura, M. (2021). Effective estimation algorithm for parameters of multivariate farlie–gumbel–morgenstern copula. *Japanese Journal of Statistics and Data Science*, **4**, 1049–1078.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**, 683–690.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, **8**, 229–231.

A. Appendix

A.1. Application data

NOTE: Well ID starting with TH are located in the northern subregion and Well ID starting with DT or TB are from the southern subregion.

Table 10: MDR groundwater data

Well ID	As (ppb)	Cl (ppm)	Eh (mv)	pH	Well ID	As (ppb)	Cl (ppm)	Eh (mv)	pH
DT7	563.9	107	-126	6.78	TB19	300.3	160.3	-120	6.68
DT6	0.5	56.1	142	6.71	TBE10	700.4	81.4	-108	7.07
DT5	0.7	46.8	199	7.04	TBE9	196.2	986.6	-110	6.72
DT3	0.4	345.2	169	6.44	TBE7	166.3	20	-84	7.16
DT4	0.1	500.1	165	6.51	TBE4	4.4	1499.6	82	6.09
DT2	1.8	632.8	101	6.66	TBE5	981.4	60.4	-110	6.87
DT1	13.1	19.7	97	7.75	TBE3	6.8	2.7	158	7.17
TB11	462.3	9.2	-114	6.92	TBE1	6.6	61.7	126	7.1
TB18	155.7	25.9	-72	6.52	TBE11	5.3	12.2	60	7.16
TB9	187.6	12.8	-128	6.94	TBE6	3.2	1527	149	6.73
TB2	850.4	10.5	-133	7.14	TH16	0.4	173.6	157	6.14
TB24	370.4	13.9	-90	7.15	TH9	0.2	275.3	253	5.84
TB26	139.9	13.8	-83	7.43	TH13	0	22.7	194	6.19
TB27	77.7	5.4	-33	7.24	TH14	0.3	113.6	184	6.02
TB21	842.1	21.1	-105	6.88	TH22	0.1	228.1	226	6.5
TB1	276.8	19.6	-92	6.85	TH21	0.3	89.9	169	6.1
TB10	377.3	8.2	-129	6.79	TH5	0.8	742.1	251	5.83
TB25	272.9	11.9	-104	7.2	TH12	2.3	182.7	127	6.31
TB13	746	72.7	-125	7.16	TH15	8.4	27.5	60	6.18
TB22	311	13.5	-130	6.63	TH1	6	544.4	210	6.08
TB15	937.7	19.3	-110	7.04	TH10	3.2	277.3	231	6
TB16	314.5	25.8	-115	6.74	TH2	2	487.6	130	5.87
TB20	746.3	6.9	-139	6.61	TH23	0.2	158.5	175	6.04
TB23	270	12.7	-110	7.01	TH3	1.5	560.2	261	6
TB17	224.2	21.5	-126	6.46	TH4	2.6	21.4	80	5.99
TB3	727	10.8	-136	7.14	TH11	8.9	479.8	158	6.56
TB12	931.5	2.9	-125	7.03	TH18	3.6	335.6	181	6.29
TB14	747.7	63.4	-115	7.15	TH8	6	253.2	235	5.85
TB5	416.3	0	-60	7.69	TH7	0.7	122.3	162	6.29
TB4	360.3	42.9	-130	7.34	TH6	0	242.8	200	6.19
TB6	315.5	61.2	-111	7.36	TH17	22.2	40.5	-13	7.03
TB7	101.1	42.7	-28	7.3	TH19	17.5	57.3	145	7.39
TB8	237.6	124.4	-98	7.17	TH20	2.4	0	24	6.51

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; 97-110

Optimum Mixture Designs in Constrained Experimental Regions - An Informative Review

Manisha Pal
Department of Statistics, Faculty of Science
St. Xavier's University, Kolkata, India

Received: 22 May 2024; Revised: 30 June 2024; Accepted: 02 July 2024

Abstract

In a mixture experiment, the response depends on the proportions of the mixing components. Canonical models of different degrees have been suggested by Scheffé (1958) to represent the mean response in terms of the mixing proportions, and optimum designs for estimation of the parameters of the models have been investigated by several authors. In most cases, the optimum design includes the vertex points of the simplex as support points of the design, which are not mixture combinations in the true non-trivial sense, and therefore are not acceptable to the practitioners. Further, in some situations, due to physical or economic limitations, the experimental region forms only a part of the simplex that does not cover the extreme points. The present paper gives a review of the available literature on optimum mixture experiments in regular subspaces of the simplex.

Key words: Mixture experiments; Restricted experimental region; Optimum designs.

AMS Subject Classifications: 62K99, 62J05

1. Introduction

A systematic study of optimum regression designs began with the pathbreaking work of Kiefer and Wolfowitz (1959). Soon after, various authors started to investigate optimality criteria for designs to estimate the model parameters (*cf.* Elfving, 1959; Karlin and Studden, 1966; Fedorov, 1971; Pukelsheim, 1993; Draper and Pukelsheim, 1996; Liski *et al.* 1998; Li *et al.*, 2005). A mixture experiment is a special case of a regression experiment, where the mean response is dependent on the mixing proportions of the ingredients in the mixture, rather than on their actual amounts. Thus, for a mixture experiment with q ingredients, the experimental region is defined by

$$\Xi = \left\{ (x_1, x_2, \ \ldots, \ x_q)^T : \ x_i \geq 0, \ i = 1\,(1)\,q, \ \sum_{i=1}^{q} x_i = 1 \right\}, \tag{1}$$

where $(x_1, x_2, \ \ldots, \ x_q)$ denote the mixing proportions. Graphically, (1) is defined by a simplex with vertices $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)$.

Corresponding Author: Manisha Pal
Email: manishapal2@gmail.com

There are high applications of mixture methodologies in different research areas, like (a) agricultural experiments, such as (i) intercropping (Dhekale *et al.*, 2003), (ii) split of total fertilizer application at different growth-stages of plants (Batra *et al.*, 1999), (iii) blend of waste water/saline water/marginal quality water for effective irrigation (Kan and Rapaport-Rom, 2012), (b) horticultural experiments, such as preparation of ready-to-serve beverages (Deka *et al.*, 2001), (c) animal nutritional experiments, such as feeding trials with several alternatives (Osborne and Mendel, 1921), (d) gasoline blending (Snee, 1981), (e) experiments with chemical pesticides (Deneer, 2000), and so on.

Mixture models, of the form $\eta_{\boldsymbol{x}} = \boldsymbol{f'(x)\beta}$, were first introduced by Scheffé (1958), who defined canonical models of degrees one to three to express the mean response in terms of the mixing proportions as follows:

Linear (homogeneous): $\eta_{\boldsymbol{x}} = \sum_i \beta_i x_i$

Quadratic: $\eta_{\boldsymbol{x}} = \sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j$

Full cubic: $\eta_{\boldsymbol{x}} = \sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k + \sum_{i<j} \delta_{ij}(x_i - x_j)$

Special cubic: $\eta_{\boldsymbol{x}} = \sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k$.

Scheffé (1958, 1963) also proposed the simplex lattice design and the simplex centroid design as suitable for parameter estimation in his proposed models. Later, other models, like the log-contrast model, Darroch-Waller quadratic mixture model, linear mixture models with synergism, were introduced.

Optimal designs for estimation of model parameters in various mixture models have been investigated by many researchers. Noteworthy are the studies by Kiefer (1961), Farrel *et al.* (1967), Atwood (1969), Galil and Kiefer (1977), Liu and Neudecker (1995), to name a few. Generally, the designs suggested for estimation and analysis in mixture experiments include the vertex points of the simplex, and such designs also turn out to be optimal designs. However, practitioners find such suggestions rather absurd and illogical as vertices of the simplex are not mixtures in the true sense, and they prefer to perform experiments excluding these points. Further, often due to physical or economic limitations, or interest of the experimenter, experiments may be confined to a sub-region of the whole experimental space. For example, when interest lies on the relationship among the ingredients, factorial arrangements can be used to analyze the response to ratios of ingredients (Kentworthy, 1963). Here only complete mixtures must be considered, that is, only mixtures where the proportion of each component is greater than zero. In agricultural/horticultural experiments, there are instances of usage of mixture experiments, and a growing interest in use of restricted subspaces of the simplex (Batra *et al.* 1999; Deka *et al.* 2001; Dhekale *et al.* 2003). Suggestion for the experimental region as a subspace of the simplex that does not include the vertex points are available in (Cornell, 2002). Though much research has been conducted to find appropriate designs for mixture experiments with restricted space, not much studies are available where the optimal design has been investigated.

This paper takes the readers on a journey through optimum designs when the experimental region is defined by a regular subspace of the simplex, such as an ellipsoid, a simplex within the simplex or a cuboid.

2. Restricted regions

The most common form of restricted region arises when one or more of the proportions of ingredients in the mixture are subjected to lower and/or upper bounds. This is very common in pharmaceutical experiments, horticulture experiments, agricultural experiments, gasoline blending, etc. The experimental region in such cases form a subset of the simplex. For example, if we have a 4-component mixture with the mixing proportions x_1, x_2, x_3 and x_4 having bounds $0.4 \leq x_1 \leq 0.6,\ \ 0.1 \leq x_2, x_3 \leq 0.5,\ \ 0.03 \leq x_4 \leq 0.08$, the experimental region is given by the bounded region within the simplex in Figure 1.

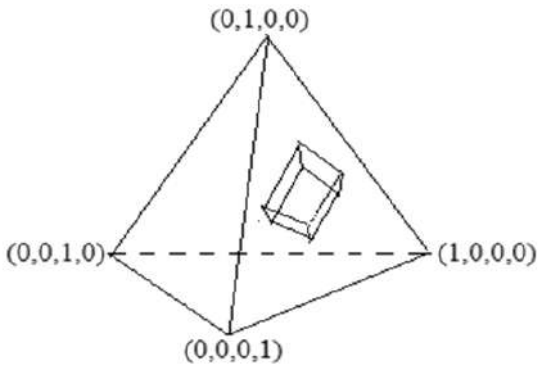


Figure 1: Experimental region within the simplex

In view of the bounds on the mixing proportions, the vertices of the simplex are excluded in the experimental region.

An interesting investigation carried out with such lower or upper bounds on mixing proportions, or on linear combinations of them, is due to Martin *et al.* (1999), who argued that theory cannot usually be used to obtain a good design. They discussed the algorithmic methods to obtain optimal designs, mainly using the D-optimality criterion, and sometimes the V – optimality criterion, and compared the algorithms using several published 3-component mixture examples. Their study was restricted to optimum designs for parameter estimation in Scheffé's canonical models. Later, Mandal *et al.* (2008) attempted to find the optimum design for estimation of the optimum mixing proportions in 2- and 3-component mixtures using Scheffé's quadratic mixture model, where one of the components is restricted by an upper bound less than unity. They used the pseudo-Bayesian approach due to Pal and Mandal (2006), and obtained the A-optimal design in the case of 2-component mixture. However, in the case of 3-component mixture, they could suggest an optimum design within six-point designs, but not within all competing designs. This instigated them to search further, and they came up with a seven-point design which was very close to the other design in terms of the criterion function. So, their suggestion was to start with any one of these designs, and use a standard numerical algorithm to reach the optimum design.

Other types of restricted experimental regions may be as given in Figure 2, As is noted, these regions have regular shapes, which are easy to study analytically, rather than very irregular regions within the simplex.

The restricted regions in Figure 2 also do not include the vertex points of the simplex.

An ellipsoidal experimental region often appears in pharmaceutical and engineering

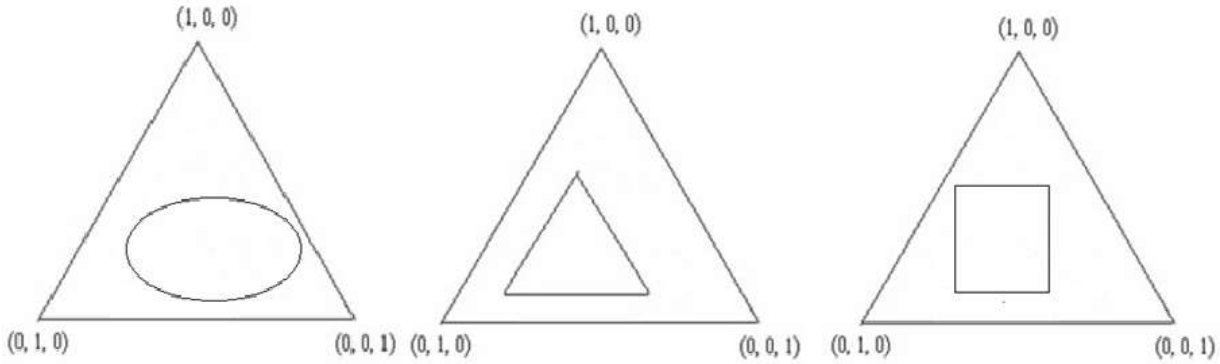


Figure 2: Some regular subspaces within the simplex forming the experimental regions

experiments. For example, Rais *et al.* (2004) used an ellipsoidal subregion of the space of mixture components for the optimization of a fluoroanhydrite-based self-levelling floor composition. A simplex within a simplex arises as an experimental region when the mixing proportions lie within a fixed range in (0,1). An example of this can be found in pharmaceutical experiment with oral tablets, where 2 or 3 polymers may be used with proportions having the same fixed non-zero bounds. Again, bounds on the mixing proportions may lead to a rectangular cuboid experimental region under certain conditions, as shown by Crosier (1990).

Restricted experimental region, ignoring the vertex points of the simplex, have been studies in mixture experiments to prescribe designs for parameter estimation (*cf.* Cornell, 2002). However, few authors attempted to find the optimum designs in such cases. Sections 3 - 5 review optimal designs for parameter estimation in Scheffé's first and second order models under regular experimental regions as indicated in Figure 2.

3. Ellipsoidal experimental region in the simplex

Mandal *et al.* (2015) were perhaps the first to attempt to find optimal design in an ellipsoidal region. For a q-component mixture experiment, they defined the constrained experimental region as

$$\Xi_0 = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_q)^T : x_i \geq 0, 1 \leq i \leq q, \sum_{i=1}^{q} x_i = 1, (\boldsymbol{x} - \boldsymbol{x}_0)^T H^{-2} (\boldsymbol{x} - \boldsymbol{x}_0) \leq 1 \right\}$$

where $\boldsymbol{x_0} = (1/q, 1/q, \ldots, 1/q)^T$ is the centroid of the simplex, and H is a non-singular diagonal matrix given by $H = diag\,(h_{11}, h_{22}, \ldots, h_{qq})$. The experimental region can be made to suit a specific situation by varying the h_{ii}'s.

The authors considered the case where $H \propto I_q$, an identity matrix. Then, for the transformation $\boldsymbol{x} \to \boldsymbol{z} = H^{-1}(\boldsymbol{x} - \boldsymbol{x_0})$, the domain of $\boldsymbol{z}$ comes out as $\{\boldsymbol{z} = (z_1, z_2, ..., z_n)^T : \boldsymbol{z}^T\boldsymbol{z} \leq 1, \boldsymbol{z}^T\mathbf{1}_q = 0\}$. A further transformation, *viz.* $\begin{bmatrix} u \\ \boldsymbol{v}^{(q-1)\times 1} \end{bmatrix} = Q\boldsymbol{z}$, where Q is an orthogonal matrix given by $\begin{bmatrix} q^{\frac{1}{2}}\mathbf{1}_q \\ P \end{bmatrix}$, and $\mathbf{1}_q$ is a $q \times 1$ vector with all elements unity, leads

to $u = 0$ and $\boldsymbol{v} = P\boldsymbol{z}$, with domain of v given by $\{\boldsymbol{v} = (v_1, v_2, \ldots, v_{(q-1)})^T : \boldsymbol{v}^T\boldsymbol{v} \leq 1\}$. One can easily express a Scheffé's model in terms of $\boldsymbol{v}$.

The problem of determining optimum designs in terms of $\boldsymbol{v}$ in its domain is a standard one in the context of response surface. Invariance structure, combined with the Loewner ordering (partial ordering of the information matrices), constitutes the Kiefer ordering, and this provides an effective tool in tackling optimal design problems of high dimension (*cf.* Pukelsheim, 1993). Using inverse transformation, it is easy to obtain the optimal design in terms of $\boldsymbol{z}$, and hence in terms of $\boldsymbol{x}$. Further, the optimum design in terms of $\boldsymbol{x}$ may not include the vertex points of the simplex, and, therefore, it will be different from the standard optimum design obtained over the whole simplex.

When Scheffé's first order mixture model is considered in the restricted space, the model in terms of $\boldsymbol{v}$ is also of first order, and to construct Kiefer optimal design on the experimental domain $\boldsymbol{v}^T\boldsymbol{v} \leq 1$ one needs to vary each of the $k = q - 1$ components of $\boldsymbol{v}$ on the two levels $\pm\ k^{-1/2}$ only. The design that assigns uniform weight to each of the 2^k vertices of $[-k^{-1/2}, k^{-1/2}]^k$ is the complete factorial design 2^k, and its optimality is established through the following lemma (*cf.* Pukelsheim, 1993):

Lemma 1: A first order design $D(n \times k)$ with k components is optimum in the sense of Kiefer ordering if $D^TD \ \propto \ I_k$.

Examples of first order optimal designs for the restricted region are obtained by exploiting the Kiefer optimal first order designs on the $\boldsymbol{v}$ – space and choice of H as follows:

(i) For $q = 2$, $v \in [-1, 1]$, and the Kiefer optimal design assigns equal mass, namely $\frac{1}{2}$, to the two extreme points $v = -1$ and $v = 1$. Accordingly, the optimal design on the original restricted domain has the support points

(a) $\left(\frac{\sqrt{3}+\sqrt{2}}{2\sqrt{3}}, \frac{\sqrt{3}-\sqrt{2}}{2\sqrt{3}}\right)$ and $\left(\frac{\sqrt{3}-\sqrt{2}}{2\sqrt{3}}, \frac{\sqrt{3}+\sqrt{2}}{2\sqrt{3}}\right)$ when $H = 3^{-1/2}\ I_2$,

(b) $\left(\frac{\sqrt{2}+1}{2\sqrt{2}}, \frac{\sqrt{2}-1}{2\sqrt{2}}\right)$ and $\left(\frac{\sqrt{2}-1}{2\sqrt{2}}, \frac{\sqrt{2}+1}{2\sqrt{2}}\right)$ when $H = 2^{-1}\ I_2$,

In general, for $q = 2$, Lemma1 establishes the Keifer optimality of the designs obtained for all H of the form $H = hI_q$, when $h \leq 2^{-\frac{1}{2}}$.

It is to be noted that for $H = 2^{-1/2}\ I_2$ the points in the experimental region Ξ_0 are restricted by $x_1(x_1-1) \leq 0$, which leads to the optimum design in the restricted space to have support points at $(1, 0)$ and $(0, 1)$ with equal masses. This is also the case in the unrestricted case.. Draper and Pukelsheim (1999) has already established the Kiefer optimality of this design in their ingenious way.

(ii) For $q = 3$, the Kiefer optimal design in the $\boldsymbol{v}$ – space has the supports $\left(-\frac{\sqrt{3}}{2},\ \frac{1}{2}\right)$, $\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$, $(0,\ -1)$, which, on inverse transformation, gives the support points of the optimal design in the restricted space as $(1/6, 1/6, 2/3), (2/3, 1/6, 1/6)$ and $(1/6, 2/3, 1/6)$ when $H = 6^{-1/2}\ I_3$, and $P = \begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$. This design is an axial design as noted in Figure 3 below.

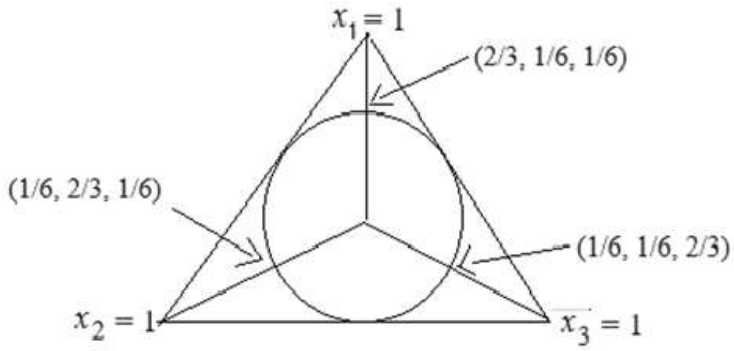


Figure 3: Support points of the optimum design in the case of 3-component mixture with ellipsoidal experimental region

Mandal *et al.* (2015) established that a design with any three points on the circumference of the circle in Figure 3, which form an equilateral triangle, is Kiefer-optimal. Thus, though the design indicated in Figure 3 is an axial design, a Kiefer optimal design is not necessarily so.

For Scheffé's quadratic mixture model, using the natural constraint $\sum_{i=1}^{k} x_i = 1$, it is possible to have a Kronecker product representation of the model, *viz.* $\eta_{\boldsymbol{x}} = (\boldsymbol{x} \otimes \boldsymbol{x})^T \boldsymbol{\beta}^*$, which makes it easier to represent the model as a quadratic model in $\boldsymbol{v}$ with its parameter vector, say τ^*, having a linear relationship with $\boldsymbol{\beta}^*$. As such, a design for estimating τ^* with Loewner Order dominance will also have Loewner Order dominance for τ^*. Pal *et al.* (2015) exploited this to obtain a Kiefer optimal design in the ellipsoidal region under Scheffé's quadratic mixture model.

Consider the central composite design (CCD) ξ^* in the $\boldsymbol{v}$ -space $\{\boldsymbol{v} : \boldsymbol{v}^T\boldsymbol{v} \leq 1\}$, which is a mixture of three blocks of designs, *viz.*

(i) cubes ξ_c , where ξ_c is a regular 2^{k-r} fraction of the full factorial design (with levels $\pm 1/\sqrt{k}$), of resolution V. (For $k \leq 5$, we have to take 2^k full factorial design);

(ii) stars ξ_s, where ξ_s is a set of star points of the form $(\pm 1, 0, 0, ..., 0), (0, \pm 1, 0, ..., 0), ..., (0, 0, ..., \pm 1)$;

(iii) centre points : $\xi_0 = \{\boldsymbol{v} : \boldsymbol{v}^T\boldsymbol{v} = 0\}$,

and ξ^* is defined as

$$\xi^* = (1 - \alpha)\xi_0 + \alpha\tilde{\xi}, \tag{2}$$

where $\tilde{\xi} = \frac{n_c\xi_c + n_s\xi_s}{n}$, $n_c = k^2, n_s = 2^{k-r}$, $n = 2^{k-r}n_c + 2kn_s, 0 < \alpha < 1$ Mandal *et al.* (2015) proved the following Theorem:

Theorem 1: The class of central composite designs (CCD), given by (2), is complete in the sense that given any design ξ, there is always a CCD of the form ξ^* given by (2) which is better in terms of

(i) Kiefer ordering

(ii) ϕ-optimality, provided it is invariant with respect to orthogonal transformation.

Through inverse transformation, it is then easy to conclude that the Kiefer optimal design for parameter estimation in Scheffé's quadratic mixture model in the ellipsoidal experimental region is obtained from a CCD, which is Kiefer optimal for the model in terms of $\boldsymbol{v}$.

For a 3-component mixture, ξ^* is obtained from the following blocks of designs:

(i) 4 star points: $(\pm 1, 0), (0, \pm 1)$

(ii) 2^2 factorial design points: $\frac{1}{\sqrt{2}}(-1, 1), \frac{1}{\sqrt{2}}(-1, 1), \frac{1}{\sqrt{2}}(1, -1), \frac{1}{\sqrt{2}}(1, 1)$

(iii) centre point: $(0, 0)$.

Then, for $H = \sqrt{6}I_3$, the optimal design in the restricted experimental region has the supports

$(1)\left(\frac{1}{3} + \frac{1}{2\sqrt{3}}, \frac{1}{3}, \frac{1}{3} - \frac{1}{2\sqrt{3}}\right); (2)\left(\frac{1}{3} - \frac{1}{2\sqrt{3}}, \frac{1}{3}, \frac{1}{3} + \frac{1}{2\sqrt{3}}\right); (3)\left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right);$

$(4)\left(\frac{1}{2}, 0, \frac{1}{2}\right); (5)\left(\frac{1}{3} + \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}, \frac{1}{3} + \frac{1}{3\sqrt{2}}, \frac{1}{3} - \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}\right);$

$(6)\left(\frac{1}{3} - \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}, \frac{1}{3} + \frac{1}{3\sqrt{2}}, \frac{1}{3} + \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}\right);$

$(7)\left(\frac{1}{3} + \frac{1}{2\sqrt{6}} + \frac{1}{6\sqrt{2}}, \frac{1}{3} - \frac{1}{3\sqrt{2}}, \frac{1}{3} - \frac{1}{2\sqrt{6}} + \frac{1}{6\sqrt{2}}\right);$

$(8)\left(\frac{1}{3} - \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}, \frac{1}{3} + \frac{1}{3\sqrt{2}}, \frac{1}{3} + \frac{1}{2\sqrt{6}} - \frac{1}{6\sqrt{2}}\right);$

$(9)\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$

which are presented in Figure 4.

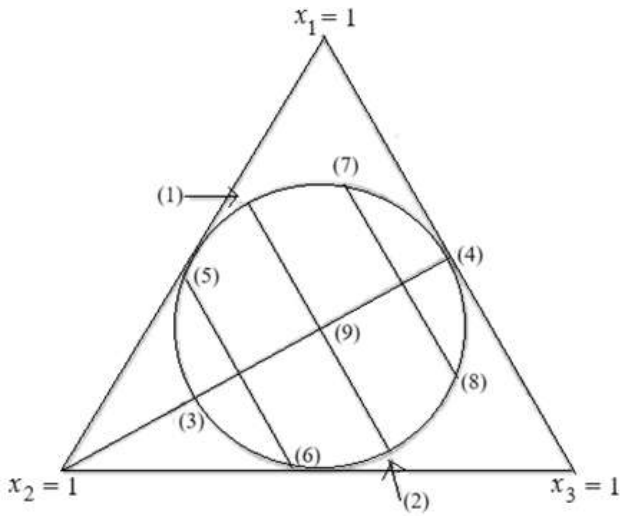


Figure 4: Support points of a Kiefer optimal design for the quadratic mixture model under ellipsoidal experimental region

It is noted that the Kiefer optimal design has 8 support points in the interior of the simplex, including the centroid, given by (9), and one on an edge, namely (4).

4. A restricted region in the form of a simplex within the unrestricted simplex

Mandal and Pal (2017) investigated the Kiefer optimal design when the experimental region is given by the form

$$\Xi_1 = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_q)^T : \frac{1}{q} - \frac{h}{q-1} \leq x_i \leq \frac{1}{q} + \frac{h}{q-1}, 1 \leq i \leq q, \sum_{i=1}^{q} x_i = 1 \right\},$$

where $h \in \left(0, \frac{q-1}{q}\right)$.

It is noteworthy that the centroid of the restricted region coincides with that of the simplex, *viz.* $\boldsymbol{x}_0 = \left(\frac{1}{q}, \frac{1}{q}, \ldots, \frac{1}{q}\right)^T$.

The transformation $\boldsymbol{x} \rightarrow \boldsymbol{z} = \frac{q-1}{qh}[\boldsymbol{x} - (\boldsymbol{x}_0 - \frac{h}{q-1}\mathbf{1}_q)]$ transforms the experimental region to

$$\Xi_{\boldsymbol{z}} = \left\{ \boldsymbol{z} = (z_1, z_2, \ldots z_q)^T : z_i \in [0.1], \ i = 1\,(1)\,q, \ \sum_{i=1}^{q} z_i = 1 \right\}, \tag{3}$$

which is same as the unrestricted experimental region Ξ.

The Kiefer optimal design in the permutation invariant class for estimation of the parameters of a first-degree or second-degree model, with unrestricted experimental region, is available in literature (Draper and Pukelsheim, 1999). This leads to the Kiefer optimal design for parameter estimation of the model in the restricted region owing to the 1:1 relation between $\boldsymbol{x}$ and $\boldsymbol{z}$.

For $q = 3$, the Kiefer optimal design in the $\boldsymbol{z}$-space has support points $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$ for Scheffé's linear homogeneous model, and $(1, 0, 0), (0, 1, 0), (0, 0, 1), (1/2, 1/2, 0)$, $(1/2, 0, 1/2)$ and $(0, 1/2, 1/2)$ for Scheffé's quadratic mixture model. This enables to find the support points of the Kiefer optimal design in the restricted $\boldsymbol{x}$-space as shown in Figure 5 for Scheffé's first order mixture model, and Figure 6 for Scheffé's quadratic mixture model. The points marked by alphabets in parentheses denote the support points.

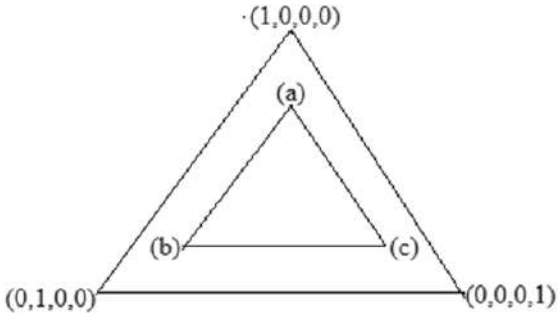
Remark: As the experimental region is well within the simplex, the vertex points of the simplex can never be included in a design.

5. Cuboidal experimental region in the simplex

A q-dimensional hypercube often defines the experimental region in industrial experimentation. In case of a cuboidal region, no result has been established so far that could help in finding the Kiefer optimal design. In view of that, Mandal and Pal (2017) attempted to find the D-optimal design for parameter estimation in the linear, homogeneous and quadratic mixture models due to Scheffé in the cuboidal region.

The restricted region within the simplex is defined by

$$\Xi_2 = \left\{ (x_1, \ldots, x_q)^T : 0 \leq x_{i0} - h_i \leq x_i \leq x_{i0} + h_i, \ \sum_{i=1}^{q} x_i = 1 \right\},$$

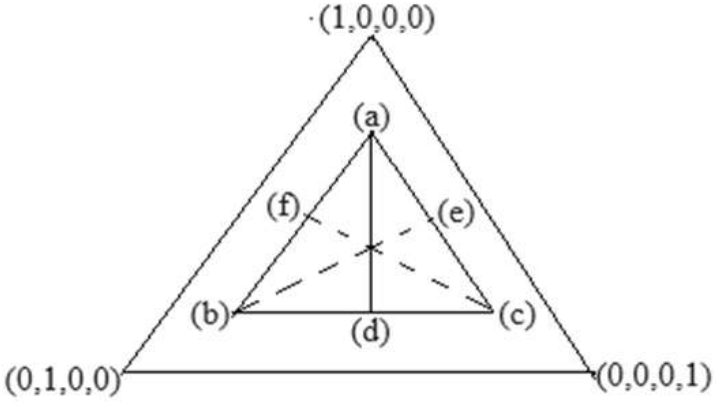


$$(\mathrm{a}) = \left(\frac{1}{3} + \frac{4}{9h}, \frac{1}{3} - \frac{2}{9h}, \frac{1}{3} - \frac{2}{9h}\right)$$

$$(\mathrm{b}) = \left(\frac{1}{3} - \frac{2}{9h}, \frac{1}{3} + \frac{4}{9h}, \frac{1}{3} - \frac{2}{9h}\right)$$

$$(\mathrm{c}) = \left(\frac{1}{3} - \frac{2}{9h}, \frac{1}{3} - \frac{2}{9h}, \frac{1}{3} + \frac{4}{9h}\right)$$

Figure 5: Support points of a Kiefer optimal design for the first-order mixture model in the experimental region Ξ_1



$$(\mathrm{a})\left(\frac{1}{3} + h, \frac{1}{3} - \frac{h}{2}, \frac{1}{3} - \frac{h}{2}\right); (\mathrm{b})\ \left(\frac{1}{3} - \frac{h}{2}, \frac{1}{3} + h, \frac{1}{3} - \frac{h}{2}\right); (\mathrm{c})\left(\frac{1}{3} - \frac{h}{2}, \frac{1}{3} - \frac{h}{2}, \frac{1}{3} + h\right)$$

$$(\mathrm{d})\left(\frac{1}{3} + \frac{h}{4}, \frac{1}{3} + \frac{h}{4}, \frac{1}{3} - \frac{h}{2}\right); (\mathrm{e})\ \left(\frac{1}{3} + \frac{h}{4}, \frac{1}{3} - \frac{h}{2}, \frac{1}{3} + \frac{h}{4}\right); (\mathrm{f})\left(\frac{1}{3} - \frac{h}{2}, \frac{1}{3} + \frac{h}{4}, \frac{1}{3} + \frac{h}{4}\right)$$

Figure 6: Support points of a Kiefer optimal design for the quadratic mixture model in the experimental region Ξ_1

where $\boldsymbol{x}_0 = (x_{10}, \ldots, x_{q0})^T$ is the centre of Ξ_2, $h_i \leq \min\left[x_{i0}, 1 - x_{i0}\right] \ \ \forall i = 1\,(1)\,q$, and it is assumed that $\boldsymbol{x}_0$ is the centroid of the simplex .

A transformation $\boldsymbol{x} \rightarrow \boldsymbol{z} = H^{-1}(\boldsymbol{x} - \boldsymbol{x}_0)$, where $H = Diag(h_1,\ h_2, \ldots, h_q)$, along with the natural constraint $\sum_{i=1}^{q} x_i = 1$, gives

$-1 \leq z_i \leq 1$, for $i = 1\,(1)\,q$, $\quad \sum_{i=1}^{q} h_i z_i = 0$.

For $H \propto I_q$, or $h_i = h$ for all i, the $\boldsymbol{z}$ – space reduces to

$$\Xi_{\boldsymbol{z}} = \left\{ \boldsymbol{z} = (z_1, z_2, \ldots\ z_q)^T : -1 \leq z_i \leq 1, ,\ i = 1\,(1)\,q,\ \sum_{i=1}^{q} z_i = 0 \right\}.$$

A further orthogonal transformation $\boldsymbol{z} \rightarrow \begin{bmatrix} u \\ \boldsymbol{v}^{(q-1)\times 1} \end{bmatrix} = \begin{bmatrix} \sqrt{q}\mathbf{1}_q \\ P \end{bmatrix} \boldsymbol{z}$, gives $u = 0$, and the range of values of v_i as $-c \leq v_i \leq c$, where $\boldsymbol{v} = (v_1, v_2, \ldots\ v_{q-1})$ and

$$c \leq c^* = \min_{1 \leq i \leq q} x_{i0} \left[\frac{1}{h_i^2} - \frac{1}{a - h_i^2} \right]^{1/2}, a = \sum_{i=1}^{q} h_i^2$$

(vide Cornell (2002), pp. 122). c^* gives the greatest possible distance from $\boldsymbol{x} = \boldsymbol{x}_0$ to the closest boundary opposite the vertex $x_i = 1$.

Expressing the Scheffé's response model in terms of $\boldsymbol{v}$, the problem of determining the D-optimum design in the domain $-c \leq v_i \leq c,\ i = 1\,(1)\,(q-1)$, is a standard one in the context of response surface and the results are well known (*cf.* Pukelsheim, 1993). Mandal and Pal (2017) made use of this to find the D-optimum design for parameter estimation in the model in $\boldsymbol{x}$ with cuboidal experimental region.

For Scheffé's first-degree model in $\boldsymbol{x}$, the model in terms of $\boldsymbol{v}$ is also a first-degree model with its parameters sharing a 1:1 relationship with the parameters of the model in $\boldsymbol{x}$, and the restricted $\boldsymbol{x}$ -space Ξ_2 is permutation invariant. Hence, If $\xi_{\boldsymbol{x}}$ is a design in Ξ_2 corresponding to a design $\xi_{\boldsymbol{v}}$ in the $\boldsymbol{v}$-space, and, if $\xi_{\boldsymbol{v}}$ is D-optimal in the $\boldsymbol{v}$–space, then $\xi_{\boldsymbol{x}}$ will also be D-optimal in Ξ_2.

For $q = 2$ and $h_i = h$ for all i, v is a single variable in the interval $[-c, c]$. In this case, the D-optimal design in the v – space assigns mass $1/2$ at each of the values $-c$ and $+c$. Accordingly, the D-optimal design in the restricted space of $\boldsymbol{x}$ puts equal masses at $(x_{01} - \frac{hc}{\sqrt{2}},\ x_{02} + \frac{hc}{\sqrt{2}})$ and $(x_{01} + \frac{hc}{\sqrt{2}},\ x_{02} - \frac{hc}{\sqrt{2}})$.

For $q = 3$, the D-optimal design assigns equal masses at its support points $(\pm c, \pm c)$. Reverse transformation gives the support points of the D-optimal design with equal masses in the restricted $\boldsymbol{x}$ – space as

$$\left(x_{01} + \frac{\left(\sqrt{3} - 1\right) hc}{\sqrt{6}}, x_{02} + \frac{2hc}{\sqrt{6}}, x_{03} - \frac{\left(\sqrt{3} + 1\right) hc}{\sqrt{6}} \right)$$

$$\left(x_{01} + \frac{\left(\sqrt{3} + 1\right) hc}{\sqrt{6}}, x_{02} - \frac{2hc}{\sqrt{6}}, x_{03} - \frac{\left(\sqrt{3} - 1\right) hc}{\sqrt{6}} \right)$$

$$\left(x_{01}-\frac{\left(\sqrt{3}+1\right)hc}{\sqrt{6}},x_{02}+\frac{2hc}{\sqrt{6}},x_{03}+\frac{\left(\sqrt{3}-1\right)hc}{\sqrt{6}}\right)$$
$$\left(x_{01}-\frac{\left(\sqrt{3}-1\right)hc}{\sqrt{6}},x_{02}-\frac{2hc}{\sqrt{6}},x_{03}+\frac{\left(\sqrt{3}+1\right)hc}{\sqrt{6}}\right)$$

These points lie within the cuboidal region Ξ_2.

Remark: Different choices of H lead to different optimal designs in the restricted space.

In the quadratic response model due to Scheffé, using similar transformations [$\boldsymbol{x} \rightarrow \boldsymbol{z} \rightarrow (0, \boldsymbol{v})$], and observing that there is a 1:1 relation between the parameters of the model in terms of $\boldsymbol{x}$ and that in terms of $\boldsymbol{v}$, the D-optimal design in the $\boldsymbol{v}$–space leads to the D-optimal design in the restricted $\boldsymbol{x}$–space through reverse transformation.

From Mandal (1989), the support points of the D-optimal design in the $\boldsymbol{v}$–space is obtained from the following result:

Theorem 2: D-optimum design in the $\boldsymbol{v}$–space is supported on the lattice of points with coordinates only 0 or $\pm c$.

For the case of 3-component mixture, the support points of the D-optimum design in the $\boldsymbol{v}$–space are the points $(0,0),(\pm c,0),(0,\pm c),(\pm c,\pm c)$.

Reverse transformation gives the support points of the D-optimum design in the restricted $\boldsymbol{x}$–space as

(i) (x_{01}, x_{02}, x_{03}) with mass 0;

(ii) $\left(x_{01}-\frac{hc}{\sqrt{2}},x_{02},x_{03}+\frac{hc}{\sqrt{2}}\right),\ \left(x_{01}+\frac{hc}{\sqrt{2}},x_{02},x_{03}-\frac{hc}{\sqrt{2}}\right),$

$$\left(x_{01}-\frac{hc}{\sqrt{6}},x_{02}-\frac{2hc}{\sqrt{6}},x_{03}+\frac{hc}{\sqrt{6}}\right),\ \left(x_{01}+\frac{hc}{\sqrt{6}},x_{02}+\frac{2hc}{\sqrt{6}},x_{03}-\frac{hc}{\sqrt{6}}\right)$$

each with mass 0.1325;

(iii) $\left(x_{01}+\frac{(\sqrt{3}-1)hc}{\sqrt{6}},x_{02}+\frac{2hc}{\sqrt{6}},x_{03}-\frac{(\sqrt{3}+1)hc}{\sqrt{6}}\right),\left(x_{01}+\frac{(\sqrt{3}+1)hc}{\sqrt{6}},x_{02}-\frac{2hc}{\sqrt{6}},x_{03}-\ \frac{(\sqrt{3}-1)hc}{\sqrt{6}}\right),$
$\left(x_{01}-\frac{(\sqrt{3}+1)hc}{\sqrt{6}},x_{02}+\frac{2hc}{\sqrt{6}},x_{03}+\frac{(\sqrt{3}-1)hc}{\sqrt{6}}\right),\left(x_{01}-\frac{(\sqrt{3}-1)hc}{\sqrt{6}},x_{02}-\ \frac{2hc}{\sqrt{6}},x_{03}+\frac{(\sqrt{3}+1)hc}{\sqrt{6}}\right)$

each with mass 0.1175.

6. Concluding remarks

The restricted experimental regions reviewed in this paper are proper subspaces of the simplex, and have suitable permutation invariance property. As such, it has been possible to characterize the optimal designs for Scheffé's linear and quadratic mixture models. Though

the derivations are non-trivial, the tools used in earlier studies have been exploited to find the optimum designs.

In case of the absence of symmetry and invariance, it would be very difficult to obtain the optimum designs. This is the case if (i) the restricted region has its centre/centroid different from the centroid of the unrestricted simplex, or (ii) $H \propto I_q$ does not hold for the ellipsoidal/cuboid region. These remain as open problems which perhaps would be very challenging to tackle analytically.

Acknowledgement

The author thanks the anonymous reviewer for the fruitful comments, which helped to improve the presentation of the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Atwood. C. L. (1969). Optimal and efficient designs of experiments. *Annals of Mathematical Statistics*, **40**, 1570-1602.

Batra, P. K., Parsad, R., Gupta, V. K., and Khanduri, O. P. (1999). A strategy for analysis of experiments involving split application of fertilizer. *Statistics and Applications*, **1**, 175–187.

Cornell, J. (2002). *Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data*. 3rd ed. Wiley, New York.

Crosier, R. B. (1990). Symmetry and Design in Mixture Experiments, Project.

Deka, B. C., Sethi, V., Parsad, R., and Batra, P. K. (2001). Application of mixtures methodology for beverages from mixed fruit juice/pulp. *Journal of Food Science and Technology*, **38**, 615–618.

Deneer, J. W. (2000). Toxicity of mixtures of pesticides in aquatic systems. *Pest Management Science*, **56**, 516–520.

Dhekale, J. S., Parsad, R., and Gupta, V. K. (2003). Analysis of intercropping experiments using experiments with mixtures methodology. *Journal of Indian Society of Agricultural Statistics*, **56**, 260–266.

Draper, N. R. and Pukelsheim, F. (1996). An overview of design of experiments. *Statistical Papers*, **37**, 1–32.

Draper, N. R. and Pukelsheim, F. (1999). Kiefer ordering of simplex designs for first and second degree mixture models. *Journal of Statistical Planning and Inference*, **79**, 325–348.

Elfving, G. (1959). Design of linear experiments. In Grenander, Ulf (ed.). *Probability and Statistics: The Harald Cramér Volume*. Almqvist & Wiksell, Stockholm; John Wiley & Sons, New York, 58–74.

Farrel, R. H., Kiefer, J., and Walbram, A. (1967). Optimum multivariate designs. In: Le Cam, L.M., Neyman, J. (Eds.), *Proc. Fifth Berkeley Symposium, Vol. 1, Berkeley*, 113-138.

Fedorov, V. V. (1971). The designs of experiments for linear optimality criteria. *Theory of Probability and its Applications*, **16**, 189–195.

Galil, Z. and Kiefer, J. (1977). Comparison of simplex designs for quadratic mixture models. *Technometrics*, **19**, 445-453.

Kan, I. and Rapaport-Rom, M. (2012). Regional blending of fresh and saline water: Is it efficient? *Water Resources Research*, **48**, W07517, doi:10.1029/2011WR011285.

Karlin, S. and Studden, W. J. (1966). Optimal experimental designs. *Annals of Mathematical Statistics*, **37**, 783-815.

Kentworthy, O. O. (1963). Factorial experiments with mixtures using ratios. *Industrial Quality Control*, **XIX**, 24-26.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Annals of Mathematical Statistics*, **30**, 271-294.

Kiefer, J. (1961). Optimum designs in regression problems, II. *Annals of Mathematical Statistics*, **32**, 298–325.

Li, K. H., Lau, T. S., and Zhang, C. (2005). A note on D-optimal designs for models with and without an intercept. *Statistical Papers*, **46**, 451–458.

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

Liu, S. and Neudecker, H. (1995). A V-optimal design for Scheffé's polynomial model. *Statistics and Probability Letters*, **23**, 253-258.

Osborne, T. B. and Mendel, L. B. (1921). Feeding experiments with mixtures of foodstuffs in unusual proportions. *Proceedings of the National Academy of Sciences*, **7**, 157–162.

Mandal, N.K. (1989). D-optimal designs for estimating the optimum point in a quadratic response surface - rectangular region. *Journal of Statistical Planning and Inference*, **23**, 243–252.

Mandal, N. K., Pal, M., Sinha, B.K., and Das, P. (2008). Optimum mixture designs under constraints on mixing components. *Statistics and Applications*, **6** (New Series), 189-205.

Mandal, N. K., Pal, M., Sinha, B.K., and Das, P. (2015). Optimum mixture designs in a restricted region. *Statistical Papers*, **56**, 105–119.

Mandal, N. K. and Pal, M. (2017). Optimum mixture designs in some constrained experimental regions. *Communications in Statistics – Theory and Methods*, **46**, 4240-4249.

Martin, R. J., Bursnall, M. C., and Stillman, E. C. (1999). Efficient designs for constrained mixture experiments. *Statistics and Computing*, **9**, 229-237.

Pal, M. and Mandal, N. K. (2006). Optimum designs for optimum mixtures. *Statistics and Probability Letters*, **76**, 1369-1379.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley, New York.

Rais, F., Jelidi, A., Kamoun, A., Chaabouni, M., Sergent, M., and Phan-Tan-Luu, R. (2004). Use of an ellipsoidal subregion of interest in the space of mixture components to the optimization of a fluoroanhydrite-based self-leveling floor composition. *Chemometrics and Intelligent Laboratory Systems*, **74**, 253–261.

Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society Series B*, **20**, 344–360.

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

Snee, R. D. (1981). Developing blending models for gasoline and other mixtures. *Technometrics*, **23**, 119–130.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 111-126

Issues in Estimating Disease Specific Incidence Rates in Long-Term Follow-up in Childhood Cancer Survivors

Deo Kumar Srivastava[1], Kirsten Ness[2], Melissa Hudson[2,3], Sarmistha Das[4,5,6] and Shesh N. Rai[4,5,6]

[1]*Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[2]*Department of Epidemiology & Cancer Control, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[3]*Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, 38105, USA*

[4]*Department of Biostatistics, Health Informatics and Data Science, University of Cincinnati, Cincinnati, Ohio, 45267, USA*

[5]*Cancer Data Science Center, University of Cincinnati College of Medicine, Cincinnati, Ohio, 45267, USA*

[6]*Biostatistics and Informatics Shared Resource, University of Cincinnati Cancer Center, Cincinnati, Ohio, 45267, USA*

Received: 18 June 2024; Revised: 10 July 2024; Accepted: 12 July 2024

Abstract

With significant advances in the treatment and supportive care, the overall 5-year survival rate for pediatric cancers in high-income countries, such as USA, exceeds 85%, see Ehrhardt, *et al.* (2023), SEER (2022) and it is expected that the number of survivors will exceed 580,000 by year 2040. However, this comes at a high cost of treatment and cancer related long-term sequalae. To better characterize and to develop interventions/screening guidelines to mitigate the long-term effects of these adverse events researchers in North America and Europe established large-cohort retrospective studies with prospective follow-up assessments, see Robison, *et al.* (2009), Winther, *et al.* (2015), Park, *et al.* (2012). However, it is logistically impossible to follow the survivors continuously and this information is usually collected through cross-sectional surveys at various times from cancer diagnosis, which leads to interval censored data since the exact time of the onset of the adverse event of interest is unknown. However, if this risk could be characterized in a continuous manner, then appropriate screening guidelines or interventions could be implemented. Our primary focus is on estimating the incidence rates (cumulative incidence) of a particular outcome of interest *e.g.* cardiovascular events using interval censored data. In this exposition we utilize SJLIFE cohort and propose the use of multi-state survival framework for modeling incidence rates and risk factors associated with it. We also highlight the use of multi-state models for analyzing more complicated relationships and identify some challenges associated with the analysis of such data.

Corresponding Author: Deo Kumar Srivastava
Email: kumar.srivastava@stjude.org

Key words: Cross-sectional survey data; Prospective follow-up; Multi-state model, Interval-censored data.

1. Introduction

The 5-year survival rate for childhood cancer survivors now exceeds 85% but they are at an increased risk of developing long-term chronic health conditions as a result of their cancer or its treatment. As the childhood cancer survivor population increases, understanding the long-term impact of cancer on health during adulthood is important to guide the development of interventions to improve the quality of life and duration of survival. The Cancer Control and Survivorship Program in the Comprehensive Cancer Center at St. Jude Children's Research Hospital is a multidisciplinary research program that strives to improve the quality of life of individuals surviving childhood cancer by translating the research findings into effective strategies to reduce treatment related complications. The St. Jude Lifetime Cohort Study (SJLIFE) is a cancer epidemiologic cohort established to facilitate longitudinal clinical evaluation of health outcomes in childhood cancer survivors across the lifespan. A detailed description of the study along with schema for longitudinally follow-up can be obtained by visiting the St. Jude Cloud portal (https://www.stjude.cloud/) and going through the Cancer Survivorship tab and clicking on St. Jude LIFE study (SJLIFE), Howell, *et al.* (2021). SJLIFE was activated in 2007 with initial eligibility for participation including 10+ year survivor of pediatric cancer, treated, or followed at St. Jude from 1962 to 2012 who were at least 18 years of age. Eligibility criteria were modified in 2015 to include 5-year survivors. Eligible survivors are periodically invited to return to St. Jude for comprehensive clinical evaluations that involves completion of questionnaire with patient-reported outcomes, collection of biological specimens, and systematic evaluation of organ function including metabolic, cognitive and neuromuscular status.

Howell, *et al.* (2021) reported an update of the cohort progress. Among 8192 eligible survivors, 6560 have agreed to participate and 5,223 have completed baseline on-campus evaluations. The median [range] age at evaluation was 32 [7.0 – 71.9]. Participants are invited to return for follow-up visits in 3–5 years intervals. Study findings from these evaluations have enabled characterization of multimorbidity experiences by survivors many years after treatment for childhood cancer. This is highlighted in the study by Bhakta, *et al.* (2017) that used St. Jude modified National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) 4.03, Hudson, *et al.* (2017), and graded 168 chronic conditions within 13 organ systems. The CTCAE grades correspond to grade 1 (mild), grade 2 (moderate), grade 3 (severe/disabling), grade 4 (life-threatening) and grade 5 (death). Often, the focus is on modeling grade 3 or higher chronic conditions. The details of all the chronic conditions within each organ system can be obtained from Bhakta, *et al.* (2017), Supplementary Table $S1$. They grouped the chronic conditions into 13 organ systems as shown in Table 1.

In addition to studying different outcomes the identification of the appropriate study cohort at risk for developing the outcomes of interest is equally important. There are multiple factors that need to be considered in the selection of eligible subjects (survivors) in the study cohort. Because this cohort was originally constructed retrospectively, cohort entry and exit

Table 1: Number of chronic conditions within each grouped category for each organ system

Organ System	Grouped Condition Category
Cardiovascular	Myocardial Infarction (1), Arrhythmias (6), Cardiovascular Dysfunctions (4), Structural Heart Defects (4), Vascular Diseases (4), Essential Hypertension/Dyslipidemia (3)
Respiratory	Asthama (1), Obstructive Respiratory Disorders (2), Functional Pulmonary Deficits (3), Respiratory Parenchymal Diseases (7)
Gastrointestinal	Esophageal Disorders (3), Disorders of the GI Tract (11), Inflammatory Disorders (8), Hepatic Disorders (6), Disorder of the Gallbladder (1)
Reproductive	Disorders of the Female Reproductive System (8), Disorders of the Male Reproductive System (5), Condition affecting the Pituitary (1)
Endocrine	Growth Hormone Deficiencies (2), Overweight/Underweight (2), Thyroid Disorders (4), Parathyroid Disorders (2), Abnormal Glucose Metabolism (1), Conditions affecting the Pituitary (3)
Renal	Kidney Injuries (2), Obstructive Urinary Disorders (3), Hematuria (1)
Musculoskeletal	Amputation (1), Osteoporosis (1), Joint Diseases (3), Peripheral Musculoskeletal Disorders (6), Spine Disorders (4)
Neurology	Strokes (3), Central Nervous System Disorders (9), Mixed Nervous System Disorders (4), Peripheral Nervous System Disorders (6), Seizure (1), Severe Headache (1)
Immunology and Infections	Immunologic Disorders (2), Frequent/Recurrent Infections (8), Chronic Infections (7)
Hematology	Hematologic Disorders (7)
Auditory	Hearing Loss (1)
Second Neoplasms	Secondary and Recurrent Malignancy (1)
Ocular	Ocular Disorders (4)

Note: The numbers in the brackets indicate the number of chronic conditions within each grouped category within an organ system.

are heterogenous, *i.e.* their follow-up times are not equally spaced, and subjects may enter the cohort or leave the cohort at any follow-up times, including baseline (T_0), longitudinal follow-up (T_L) and date of death (T_D) *etc.* These are outlined in Table 2.

A thorough understanding of these chronic conditions, their prevalence and associated risk factors can provide valuable information which could be used to improve future treatment plans. For our discussions, we will focus on Cardiovascular Dysfunction (CD) within the Cardiovascular System (22 individual chronic conditions), which has four individual chronic conditions (cardiomyopathy (CAD), Right ventricular systolic dysfunction (RVSD), Cor Pulmonale (CP) and Pulmonary Hypertension (PH). The discussion can be easily generalized to all chronic conditions within cardiovascular system or across other organ systems as well. CAD refers to problems with heart muscles that make it harder for the heart to pump blood and, if untreated, can lead to heart failure or cardiac arrest. Similarly, RVSD if untreated could lead to heart failure or myocardial infarction, CP is an alteration in the structure and function of the right ventricle of the heart caused by a primary disorder

Table 2: Issues of heterogeneity in identifying the study cohort

Time points	Challenges
T_0	1. Should $T_0 = 5$ years since the primary diagnosis (eligibility criterion). What is the rationale for 5 years, why not 2, 3, 4, 6, 7, 8, 9, 10 years? 2. Should T_0 depend on multiple factors, including patient's age, time since treatment, disease type *etc.*?
T_L	1. If we need to choose only on long-term follow-up time to collect information, should T_L be largest or shortest or somewhere in between, when patients have more than one follow-up visits.
T_D	1. Eligibility restricted to 5-year survivors potentially introduces survival bias 2. Not every eligible patient in the the SJLIFE cohort visits clinics. They may be lost to follow-up (true censoring) or died but this information may not be accurately recoded. Should T_D and patient characteristics, including comorbidity at the time of death be obtained from other sources (such as death registry).

of the respiratory system, and PH is a condition that affects blood vessels in the lungs and makes heart work harder than normal to pump blood into lungs.

Among childhood cancer survivors, cardiovascular events (CEs) are among the top nonmalignant causes of death (Armstrong, *et al.* (2009)). This is due to the damage to cardiomyocytes caused by chemotherapy and chest radiation therapy received during the cancer treatment (Hammoud, *et al.* 2024). Even though certain chemotherapy exposures such as anthracycline are well known for associations with cardiotoxicity (Ehrhardt, *et al.* 2023), they continue to be used to treat cancer because of their curative benefits. Improved characterization of the cumulative incidence of CEs may facilitate opportunities for intervention to improve/preserve cardiac health. This motivates us to estimate the cumulative incidence (CI) of the CEs in childhood cancer survivors who completed their baseline evaluation because such information could be used to help researchers identify the best time to intervene. More information regarding the causes, treatment, and prevention of cardiotoxicity can be found in a comprehensive review by Koutsoukis, *et al.* (2018).

For our discussion we will focus on CAD (a CD), whose exact timing of development is unknown since SJLIFE participants are not followed continuously in real time. However, health related information is ascertained when participants came to campus for their baseline and follow-up visits, either on their scheduled visits or in an ad hoc manner to participate in ancillary studies. Although the exact time of CAD is unknown, the current status is available and from there we know that CAD symptoms that motivated medical attention occurred sometimes in the interval between the two follow-up visits. Such dataset can be characterized as case I interval-censored data, see Sun (2006), Rai (2008).

Remark: It may be noted that the approaches presented in this article would be applicable to the phase IV clinical trials in the context of drug development where the focus would be on monitoring for long term toxicities/side effects of an approved drug, see Zhang, *et al.* (2016).

We propose the use of multi-state models for estimating the cumulative incidence

rate for the event of interest. Multi-state models are extensions of the survival models, which are usually analyzed under the Cox proportional hazard model assumption. However, multi-state models have the advantage of providing more insight into disease process and progression as each transition, from one state to another, can be modeled and the covariates could be incorporated *e.g.* using Cox proportional hazards model. It further allows for simultaneous modeling of competing causes of death or morbidities; see Eulenburg, *et al.* (2015). The simplest extension of the survival model is the continuous time progressive three-state model *e.g.* see van den Hout (2017, Chapter 3). It is assumed that the survivors are followed longitudinally and the status of the survivors alive at the pre-specified observation time is available and that follow-up time for survivors on study can vary. In a progressive illness/death model as illustrated in Figure 1. For example, in the context of the study discussed above all survivors who survived for at least 5 years from their date of diagnosis will be assumed to be in State 1 (all are alive with no CAD) and then after a median follow-up of about 25 years were enrolled in SJLIFE and systematically evaluated for CAD. At that point each survivor could take one of the three paths: 1) remain alive with no CAD (State 1); 2) develop CAD but remain alive (moving from State 1 to State 2); 3) progress to death due to CAD or otherwise (moving from State 1 to State 3). For patients that reached State 2, they could also have two options: 1) remain alive with CAD (State 2); or 2) progress towards death or cardiac failure (moving from State 2 to State 3).

$\lambda_1(u)$ represents the transition intensity from State 1 to State 2; $\lambda_2(t)$ is the transition intensity from State 1 to State 3; $\lambda_3(t|u)$ is the transition intensity from State 2 to State 3. In general, the focus could be on estimating $\lambda_1(u)$, the transition intensity rate for patients

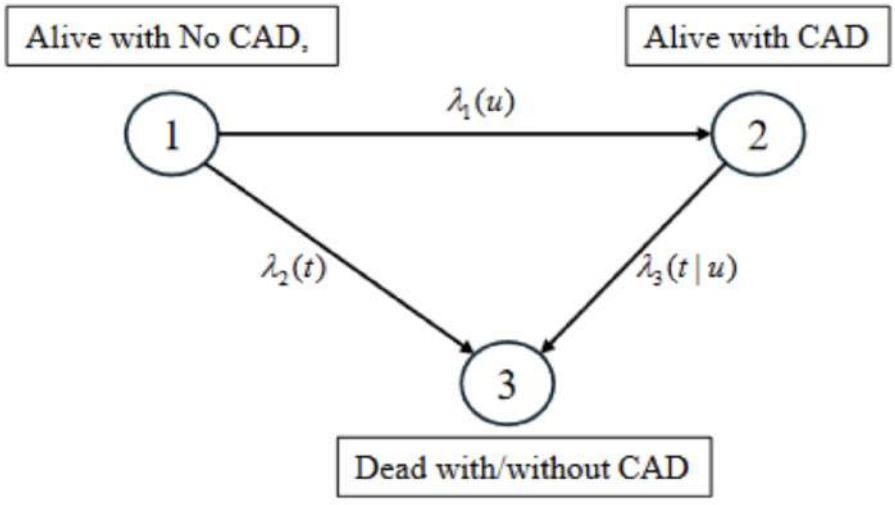


Figure 1: Three-states illness/death model of patients developing CAD

progressing from the initial state (State 1) of being normal to developing CAD (State 2) but one could be interested in estimating the incidence rates $\lambda_2(t)$ or $\lambda_3(t|u)$. This can be done using a parametric or a semi-parametric framework. In section 2 and 3 we will outline the parametric and semi-parametric approaches, respectively. In section 4, we outline the complexities involved when the interest could be in estimating the incidence rates of two or more events of interest.

2. Parametric modeling

Let $\{X(t)\}$ represent the stochastic process which identifies the state occupied a survivor at time t if we start with n survivors in the cohort then at time $t = 0$ all survivors will be in state 1 and has not experienced CAD (the event of interest). Let T be a random variable denoting the observation time (death, CAD or no event observed in the study duration) and U denotes the time to CAD. Thus, $X(t) = 1$, $X(t) = 2$ and $X(t) = 3$ represent the current status of the survivor; alive without CAD, alive with CAD or dead with or without CAD, respectively. We assume that the occurrence of CAD is irreversible. With the intensity functions shown in Figure 1 one can easily obtain the pseudo-survival functions, see Rai, *et al.* (2013) and van den Hout (2017), as follows:

$$Q_i(t) = exp\{-\int_t^0 \lambda_i(\nu)d\nu\}, \text{for } i = 1, 2 \tag{1}$$

and

$$Q_3(t|u) = exp\{-\int_t^u \lambda_3(\nu|u)d\nu\} \tag{2}$$

It may be noted that probability of surviving without experiencing CAD or death beyond time t can be represented as,

$$Q(t) = exp\{-\int_0^t (\lambda_1(\nu) + \lambda_2(\nu))d\nu\} = Q_1(t)Q_2(t), \tag{3}$$

and the survival function can be obtained as,

$$\begin{aligned} S(t) &= P(X(t) = 1) + P(X(t) = 2) \\ &= Q(t) + \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du \end{aligned} \tag{4}$$

2.1. Construction of the likelihood

The likelihood for the three-state model can be constructed in the following manner. Let θ denote the vector of parameters including transition intensities. Let t be the realization of the *r.v.* T and let Δ_i denote the contribution to the likelihood for the i^{th} survivor for $i = 1, 2, \ldots, n$. Then the likelihood function $L(\theta) = \prod_{i=1}^n \Delta_i$ Within this framework the survivor will be in one of the four distinct types of observations and their contribution to the likelihood will be as follows:

(i) Death without CAD, $T = t$, $X(t^-) = 1$, and $L_1(t) = \lambda_2(t)Q(t)$,

(ii) Alive without CAD, $T > t$, $X(t) = 1$, and $L_2(t) = Q(t)$,

(iii) Death with CAD, $T = t$, $X(t^-) = 2$, and $L_3(t) = \int_0^t \lambda_1(u)Q(u)\lambda_3(t|u)Q_3(t|u)du$,

(iv) Alive with CAD, $T > t$, $X(t) = 2$, and $L_4(t) = \int_0^t \lambda_1(u)Q(u)Q_3(t|u)du$

The likelihood function depends on in addition to the observation time and status, but is suppressed for convenience.

Parametric modeling is appealing as one can easily obtain the estimates and perform the inference using likelihood approaches. Among the class of parametric distributions, *e.g.* see Srivastava, *et al.* (2018), ven den Hout (2017), commonly used distributions are exponential, piecewise exponential and Weibull but other distributions such as log-normal, Gamma, log-logistic or Gompertz distributions could also be used. The data observed for each survivor i, $i = 1, 2, \ldots, n$, at a particular time, consists of triplet $(t_i, \delta_i, \gamma_i)$ where t_i is the observation time and,

$$\delta_i = \begin{cases} 1, & \text{if } i^{th} \text{ survivor died at time } t_i \\ 0, & \text{if } i^{th} \text{ survivor was alive without CAD at time } t_i \end{cases}$$

and

$$\gamma_i = \begin{cases} 1, & \text{if } i^{th} \text{ survivor had CAD at time } t_i \\ 0, & \text{if } i^{th} \text{ survivor did not have CAD at time } t_i \end{cases}$$

Then, the log-likelihood function can be written as,

$$l(\theta) = \sum_{i=1}^{n} [a_i log L_1(t_i) + b_i log L_2(t_i) + c_i log L_3(t_i) + d_i log L_4(t_i)], \tag{5}$$

where, $a_i = \delta_i(1 - \gamma_i)$, $b_i = (1 - \delta_i)(1 - \gamma_i)$, $c_i = \delta_i\gamma_i$, and $d_i = (1 - \delta_i)\gamma_i$.

The contributions that a survivor makes to the likelihood depends on the status and the underlying distribution. It may be worth noting that often the interest would be in estimating the cumulative incidence of the event of interest (CAD) at particular time points, *e.g.* 5-year of 10-year. Exponential distribution is the simplest model because it assumes the intensity function to be constant across time. However, this is not a plausible assumption when the follow-up time is long and there is the possibility of the intensity function changing over time. To provide more flexibility in modeling such data Rai, *et al.* (2013) proposed to use piecewise exponential distribution in estimating $\lambda_1(u)$. However, this poses the problem of knowing exactly when the incidence rate changes, how many change points are needed and that the incidence rates are constant within each piece. Srivastava, *et al.* (2018) used Weibull distribution to circumvent these limitations. Pradhan and Kundu (2014) also used Weibull distribution as the underlying lifetime distribution for the interval-censored data but suggested using the EM algorithm approach.

For exponential distribution the intensity rates are constant, *i.e.* $\lambda_i(t) = \lambda_i$ for $i = 1, 2$ and $\lambda_3(t|u) = \lambda_3$. Then, the contributions to the likelihood will be as follows:

$$L_1(t) = \lambda_2 \exp(-(\lambda_1 + \lambda_2)t) \tag{6}$$

$$L_2(t) = \exp(-(\lambda_1 + \lambda_2)t) \tag{7}$$

$$L_3(t) = \frac{\lambda_1\lambda_3}{\lambda_1 + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_1 + \lambda_2)t)) \tag{8}$$

$$L_4(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_1 + \lambda_2)t)) \tag{9}$$

Now if one assumes that the incidence rate for CAD may change say at t_c years (say,

$t_c = 5$) then one could use piecewise exponential distribution that would imply that $\lambda_1 = \lambda_{11}$ if $t < t_c$ and λ_{12} if $t \geq t_c$. The piecewise exponential assumption could be extend to other pieces of multi-state model if there is evidence of shift in incidence rates over time. The contributions to the likelihood will be as follows:

$$
\begin{aligned}
&L_1(t) = \lambda_2 \exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_{12}t) \exp(-\lambda_2 t) \\
&L_2(t) = \exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_{12}t) \exp(-\lambda_2 t) \\
&L_3(t) \\
&= \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3}(\exp(-\lambda_3 t) - \exp(-(\lambda_{11} + \lambda_2 + \lambda_3)t)) \text{ if } t < t_c \\
&= \frac{\lambda_{11}\lambda_3}{\lambda_{11} + \lambda_2 - \lambda_3} \exp(-\lambda_3 t)(1 - \exp(-(\lambda_{11} + \lambda_2 + \lambda_3)t_c) + \frac{\lambda_{12}\lambda_3}{\lambda_{12} + \lambda_2 - \lambda_3} \\
&\exp(-(\lambda_{11} - \lambda_{12})t_c - \lambda_3 t)(\exp(-(\lambda_{12+\lambda_2-\lambda_3})t_c)) - \exp(-(\lambda_{12} + \lambda_2 - \lambda_3)t) \text{ if } t \geq t_c \\
&L_4(t) = L_3(t)/\lambda_3
\end{aligned}
$$

For Weibull distribution the intensity function can be defines by $\lambda_i(t) = \eta_i \omega_i t^{\omega_i - 1}$, for $i = 1, 2$ and $\lambda_3(t|u) = \lambda_3^{SM}(t|u) = \eta_3\omega_3(t)^{\omega_3 - 1}$ under the assumption of a semi-Markov process or $\lambda_3(t|u) = \lambda_3^{M}(t|u) = \eta_3\omega_3(t-u)^{\omega_3 - 1}$ under the assumption of a Markov process. This leads to $Q_3^{SM}(t|u) = \exp(-\eta_3(t^{\omega_3} - u^{\omega_3}))$ under semi-Markov assumption and $Q_3^{M}(t|u) = \exp(-\eta_3(t-u)^{\omega_3})$ under Markov assumption, see Kalbfleisch and Lawless (1985) and Hazerlak, *et al.* (2003). The contribution to the likelihood is provided below:

$$
\begin{aligned}
&L_1(t) = \eta_2\omega_2 t^{\omega_2 - 1} \exp(-\eta_1 t^{\omega_1} - \eta_2 t^{\omega_2}) \\
&L_2(t) = \exp(-\eta_1 t^{\omega_1} - \eta_2 t^{\omega_2}) \\
&L_3(t) = \int_0^t Q(u)\lambda_1(u) Q_3^{SM}(t|u)\lambda_3^{SM}(t|u)du \quad \text{under semi-Markov assumption} \\
&= \int_0^t Q(u)\lambda_1(u) Q_3^{M}(t|u)\lambda_3^{M}(t|u)du \quad \text{under Markov assumption} \\
&L_4(t) = \int_0^t Q(u)\lambda_1(u) Q_3^{SM}(t|u)du \quad \text{under semi-Markov assumption} \\
&= \int_0^t Q(u)\lambda_1(u) Q_3^{M}(t|u)du \quad \text{under Markov assumption}
\end{aligned}
$$

Now, using the above contributions to the likelihood one can perform the likelihood estimation and obtain confidence intervals.

2.2. Incorporating covariates

As we have noted before, one of the long-term consequences of cancer patients treated with cardiotoxic therapy (treated with anthracycline and/or chest radiation) is that they are at a very high risk of developing CAD. Let us denote this group as AR (At Risk group), and let NR represent the groups of survivors who were not treated with cardiotoxic therapy.

Within this context, it would be important not only to know the onset time of these CADs but it would be important to know, from a clinician's perspective, the risk associated with therapeutic exposure and other risk factors such as age at diagnosis or sex *etc.* Thus, incorporation of covariates is an important issue. van den Hout (2017) proposes to use proportional hazard type to regression framework represented as

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta'X}) \tag{10}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of parameters and $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ is the vector of fixed covariates without an intercept term and $h_0(t)$ denotes the baseline hazard which could be modeled using any of the above mentioned parametric distributions. Alternatively, one can assess the impact of the covariates on the incidence rates by modeling the scale parameter as function of the parameters. For exponential distribution it would be equivalent to modeling $\lambda = \exp(\boldsymbol{\beta'X})$ and for Weibull distribution it would amount to modeling $\eta = \exp(\boldsymbol{\beta'X})$.

3. Semi-parametric modeling

Both Rai, *et al.* (2013) and Srivastava, *et al.* (2018) used a parametric approach with the assumption that the transition intensity rates from different risk groups are independent. Subjects in different groups are indeed independent, but the rates can be modeled assuming a dependence structure. In other words, the parametric approach may not be suitable if the underlying parametric assumptions do not hold. To overcome this limitation, we propose a semi-parametric approach with EM algorithm to model CI within the illness-death framework. Rai and Matthews (1997) introduced discrete scale models for estimating the transition intensity rate in a survival-sacrifice experiment using EM algorithm. Later, Rai, *et al.* (2000) extended a similar methodology to a mixed-scale model with an EM algorithm approach as well. EM algorithm is a powerful procedure to use when no closed-form solution can be obtained from the likelihood function. Besides the traditional EM algorithm, Rai and Matthews (1993) introduced a modified approach that could potentially save time (fewer iterations). A detailed application demonstration of EM algorithm can be found in Gunaratnam and Rai (2019). Additionally, when dealing with high dimensional parameters, the regular simple case likelihood function will not work. Rather, the use of profile likelihood should be considered, see Murphy and Van Der Vaart (2000).

Given that the data on CAD is collected intermittently the occurrence of the events falls in the category of interval-censored data as discussed above. To better understand the problem at hand, let us define Year 1 as 5 years after treatment completion and Year 2 as 6 years after treatment completion, a patient can choose to participate in the study either at Year 1 or Year 2, or even later. The challenge comes when we build the likelihood functions that will be discussed in the next section.

The following table reflects the data characteristics in a complete data setting. Table 3 corresponds to survivors who have come for SJLIFE evaluations for the first time and assume the maximum follow-up time is 5 years after cohort entry.

It is assumed that we are interested in looking at intensity rates up to five years. Let λ_i, $i = 1, 2, \ldots, 5$ be the intensity rates at Year i (Year 1, Year 2, *etc.*). Let n_{+i} $i = 1, 2, \ldots, 5$

Table 3: Data characteristics in a complete data setting

Year	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n_{+5}	Rate
1	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	λ_1
2		r_{22}	r_{23}	r_{24}	r_{25}	λ_2
3			r_{33}	r_{34}	r_{35}	λ_3
4				r_{44}	r_{45}	λ_4
5					r_{55}	λ_5
	r_{+1}	r_{+2}	r_{+3}	r_{+4}	r_{+5}	

be the number of subjects that came to the clinic in year i. For example, n_{+1} represents the patients who came to the clinic with 1 year of follow-up, and n_{+2} represents the patients who came to the clinic with 2 years of follow-up. Within any year i let there be r_{+i}, $i = 1, 2, \ldots, 5$ events (survivors with CAD, abnormal). For example, when an abnormal survivor visited in Year 2, the survivor may have become abnormal either in Year 1 or in Year 2, but that information is unknown. Therefore, we define $r_{+2} = r_{12} + r_{22}$ as the total number of events in Year 2 in which r_{11} represents the events that occurred during Year 1 and r_{22} represents the events that occurred during Year 2.

In summary, estimating transition intensity rates in a three-states illness-death model, such as the SJLIFE study, is not a simple task. The existing approach such as constructing the likelihood function is very complicated since the function is not in a closed form. Furthermore, sometimes the exponential model and the Weibull model might not be appropriate as the underlying assumptions might not always hold. This motivates us to construct a non-parametric model. In combination with EM algorithm, the transition intensity rate can be easily obtained.

Likelihood based approach

Including all the characteristics of the data in the likelihood is somewhat challenging in our situation. However, here we present an approach that would be appropriate for the data that we have in hand. For simplicity purposes, we will only show time points up to three years ($M = 3$).

We define the table above as complete data since we specifically know which survivors are abnormal. For incomplete data, we define (n_{+k}, r_{+k}), $k = 1, 2, \ldots, M$ and $r_{+k} = \sum_{j=1}^{k} r_{jk}$, $k = 1, 2, \ldots, M$ represents the total number of survivors with abnormality with k years of follow-up and n_{+k} represents the survivors who come to the clinic with k years of follow-up. In our case, we can write it as $r_{+1} = r_{11}$, $r_{+2} = r_{12} + r_{22}$, $r_{+3} = r_{13} + r_{23} + r_{33}$. In the incomplete data setting, we only know r_{+k} but not the actual number of abnormal cases within each year of follow-up. Since we have two independent groups, AR (At Risk) and NR (Not at Risk), the likelihood function can be defined for each group independently, which have a similar form. Let

$$
\begin{aligned}
z_{+j} &= n_{+j} - r_{+j}, j = 1, 2, 3 \\
p_{+1} &= \lambda_1 \\
p_{+2} &= \lambda_1 + (1 - \lambda_1)\lambda_2 \\
p_{+3} &= \lambda_1 + (1 - \lambda_1)\lambda_2 + (1 - \lambda_1)(1 - \lambda_2)\lambda_3
\end{aligned}
$$

Thus, in general,

$$p_{+j} = \lambda_1 + \sum_{k=1}^{j-1} \Big[\prod_{i=1}^{k} (1-\lambda_i) \Big] \lambda_{k+1} \tag{11}$$

z_{+j} represents the number of subjects at risk, p_{+j} is the prevalence of events occurring in the j^{th} year. For incomplete data, the likelihood functions corresponding to the first three time points will be:

$$L_1^{IC} = p_{+1}^{r_{+1}}(1-p_{+1})^{z_{+1}} = \lambda_1^{r_{+1}}(1-\lambda_1)^{n_{+1}-r_{+1}}$$
$$L_2^{IC} = p_{+2}^{r_{+2}}(1-p_{+2})^{z_{+2}} = [\lambda_1 + (1-\lambda_1)\lambda_2]^{r_{+2}}[(1-\lambda_1)(1-\lambda_2)]^{n_{+2}-r_{+2}}$$
$$L_3^{IC} = p_{+3}^{r_{+3}}(1-p_{+3})^{z_{+3}} = [\lambda_1 + (1-\lambda_1)\lambda_2 + (1-\lambda_1)(1-\lambda_2)\lambda_3]^{r_{+3}}[(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)]^{n_{+3}-r_{+3}}$$

The generalized form of the likelihood function at each time point can be written as:

$$L_j^{IC} = \Big[\lambda_1 + \sum_{k=1}^{j-1}\prod_{i=1}^{k}(1-\lambda_i)\lambda_{k+1}\Big]^{r_{+j}} \Big[\prod_{k=1}^{j}(1-\lambda_j)\Big]^{z_{+j}}, j = 1, 2, \ldots, M \tag{12}$$

For complete data, assuming the number of events to follow a multinomial distribution, the likelihood functions of the first three time points can be presented as:

$$L_1^C = \lambda_1^{r_{11}}(1-\lambda_1)^{z_{+1}}$$
$$L_2^C = \lambda_1^{r_{12}}[(1-\lambda_1)\lambda_2]^{r_{22}}[(1-\lambda_1)(1-\lambda_2)]^{z_{+2}}$$
$$L_3^C = \lambda_1^{r_{13}}[(1-\lambda_1)\lambda_2]^{r_{23}}[(1-\lambda_1)(1-\lambda_2)\lambda_3]^{r_{33}}[(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)]^{z_{+3}}$$

The generalized form of the likelihood function at each time point can be written as:

$$L_j^C = \lambda_1^{r_{1j}}[(1-\lambda_1)\lambda_2]^{r_{2j}} \ldots [\prod_{k=1}^{j-1}(1-\lambda_k)\lambda_j]^{r_{jj}}[\prod_{k=1}^{j}(1-\lambda_k)]^{z_{+j}} \tag{13}$$

Now assuming the intensity rates to be λ_i and λ_i^*, $i = 1, 2, \ldots, M$, for the AR and NR groups respectively, one can obtain the complete likelihood functions, for details see Qian, *et al.* (2023). Qian, *et al.* (2023) used logit link, see Agresti (2013), Rai and Matthews (1997), to establish a relationship between and to provide for a parsimonious modeling of the data. Specifically, they assumed,

$$\lambda_k^* = \frac{e^\beta \lambda_k}{1 + (e^\beta - 1)\lambda_k}, \text{ and } (1-\lambda_k^*) = \frac{1-\lambda_k}{1+(e^\beta-1)\lambda_k} \tag{14}$$

It may be noted that within this framework other covariates of interest could be modeled by replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta'X}$ in the above equation. Now it is easy to see that r_{+j} is known in the incomplete data and follows a binomial distribution,

$$r_{+j} \sim B(n_{+j}, 1 - \prod_{k=1}^{j}(1-\lambda_k)), \;\; j = 1, 2, \ldots, M \tag{15}$$

and the conditional distribution of r_{kj} given r_{+j} will also follow a binomial distribution given

by,

$$r_{kj}|r_{+j} \sim B(r_{+j}, \frac{\prod_{i=1}^{k-1}(1-\lambda_i)\lambda_k}{1-\prod_{i=1}^{j}(1-\lambda_i)}), \;\; k = 1, 2, \ldots, j; j = 1, 2, \ldots, M \tag{16}$$

Qian, *et al.* (2023) use the fact the sufficient statistics corresponding to the complete log-likelihood $l^C(\lambda, \beta)$ are:

$$\sum_{j=k}^{M}(r_{kj} + r^*_{kj}), \;\; \text{and} \;\; \sum_{j=k}^{M} r^*_{kj} \tag{17}$$

and propose to estimate the parameters using EM algorithm. The basic steps are outlined below:

E-step: Start with the initial estimates of $\lambda = \lambda^{(0)}$ and $\beta = \beta^{(0)}$ then one can obtain the values of $r^{(1)}_{kj}$ for $k = 1, 2, \ldots, j$ and for all $j = 1, 2, \ldots, M$.

M-step: Then, using the $r^{(1)}_{kj}$'s and the initial value $\lambda^{(0)}$ and one can obtain the estimate of β, $\beta^{(1)}$, using profile likelihood. Then, using $r^{(1)}_{kj}$'s and $\beta^{(1)}$ obtain updated estimate $\lambda^{(1)}$ using the complete likelihood. This iterative process continues until the distance between ($\lambda_1^{(q)}$, $\lambda_2^{(q)}$, $\lambda_3^{(q)}$, $\beta^{(q)}$) and ($\lambda_1^{(q-1)}$, $\lambda_2^{(q-1)}$, $\lambda_3^{(q-1)}$, $\beta^{(q-1)}$) at the q^{th} iteration is smaller than a pre-specified constant C_0. Qian, *et al.* (2023) also performed simulation studies to show that the performance of the EM approach is reasonable.

4. Extension of multi-state model for competing events

So far, our focus has been on estimating the CI for CAD. However, there are multiple events of interest such as RVSD and PH making the modeling becomes even more complicated. In this section we outline some of the issues in modeling such data and discuss some analytical approaches. For simplicity let us first assume that all three types of CEs are mutually exclusive and if the interest is in estimating CI for all three types of events, then we can proceed to model it according to the illness-death model proposed below.

In the above setting we assume that the three CDs of interest are mutually exclusive, and each survivor can have only one event during the follow-up time. In such situations, one can extend the parametric models proposed in Section 2 using either exponential or Weibull distribution. Although, in principle, the approach could be easily implemented but the likelihood representation may be somewhat complicated, and the estimations process could be computationally more involved. The pseudo survival functions can be obtained as,

$$Q_{1a}(t) = \exp\{-\int_0^t \lambda_{1a}(\nu)d\nu\}, Q_{1b}(t) = \exp\{-\int_0^t \lambda_{1b}(\nu)d\nu\}, \;\; \text{and}$$
$$Q_{1c}(t) = \exp\{-\int_0^t \lambda_{1c}(\nu)d\nu\}, Q_2(t) = \exp\{-\int_0^t \lambda_2(\nu)d\nu\}$$

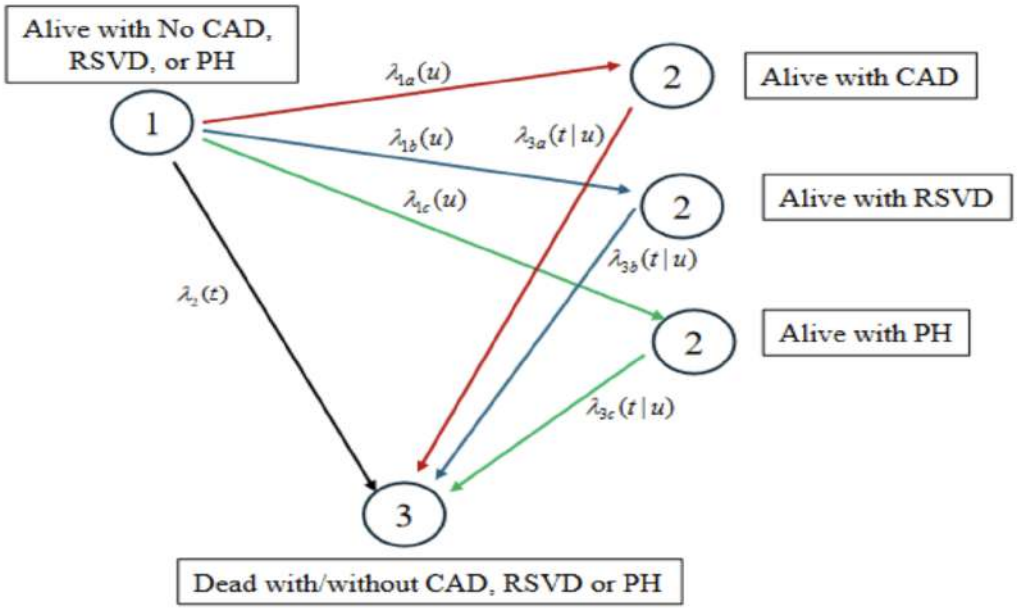


Figure 2: Three-states illness/death model of patients developing cardiovascular dysfunction

and

$$Q_{3a}(t|u) = \exp\{-\int_u^t \lambda_{3a}(\nu|u)d\nu\}, Q_{3b}(t|u) = \exp\{-\int_u^t \lambda_{3b}(\nu|u)d\nu\},$$
$$Q_{3c}(t|u) = \exp\{-\int_u^t \lambda_{3c}(\nu|u)d\nu\}$$

It may be noted that probability of surviving without experiencing and any CD or death beyond time t can be represented as,

$$Q(t) = \exp\{-\int_0^t (\lambda_{1z}(\nu) + \lambda_{1b}(\nu) + \lambda_{1c}(\nu) + \lambda_2(\nu))d\nu\} = Q_{1a}(t)Q_{1b}(t)Q_{1c}(t)Q_2(t) \quad (18)$$

and the survival function can be obtained as,

$$\begin{aligned} S(t) &= P(X(t) = 1) + P(X(t) = 2) \\ &= Q(t) + \int_0^t \lambda_{1a}(u)Q(u)Q_{3a}(t|u)du + \int_0^t \lambda_{1b}(u)Q(u)Q_{3b}(t|u)du + \\ &\int_0^t \lambda_{1c}(u)Q(u)Q_{3c}(t|u)du \end{aligned} \quad (19)$$

Then, using the above pseudo survival functions one can obtain the contributions to the likelihood made by each survivor depending on their outcome which can be described as follows:

(i) Death without any CD, $T = t$, $X(t^-) = 1$, and $L_1(t) = \lambda_2(t)Q(t)$,

(ii) Alive without any CD, $T > t$, $X(t) = 1$, and $L_2(t) = Q(t)$,

(iii) Death with CAD, $T = t$, $X(t^-) = 2$, and $L_{3a} = \int_0^t \lambda_{1a}(u)Q(u)\lambda_{3a}(t|u)Q_{3a}(t|u)du$

(iv) Alive with CAD, $T > t$, $X(t) = 2$, and $L_{4a} = \int_0^t \lambda_{1a}(u)Q(u)\lambda_{3a}(t|u)du$

(v) Death with RSVD, $T = t$, $X(t^-) = 2$, and $L_{3b} = \int_0^t \lambda_{1b}(u)Q(u)\lambda_{3b}(t|u)Q_{3b}(t|u)du$

(vi) Alive with RSVD, $T = t$, $X(t) = 2$, and $L_{4b}(t) = \int_0^t \lambda_{1b}(u)Q(u)Q_{3b}(t|u)du$

(vii) Death with PH, $T > t$, $X(t^-) = 2$, and $L_{3c}(t) = \int_0^t \lambda_{1c}(u)Q(u)\lambda_{3c}(t|u)Q_{3c}(t|u)du$

(viii) Alive with PH, $T = t$, $X(t) = 2$, and $L_{4c}(t) = \int_0^t \lambda_{1c}(u)Q(u)Q_{3c}(t|u)du$

Based on the above quantities one can write down the likelihood function and obtain the estimates of the parameters of interest using the theory of maximum likelihood or EM algorithms as discussed in Sections 2 and 3. Extension of semi-parametric approach for simultaneous modeling of the three types of CDs requires more theoretical development and is proposed as future work.

5. Conclusions and discussions

In this manuscript, we have provided an overview of the parametric and semi-parametric approaches that could be adopted for modeling CI of one or more competing events of interest with death being an absorbing state.

When the survivors are followed longitudinally then, under the assumption of continuous time Markov process, one can easily adopt the likelihood approach to model the transition probabilities as discussed in van den Hout (2017, Chapter 4). The development of semi-parametric approach needs to be developed and is left as future work.

The SJLIFE cohort study is a unique study to evaluate the association of childhood cancer treatment with the long-term adverse effect. The discussed approach can be extended to any interval-censored data or any multi-state models and could be extremely useful in the prediction of adverse outcomes.

There are multiple challenges to drawing statistical inference from such studies. Robustness of results may depend on the selection of the cohort and time of data collection as discussed in Table 2. Estimation can be based on parametric, non-parametric or semi-parametric models. The number of parameters not only depends on estimation procedure but also number of stages as included in Figures 1 and 2. Incorporating covariate effects on different parameters in Figures 1 and 2 makes inference much more cumbersome.

Funding

S.N. Rai was partly supported by the University of Cincinnati Cancer Center, College of Medicine. D.K. Srivastava, K. Ness and M. Hudson were in part supported by the St. Jude Children's Research Hospital Cancer Center Support Grant No. 5P30CA021765-33, the St. Jude Lifetime Cohort Study Grant No. U01 CA195547, and the American Lebanese Syrian Associated Charities (ALSAC).

Acknowledgements

We thank Dr. Vinod Gupta, Chair Editor, Statistics and Applications President, Society of Statistics, Computer and Applications (SSCA); ssca.org.in/ and President, Governing Body, Institute of Applied Statistics and Development Studies.

References

Agresti, A. (2013) *Categorical Data Analysis*. Wiley-Interscience, A John Wiley & Sons, INC., New York.

Armstrong, G. T., Liu, Q., Yasui, Y., Neglia, J. P., Leisenring, W. L., *et al.* (2009). Late mortality among 5-year survivors of childhood cancer: a summary from the childhood cancer survivor study. *Journal of Clinical Oncology*, **27**, 2328-2338.

Bhakta, N., Liu, Q., Ness, K. K., Bassiri, M., Eissa, H., *et al.* (2017). The cumulative burden of surviving childhood cancer: An initial report from the St. Jude Lifetime Cohort Study. *LANCET*, **390**, 2569-2582.

Ehrhardt, M. J., Krull, K. R., Bhakta, N., Liu Q., Yasui Y., *et al.* (2023). Improving quality and quantity of life for childhood cancer survivors globally in the twenty-first century. *Nature Reviews Clinical Oncology*, **20**, 678-698.

Eulenburg, C., Manher, S., Woeiber, L., and Wegscheider, K. (2015). A systematic model specification procedure for an illness-death model without recovery. *PLoS One*, **10**, e0123489.

Gunaratnam, B. and Rai, S. N. (2019). Comparing the variability using Louis' method and resampling methods. *Journal of Biometrics and Biostatistics*, **10**.

Hammoud, R. A., Liu, Q., Dixon, S. B., Onerup, A., Mulrooney, D. A., *et al.* (2024). The burden of cardiovascular disease and risk for subsequent major adverse cardiovascular events in survivors of childhood cancer: a prospective, longitudinal analysis from the St. Jude Lifetime Cohort Study. *The Lancet Oncology*, **25**, 811-822.

Harezlak, J., Gao, S., and Hui, S. L. (2003). An illness-death stochastic model in the analysis of longitudinal dementia data. *Statistics in Medicine*, **22**, 1465-1475.

Howell, C.R., Bjornard, K. L., Ness, K. K., Alberts, N., Armstrong, G. T., *et al.* (2021). Cohort Profile: The St. Jude Lifetime Cohort Study (SJLIFE) for pediatric cancer survivors. *International Journal of Epidemiology*, **50**, 39-49.

Hudson, M. M., Ehrhardt, M. J., Bhakta, N., Baassiri, M., Eissa, H., *et al.* (2017). Approach for Classification and Severity Grading of Long-term and Late-Onset Health Events among Childhood Cancer Survivors in the St. Jude Lifetime Cohort. *Cancer Epidemiology, Biomarkers & Prevention*, May; **26**, 666-674. doi: 10.1158/1055-9965.EPI-16-0812. EPI-16-0812. Epub 2016 Dec 29. PMID: 28035022; PMCID: PMC5413397.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association, Theory and Methods*, **80**, 863-871.

Koutsoukis, A., Ntalianis, A., Repasos, E., Kastritis, E., Dimopoulos, M. A., *et al.* (2018). Cardio-oncology: A Focus on Cardiotoxicity. *European Cardiology Review*, **13**, 64-69.

Murphy, S. A. and Van der Vaart, A. W. (2000). On Profile Likelihood. *Journal of the American Statistical Association*, **95**, 449–465.

Park, J.R., Bagatell, R., London, W.B., Maris, J.M., Cohn, S.L., *et al.* (2012). Children's Oncology Group's 2013 blueprint for research: neuroblastoma. *Pediatric Blood Cancer*, **60**, 985-993.

Pradhan, B. and Kundu, D. (2014). Analysis of Interval-Censored Data with Weibull Lifetime Distribution. *Sankhya B*, **76**, 120–139.

Qian, C., Srivastava, D. K., Pan, J., Hudson, M. M., and Rai, S. N. (2023). Estimating transition intensity rate on interval-censored data using semi-parametric with EM algorithm approach. *Communications in Statistics - Theory and Methods* **53**, 1-17.

Rai, S. N. and Matthews, D. E. (1993). Improving the EM algorithm. *Biometrics*, **49**, 587-591.

Rai, S. N. and Matthews, D. E. (1997). Discrete scale models for survival-sacrifice experiments. *Applied Statistics*, **48**, 93-109.

Rai, S. N., Matthews, D. E., and Krewski, D. R. (2000). Mixed-scale models for survival/sacrifice experiments. *The Canadian Journal of Statistics*, **28**, 65-80.

Rai, S. N. (2008). Analysis of occult tumor studies. In: Tan WY and Hanin L (Ed.), *Handbook of Cancer Models with applications.* World Scientific Press.

Rai, S. N., Pan, J., Sun, J., Hudson, M. M., and Srivastava, D. K. (2013). Estimating incidence rate on current status data with application to a Phase IV cancer trial. *Communications in Statistics: Theory and Methods*, **42**, 2417-2433.

Robison, L. L., Armstrong, G. T., Boice, J. D., Chow E. J., Davies, S. M., *et al.* (2009). The childhood cancer survivor study: a National Cancer Institute – supported resource for outcome and intervention research. *Journal of Clinical Oncology*, **27**, 2308-2318.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER *Stat Database: Incidence – SEER Research Data, 8 Registries, Nov 2021 Sub (1975-2019) – Linked To County Attributes – Time Dependent (1990-2019) Income/Rurality 1975-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 submission.

Srivastava, D. K., Zhu, L., Hudson, M. M., Pan, J., and Rai, S. N. (2018). Robust estimation and inference on current status data with applications to Phase IV cancer trial. *Journal of Modern Applied Statistical Methods*, **17**, Article no. 18, 1-20.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer: New York.

van den Hout A. (2017). *Multi-State Survival Models For Interval Censored Data.* CRC Press, Taylor and Francis Group, New York.

Winther, J. F., Kenborg, L., Byrne, J., Hjorth, L, Kaatsch, P., *et al.* (2015). Childhood cancer survivor cohorts in Europe. *Acta Oncologica*, **54**, 655-668.

Zhang, X., Zhang, Y., Ye, X., Guo, X., Zhang, T., *et al.*. (2016). Overview of phase IV clinical trials for postmarket drug safety surveillance: a status report from the ClinicalTrials.gov registry. *British Medical Journal (open)*, **6**, e010643.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 127-141

Text Representation: A Journey from Traditional Vector Space Model to LLM

Sharad Verma[1], Pragati Bhatnagar[2] and Aditi Sharan[3]
[1]*Department of Information Technology*
Rajkiya Engineering College Ambedkar Nagar,224122
[2]*Jain Vishwabharti University, ladnun, Rajasthan*
[3]*SC&SS, Jawaharlal Nehru University, New Delhi, 110067*

Received: 01 July 2024; Revised: 15 July 2024; Accepted: 17 July 2024

Abstract

In the era of language processing and artificial intelligence, people are amazed by the capabilities of ChatGPT. However, ChatGPT is not the end but the beginning of an era where much more is yet to happen. The backbone of ChatGPT is generative AI or large language models (LLMs). One of the most difficult challenges has been dealing with the semantics or meaning of language. This is what LLMs have been able to achieve to a certain extent: enabling computers to understand the semantics of text. However, LLMs are not the effort or product of a single person or community. They are the result of ongoing community efforts involving numerous scientists, researchers, and professionals who have worked for decades across the globe in various interdisciplinary fields, including statistics, computer science, and linguistics. Therefore, it is particularly important for the computer science and statistics communities to systematically understand the evolution of language models. This is the main objective of our paper. Our paper addresses the fundamental issue of incorporating the semantics of natural language text into its representation, which is the core of language models, including LLMs, and has revolutionized the field of natural language processing (NLP).

Key words: Attention based neural network; BERT; Deep learning based language embeddings; Large Language model; LSTM.

1. Introduction

Since the start of the digital processing of natural language text, text representation has been the most primitive requirement but, at the same time, the most complicated task. Some of the parameters are well known in the context of the difficulty in text representation viz: unstructured nature, ambiguity, how to represent meaning of text, high dimensionality, etc. However, another major issue evolved with the use of machine learning techniques for processing natural language text. As all the Machine learning algorithms work naturally

Corresponding Author: Dr. Aditi Sharan
Email: aditisharan@mail.jnu.ac.in

with the vector data, the vector representation of the text is the first step in dealing with the text data. Not only is it the first step, but it also impacts the text mining process and their outcomes from the machine learning algorithm, accuracy and the biases in the results. Though natural language processing has been an important research area for more than a decade, today, it has encroached on our day-to-day life, intentionally or unintentionally, in the form of web searches, recommendation systems, chatbots, etc. People are amazed by the potential of ChatGPT. However, along with the utility of ChatGPT in NLP, many issues are emerging regarding the accuracy, reliability, authenticity, and biases of the results. To be aware of these intricacies, one must know what is happening at the backend of ChatGPT as a specific case. Another term associated with the emergence of ChatGPT that became popular in the scientific and general community is Large Language Models(LLMs), more popularly termed generative AI models. These models form the core essence of the working of ChatGPT. Thus, it becomes important for the statistical and computer science community to understand the notion of LLMs. The core of LLM's efficiency is the efficient text representation and text generation. Here, we will focus on text representation. In fact, text generation can be considered as an outcome of text representation. The representation of text in the form we see today is not a sudden discovery, but actually, it is a success story of a long, tough but consistent effort of the NLP community, including computer scientists, mathematicians, statisticians and a lot of contributions from linguists. The objective of this paper is to present the evolution of transformer-based generative LLMs starting from one of the earliest tf-idf based traditional vector space models. The paper provides a systematic review of the important models focusing on the emergence of transformer-based models. It will address the basic issue of incorporating semantics of a natural language text in the representation itself, which is the core of deep learning algorithms, including LLMs and has revolutionized the field of NLP. Additionally, we discuss the unresolved issues and challenges in this field. We also present some of our efforts to address these challenges.

The paper is organized as follows: Section 2 introduces the traditional TF-IDF-based Vector Space Model. Section 3 delves into the Distributional Hypothesis, laying the groundwork for the subsequent discussion on machine learning techniques for text representation in Section 4. Section 5 presents deep learning-based language modeling, and Section 6 concludes the paper.

2. Traditional TF-IDF based Vector Space Model (VSM)

One of the earliest attempts to represent a text document by a vector emerged in the form of the Classical Vector Space Model (VSM)Salton *et al.* (1975). Though not very efficient for representing the text, it is one of the simplest and computationally efficient models. Also one should keep in mind that it presents one of the earliest attempts to represent a text by a vector. The model is based on the bag of words (BOW) approach. To understand the BOW-based VSM model, let us consider that we have a corpus of text documents and the objective is to represent each text document by a vector. The initial step involves preprocessing of the text, mainly involving stop word removal and stemming to remove irrelevant/nonessential words. After preprocessing, the outcome results in a collection of words that form the vocabulary of the text corpus. A text document matrix (TDM) can then be constructed with $M * N$ size, where rows represent documents, M being number of documents and columns represent terms, N being the size of the vocabulary. Further,

each element w_{ij} of TDM represents the weight of j_{th} term in i_{th} document. Thus each row of the matrix represents the vector corresponding to the document, the size of the vector being the vocabulary size. Ideally, the weight should represent the importance of a term in the document. But importance itself is a subjective term and may depend on the task to be performed. However, some objective criteria is required. Various Now weights can be assigned in different ways, however, tf-idf (term frequency-inverse document frequency) based weighting emerged as the most popular technique for traditional VSM based model.

2.1. Weighting using TF-IDF (Term Frequency-Inverse Document Frequency)

The words in VSM need to be given weight so as to reflect their importance in the document. Documents with higher weights are more important. The most obvious way of giving weights can be related to the frequency of words in the document. The weight may correspond to frequency count of words in the document. Table 2 represents the matrix using term frequency count for the same example as presented for binary vector. The raw frequency may not be a statistically stable value, so a normalized measure might be used.However, it can easily be observed that frequency of the word, though important, may not be the only measure for assigning the weight to a word. The simplest example to understand this is that stopwords are most frequent but are least important. This leads to the notion of TF-IDF based measure in VSM. It is a statistical measure that reflects how important a word is to a document in a collection or corpus. TF-IDF consists of two main components: Term frequency (TF) and Inverse document frequency (IDF). Term frequency (TF) measures the number of times a term (word) is present in a document. Instead of raw value , the value may be normalized. Inverse document frequency (IDF) measures how rare a term is across the corpus. Some variants of IDF exist, as far as it is inversely proportional to the no. of documents in which it is appearing, thus giving higher weightage to rare terms. One way of calculating IDF is as follows: For each term t in the corpus:

$$IDF(t) = log(N/(df(t))) \tag{1}$$

where N: Total number of documents in the corpus. $df(t)$: Number of documents in the corpus that contain the term t.

TF is a local measure (calculated for each document), whereas IDF is a global measure (calculated for the entire corpus). TF-IDF is then calculated as the product of TF and IDF. The resulting value represents the importance of the term in the document and the corpus as a whole. High TF-IDF values indicate that a term is important to a document and the corpus, while low values indicate that the term is less important or common. Also frequently stated as frequent and rare terms are more important. TF-IDF is often used for text classification, information retrieval, and content-based recommendation systems. It is a popular technique because it is simple, efficient, and effective in identifying important terms in a document or corpus. The TF-IDF score can be calculated by the formula:

$$TF - IDF(t, d) = TF(t, d) * IDF(t) \tag{2}$$

where $TF(t, d)$ represents the frequency of term t in document d.

Let us take an example of three Documents for a better understanding:-

Table 1: Sentences from medical domain

Document	Sentence
Document 1	Chemotherapy is used to treat cancer.
Document 2	Cancer cells can grow uncontrollably.
Document 3	Radiation therapy damages cancer cell

After some preprocessing (stopword removal and case conversion), the vocabulary of the corpus in this case may be represented as [cancer, therapy, cells, chemotherapy, radiation, grow, uncontrollably, damages]

Table 2 contains TF-IDF Scores of each term in the document.

Table 2: TF-IDF Scores

Term	Document 1	Document 2	Document 3
cancer	0	0	0
therapy	0.4055	0	0.4055
cells	0	0.4055	0.4055
chemotherapy	1.0986	0	0
radiation	0	0	1.0986
grow	0	1.0986	0
uncontrollably	0	1.0986	0
damages	0	0	1.0986

It can be observed that medical domain terms present in the document are important.

Traditional VSM is a sort of one-hot encoding where each word has its own space and index. One hot representation has a lot of limitations and problems. Some of the important problems with one hot encoding are as follows :

- Viewing all the words as discrete units is not ideal.
- High dimensionality problem, since there can be hundreds of thousands of words in a given language, representing and storing words as one-hots can be extremely expensive.
- Sparsity Problem.
- All sequencing information is lost.
- Lack of an inherent similarity notion. a simple way of measuring the similarity between two vectors is using the cosine similarity. But since the one-hot vectors of any two different words are necessarily orthogonal, taking the dot product of even two synonyms would yield a similarity score of 0.
- Can not deal with contextual similarity between sentences.

The Distributional Hypothesis addresses some of these limitations by positing that words with similar distributions in a large corpus are likely to have similar meanings. By

leveraging statistical patterns in language usage, the Distributional Hypothesis allows VSM to capture semantic similarities more effectively across varied contexts and improve the representation of word meanings beyond mere co-occurrence.

3. Distributional hypothesis

Now there has been much talk about incorporating semantics in text representation. This tends to consider meaning of words and the notion of synonyms in the text representation. It brings into the picture the notion of dictionary, thesaurus, ontology, *etc* . All these resources are beneficial for understanding the meaning of the text but fail to provide a computational model for representing the meaning of the text. Thus came the idea of the Distributional hypothesis, the same idea repeated in different ways by various researchers Harris (1954). A very old and popular idea in the Linguistic domain – You shall know a word by the company it keeps Firth (1957). In particular in the modern NLP context – A word is defined by its environment (the context words around it). But this is again based on the sound foundation of linguists Harris (1954): If A and B have almost identical environments we say that they are synonyms. For a long time, computational linguists have been focusing on the representation of the context of a word that can assist in incorporating the semantics of the word in the representation itself. The base of this again lies in the hypothesis that words with high similarity (such as synonyms) occur in the same context.

3.1. Representing words using co-occurrence statistics

In the previous section, we saw an example of representing text as a vector but in many applications, we are interested in finding an appropriate representation of a word as a vector. One of the most primitive ways is to consider each column of the TF-IDF matrix as a word vector. Thus words can be represented as vectors in document dimension. Here two word vectors are similar if they share common documents. This word vector captures the information of words based on their presence in documents that are not very meaningful. There are no word-to-word associations. Based on the distributional hypothesis it was thought to represent word vectors in a way that might capture this association. One of the earliest attempt is reflected in the form of mutual information-based association. Mutual information (MI) is a measure of how often two events x and y occur, compared with what we would expect if they were independent. The standard formula is

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3}$$

Based on MI an important measure in NLP is Pointwize MI (PMI). PMI measures the the chances of co-occurrence of two words PMI between two words (we we may call them word w and context word c) w and c can be given by

$$PMI(w,c) = log_2 \frac{P(w,c)}{P(w)P(c)} \tag{4}$$

However, PMI value may range from positive to negative infinity. Negative informa-

tion may create misinformation in many cases. Thus a refined measure is called as Positive PMI(PPMI)

$$PPMI(w,c) = max(log_2 \frac{P(w,c)}{P(w)P(c)}, 0) \tag{5}$$

Let us try to understand its practical application. Let us consider some words and selective context words. The words considered are computer, data, result, pie, and sugar. The context words are cherry, strawberry, digital, and information. We can see the co-occurrence of the context words in a large corpus of the order of say Wikipedia. For determining the co-occurrence of context words, we need to fix the window size say 4, then observe the count of context words in the neighborhood of all occurrences of a word considering 4 neighbors on the left and 4 on the right. Assume that it results in the following matrix between word and context.

Table 3: Co-occurence statistics

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	344
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

The resultant PPMI comes out to be :

Table 4: PPMI scores

computer	data	result	pie	sugar	count(w)
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Now a word is represented as a vector in word dimension. This was a small example. If all words in the vocabulary are included. We can have a $n * n$ matrix where n is size of the vocabulary. Thus each word is a vector in the dimension of vocabulary. One can make out an interesting observation. One can find the cosine similarity between the words and you can observe from the table that cherry and strawberry are more similar and similarly digital and information. For other pairs, the similarity is zero. The obvious observation is that digital and information share computer and data, while cherry and strawberry share pie and sugar. Thus PPMI based captures the co-occurrence information in the word vector representation that is quite meaningful. However, this matrix is again sparse as in the case of TF-IDF based measure. Also, it does not involve any learning. With the advancement in machine learning approaches for NLP, now machine learning is being used for text representation,

text processing, text mining, and finally text generation. In the next part, our focus is on machine learning based vector representation of text, frequently called word embedding and text embedding.

4. Machine Learning for text representation

The previous section presented Count based context vector creation. Count based context representation does not involve any learning so has a limited usage in era of machine learning , where everything can be learnt and thus can be predicted. Prediction based Context word model can predict the co-occurrence probabilities , thus they can predict the context words corresponding to a given word. While we are learning to predict context words, the sole objective is not only context word prediction. In fact such a trained model leads to the notion of representational learning, where the vectors representing the words can be learnt through neural network based models.

4.1. Word2Vec

Word2Vec is a groundbreaking model introduced by Tomas Mikolov Mikolov *et al.* (2013a,b) and his team at Google in 2013. It revolutionized the field of natural language processing by enabling the creation of dense vector representations of words in a continuous vector space. This model uses a shallow, two-layer neural network to process vast amounts of text data and learn the relationships between words based on their context. There are two main architectures used in Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a target word from its surrounding context words, while Skip-gram does the reverse, predicting the context words given a target word. These approaches allow Word2Vec to capture semantic relationships between words effectively, such that words with similar meanings are positioned close to each other in the vector space.

One of the most compelling features of Word2Vec is its ability to capture linear relationships between words. For example, the model can understand analogies like “King” is to “queen” as “man” is to “woman” by performing vector arithmetic: vector(“king”) - vector(“man”) + vector(“woman”) $\approx$ vector(“queen”). This ability stems from the model’s training process, which uses a technique called negative sampling to efficiently differentiate relevant context words from irrelevant ones. The learned vectors from Word2Vec have been widely used in various applications, including machine translation, text classification, and sentiment analysis, due to their effectiveness in encoding semantic meanings and relationships in a computationally efficient manner.

4.2. GloVe

Global Vectors for Word Representation(GloVe) Pennington *et al.* (2014) is a powerful word embedding technique developed by researchers at Stanford University. Unlike traditional methods that solely rely on local context (like Word2Vec), GloVe combines the benefits of both global matrix factorization and local context window methods. It constructs word vectors by aggregating global word-word co-occurrence statistics from a large corpus. Specifically, it utilizes a co-occurrence matrix where the entries represent the frequency of word pairs appearing together within a certain window. By factorizing this matrix, GloVe generates dense vector representations where the semantic relationships between words are

captured. For instance, the difference between the vectors for "king" and "queen" is similar to the difference between "man" and "woman", reflecting meaningful linear substructures. These pre-trained embeddings have been extensively used in various natural language processing tasks due to their ability to capture both syntactic and semantic word relationships effectively.

Since this method captures the global statistics in the corpus, it is named Global Vectors (in short GloVe). These methods have a very good performance on various tasks like word analogy, word-similarity, and named entity recognition.

5. Deep Learning based Language modeling

In addition to many problems associated with the computational processing of natural language, another major issue is the temporal nature of language that is reflected in language flow, continuity, etc. Thus a language is sometimes considered as a sequence that unfolds in time. Most of the deep learning models perform the task of language modeling. At the most primitive level, language modelling involves the prediction of the next word in the sequence. How accurately the model predicts the word forms the base for the performance of the language model. It seems a well-defined task, but it is too complicated, as the correct prediction of words can only be done if the text is understood properly. However, there is no model for the representation of meaning. This actually leads to modeling the distributional hypothesis, the well-known hypothesis given by linguists. These models try to capture/learn the context vectors of the words. Based on the context vector, the model tries to predict the next word in sequence. But the question may come to mind: What would be so great if we were able to predict the next word? The answer lies in the statement that correct prediction is not possible without an understanding of the text(for which there is no explicit mechanism). Thus, correct prediction involves some understanding of the text. In other words, deep learning models are able to capture the semantics of text using the notion of statistics and probabilities. With time, various deep learning models have evolved, but in terms of architecture, we have three categories: Sequential model, encoder decoder-based architecture, and transformer-based model.

5.1. Sequential Models: The Era of RNN and LSTM

The era of sequential models in artificial intelligence has been significantly shaped by recurrent neural networks (RNNs) Elman (1990) and their refined variant, Long Short-Term Memory networks (LSTMs). These models excel in processing sequential data, such as time series, text, and speech, by maintaining an internal state that evolves as new inputs are processed. RNNs were among the first neural architectures capable of capturing temporal dependencies, making them pivotal in tasks like speech recognition, language modeling, and machine translation. However, traditional RNNs are prone to the vanishing and exploding gradient problems, hindering their ability to learn long-term dependencies effectively. The introduction of LSTMs by Hochreiter and Schmidhuber in 1997 addressed these challenges by incorporating memory cells and gating mechanisms that regulate information flow, allowing them to remember information over long sequences and selectively forget irrelevant details. This innovation marked a breakthrough in sequential modeling, enabling more robust learning and improved performance in tasks requiring nuanced understanding of context and continuity over time. Despite their successes, both RNNs and LSTMs have limitations

in handling very long sequences due to computational constraints and struggles with capturing hierarchical dependencies. As the field advances, newer architectures like transformers, which rely on attention mechanisms, have emerged to address these shortcomings and push the boundaries of sequential modeling in modern AI applications.

5.2. Limitation and constraints of LSTM

Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber (1997)networks are a type of recurrent neural network (RNN) specifically designed to overcome the limitations of traditional RNNs, such as the vanishing gradient problem, by introducing a more sophisticated memory architecture. The LSTM unit consists of a cell state C_t and three gates that regulate the flow of information: the input gate i_t, the forget gate f_t, and the output gate o_t.The equations governing these gates are as follows:

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ C_t^h &= \tanh(W_C[h_{t-1}, x_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * C_t^h \\ h_t &= o_t * tanh(C_t) \end{aligned} \tag{6}$$

Here, x_t represents the input at time step t, h_t is the hidden state, W and b denote the weights and biases for the respective gates, σ is the sigmoid activation function, and $*$ indicates element-wise multiplication. The forget gate f_t determines which information from the previous cell state C_{t-1} should be discarded, the input gate i_t decides which new information should be added to the cell state, and the output gate o_t controls what part of the cell state is output as the hidden state h_t. The combination of these gates allows LSTMs to maintain and update the cell state over long sequences, making them highly effective for tasks that require learning from temporal patterns, such as natural language processing, speech recognition, and time-series prediction. Long Short-term Memory (LSTM) networks, despite their advancements in handling sequential data, have several limitations. Firstly, they require substantial computational resources and time for training due to their complex architecture and numerous parameters, which can be a bottleneck for large datasets. Additionally, LSTMs can struggle with very long sequences, where even their memory cells might not effectively capture dependencies over extremely long periods, potentially leading to gradient vanishing or explosion issues. Moreover, fine-tuning LSTM models requires considerable expertise, as the process involves balancing many hyperparameters, such as learning rates and the number of layers. They also tend to overfit if not regularized properly. Lastly, LSTMs are less interpretable compared to simpler models, making it challenging to understand and trust their decision-making processes, which is crucial in applications where model transparency is essential.

Vanishing and exploding gradients Bengio *et al.* (1994) pose significant challenges in training RNNs, particularly due to their deep sequential nature. In standard RNNs, the vanishing gradient problem arises because gradients can diminish exponentially over time steps, especially in long sequences, making it difficult for the model to learn dependencies

over distant time steps. Conversely, exploding gradients can occur due to unstable weight updates, leading to numerical instability during training. Both problems can severely impact the ability of neural networks to learn and generalize from data effectively. Addressing these issues often involves careful initialization of parameters, using activation functions that mitigate gradient saturation, and employing techniques like gradient clipping to stabilize training dynamics.

While LSTM networks have been instrumental in addressing the vanishing gradient problem in RNNs, they are not without their shortcomings. One notable limitation is their computational complexity, which arises from the multiple gates and memory cells that need to be managed per unit. This complexity can lead to slower training times and increased memory requirements, making LSTMs less scalable for very large datasets or when deploying models in resource-constrained environments. Additionally, LSTMs may struggle with capturing fine-grained dependencies within sequences, as their gating mechanisms are designed to capture long-term dependencies rather than focusing on specific, short-term relationships.

The emergence of attention mechanisms represents a significant advancement in addressing these shortcomings. Attention networks, such as the Transformer model, allow neural networks to dynamically focus on different parts of the input sequence. Unlike LSTMs, attention mechanisms do not impose a fixed-length context window and can adaptively attend to relevant parts of the input sequence. This flexibility enables attention networks to capture both short-term and long-term dependencies more effectively without the computational overhead of managing complex gating mechanisms. Moreover, attention mechanisms have shown superior performance in tasks like machine translation, where aligning and translating words or phrases across languages require capturing nuanced relationships within and between sentences.

5.3. Attention

An attention network, often referred to simply as an attention mechanism is a component of neural networks designed to dynamically focus on specific parts of the input data when making predictions. Attention networks were developed to address some of the limitations of LSTMs, particularly their difficulty in capturing long-range dependencies and their inefficiency with very long sequences. The attention mechanism allows the model to weigh the importance of different input elements, enabling it to handle long-range dependencies and improve performance on tasks such as machine translation, image captioning, and more.

The primary idea of attention is to assign different weights to different parts of the input sequence. When making a prediction, the model can then focus more on the relevant parts and less on the irrelevant ones. This selective focus helps capture relationships and dependencies that span long distances in the input data.

Several types of attention mechanisms are designed to address specific needs or improve computational efficiency. Here's an overview of the most commonly used attention mechanisms:

Additive (Bahdanau) Attention: Bahdanau attentionBahdanau *et al.* (2014), also known as additive attention, was introduced by Dzmitry Bahdanau and colleagues in their 2014 paper to improve the performance of sequence-to-sequence (seq2seq) models, particu-

larly for machine translation. The main goal of Bahdanau's attention is to allow the model to focus on different parts of the input sequence dynamically when generating each part of the output sequence. It combines the decoder hidden state and the encoder hidden states using a trainable weight matrix and a non-linear activation function (typically tanh).

$$e_{t,i} = v^T \tanh(W_s s_t + W_h h_i) \tag{7}$$

where W_s and W_h are weight matrices, v is a weight vector, s_t is the decoder hidden state at time t, and h_i is the encoder hidden state at time i.

Pros: Flexible and can capture complex relationships between the encoder and decoder states.

Cons: Computationally expensive due to the non-linear transformation.

Multiplicative (Luong) Attention: Proposed by Luong et al.Luong *et al.* (2015), this mechanism computes attention scores using a dot product (multiplicative) approach, which is computationally more efficient than additive attention.

$$e_{t,i} = s_t^T h_i \tag{8}$$

Pros: More efficient than additive attention.

Cons: May not perform as well as additive attention when the dimensions of s_i and h_i differ significantly.

Self Attention: Self-attention is a mechanism used in various neural network architectures, particularly in transformers Vaswani *et al.* (2017), to enable models to focus on different parts of the input sequence when processing each token. This mechanism allows the model to capture dependencies regardless of their distance in the sequence, making it highly effective for tasks like language modeling and machine translation.

In self-attention, each token in the input sequence is transformed into three vectors: Query (Q), Key (K), and Value (V), as shown in fig 1. These vectors are derived through learned linear transformations. Assume we have an input sequence of length n, represented by the matrix $X \epsilon R^{n*d}$, where d is the dimensionality of the input embeddings. The input matrix X is multiplied by three weight matrices to produce the Query, Key, and Value matrices as follows:

$$\begin{aligned} Q &= XW^Q, \\ K &= XW^K, \\ V &= XW^V \end{aligned} \tag{9}$$

where $W^Q, W^K, W^V \epsilon R^{d*d_k}$ are weight matrices, and d_k is the dimensionality of Query, Key, and Value vectors.

The core of the self-attention mechanism involves computing a score for each pair of tokens in the sequence to determine how much focus one token should have on another. This is done using the Query and Key matrices. The score for each pair is calculated as the dot

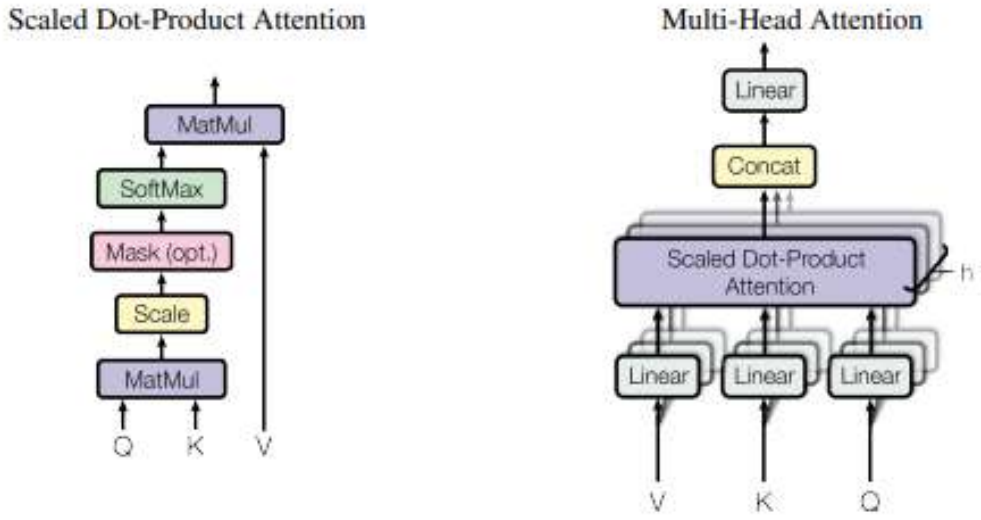


Figure 1: Scaled Dot Product and Multi-head Attention

product of their Query and Key vectors, scaled by the square root of d_k to stabilize gradients as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{10}$$

It is also known as scaled dot-product attention. The result of the self-attention mechanism is a new representation of the input sequence, where each token now includes information from all other tokens, weighted by their relevance. This process can be repeated in multiple layers to capture increasingly complex dependencies.

To allow the model to focus on different parts of the sequence simultaneously, the self-attention mechanism is often extended to multi-head attention. This involves using multiple sets of Query, Key, and Value weight matrices:

$$MultiHead(Q, K, V) = Concat(head_1, head_2..., head_h)W^O \tag{11}$$

Each head independently performs the self-attention operation, and the results are concatenated and linearly transformed using the weight matrix W^O.

The Transformer architecture has revolutionized the field of natural language processing by enabling more efficient and effective processing of sequential data, notable for its innovative encoder-decoder structure. This architecture is designed to handle sequence-to-sequence tasks such as translation, summarization, and question-answering with unparalleled efficiency and performance. The encoder is composed of multiple identical layers, each with two key sub-layers: multi-head self-attention mechanisms and position-wise feed-forward networks. The self-attention mechanism allows the encoder to weigh the importance of different words in a sentence relative to each other, capturing complex dependencies and

relationships. Each encoder layer also includes residual connections and layer normalization, ensuring stability and improving gradient flow during training. The decoder mirrors the encoder but includes an additional sub-layer for masked multi-head self-attention, which ensures that predictions for a given word depend only on previous words in the sequence, maintaining causality. The decoder also integrates multi-head attention over the encoder's output, allowing it to focus on relevant parts of the input sequence when generating each word of the output. This dual structure of the Transformer, with its powerful self-attention mechanisms and ability to process input and output sequences in parallel, represents a significant advancement over traditional recurrent models, offering superior scalability and the ability to capture long-range dependencies more effectively.

This architecture is highly scalable and can be parallelized, leading to faster training times and improved performance on large datasets. The Transformer has become the foundation for many state-of-the-art models, including BERT, GPT, and T5, driving significant advancements in tasks such as machine translation, text generation, and sentiment analysis.

5.4. BERT

Bidirectional Encoder Representations from Transformers(BERT) embeddings Devlin *et al.* (2018) have revolutionized natural language processing by providing deep, contextualized representations of words. Unlike traditional embeddings that generate a fixed vector for each word regardless of context, BERT dynamically produces word vectors based on the entire sentence, capturing intricate details of the language. BERT's pre-training involves two key subtasks: the masked language model (MLM) and next sentence prediction (NSP).

In MLM, a portion of the input tokens is randomly masked, and the model is trained to predict these masked tokens based on the surrounding context. For example, in the sentence "The quick brown fox jumps over the lazy [MASK]," BERT attempts to predict the masked word "dog" using the context provided by the rest of the sentence. This task forces BERT to develop a bidirectional understanding of language, considering both the left and right contexts of each word.

NSP is designed to train BERT to understand the relationship between sentences. During pre-training, BERT receives pairs of sentences. Some pairs are actual consecutive sentences from the corpus, while others are random pairs. The model learns to classify whether the second sentence logically follows the first. For instance, given the sentence pair *The man went to the store. He bought a gallon of milk*, BERT should identify this as a logical sequence. Conversely, for the pair *The man went to the store. Penguins are great swimmers*, BERT should recognize this as a random pairing.

By combining these two subtasks, BERT achieves a robust understanding of language context and sentence relationships. MLM helps BERT learn to predict words based on context, enhancing its ability to generate accurate word embeddings in various contexts. NSP, on the other hand, improves BERT's grasp of coherence and logical flow between sentences, which is essential for tasks requiring sentence-level comprehension, such as text summarization and question answering.

These pre-training tasks enable BERT to produce embeddings that are highly effective for a wide range of natural language processing tasks, leading to state-of-the-art performance

in many benchmarks.

6. Applications of Text Mining

The advancements in text mining and deep learning have given rise to Transformer-based models with numerous real-time applications. These applications are transforming our daily lives and benefiting society. Transformer models, such as BERT and GPT, utilize self-attention mechanisms to efficiently manage dependencies across long sequences, significantly enhancing performance in various NLP tasks. The encoder-decoder architecture in transformers, seen in models like T5 and BART, uses an encoder to process input sequences and a decoder to generate outputs, enabling tasks like translation and summarization. This architecture allows for parallel processing, significantly accelerating training and inference.

In translation tasks, these models go beyond sequential translation, encoding based on the semantics of the source language text and the context of both source and target languages. This approach allows for the creation of multilingual translation models using the same architecture with ample examples of translations. Additionally, Transformer models have numerous applications in the public health domain, including drug recommendation, drug design, health chatbots, personalized health recommendations, and telemedicine.

7. Conclusion

This paper has traced the development of text representation methodologies from the foundational Vector Space Model to advanced Attention-based architectures. Beginning with the Vector Space Model, which utilized TF-IDF to numerically represent text, we observed its limitations in capturing contextual semantics. The Distributional Hypothesis provided a theoretical basis for more nuanced vector representations like Word2Vec, significantly enhancing our ability to capture word meanings based on context. RNNs marked a significant advancement in processing sequential data but were constrained by issues like vanishing gradients, which were effectively addressed by LSTM networks. LSTMs improved the handling of long-term dependencies, making them vital for various NLP tasks. The introduction of the Attention mechanism and the subsequent Transformer architecture revolutionized text representation by allowing models to selectively focus on different parts of the input, capturing complex dependencies with unprecedented accuracy. This evolution highlights the remarkable strides made in text representation, culminating in sophisticated models that continue to push the boundaries of natural language processing.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2), 157–166.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**(2), 179–211.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 10–32.

Harris, Z. S. (1954). Distributional structure. *Word*, **10**(2-3), 146–162.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, **26**.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**, 613–620.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, **30**.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 143-160

Exploring COVID-19 Spatial Patterns in Indian Districts: Ridge and Lasso Geographic Weighted Models for Spatial Heterogeneity and Multicollinearity

Megha Sharma[1] **and Shalini Chandra**[1]
[1]*Department of Mathematics and Statistics*
Banasthali Vidyapith, Rajasthan, India

Received: 22 June 2024; Revised: 28 July 2024; Accepted: 31 July 2024

Abstract

This study conduct a comprehensive spatial analysis of COVID-19 across districts in India utilizing data from www.covidindia.org for confirmed cases and deaths, and integrating population characteristics from the National Family Health Survey 5 (2019-2021) and supplementary sources. The objective of the study is to uncover risk factors through spatial modelling while mitigating multicollinearity using the concept of LASSO and ridge regression. Employing spatial analysis, we identify COVID-19 hotspots and coldspots across districts. High-impact districts including Mumbai, Pune, Chennai, Kolkata, and Bengaluru are highlighted, along with lesser-affected districts in central and north-eastern regions. Analysis used geographical weighted regression (GWR) models, incorporating ridge and LASSO techniques to assess the impact of demographic, socioeconomic, climatic, and comorbidity factors on COVID-19 while accounting for spatial relationships. Notably, the GWR with LASSO (GWL) outperforms the other models, with lower RMSE and a notably higher R^2 value. This study reveal significant risk factors such as sanitation facilities, healthcare amenities, women's education, tobacco/alcohol usage, urban population and density, comorbidity, as well as climatic conditions. The GWL model's localized coefficients offer valuable insights into predictor relationships within each spatial unit.

Key words: COVID-19; Geographic weighted model; LASSO regression; Spatial association; Ridge regression.

1. Introduction

The COVID-19 pandemic began in Wuhan, China, in December 2019, caused by the SARS-CoV-2 virus (Li *et al.*, 2020). The COVID-19 pandemic has had a profound impact on individuals' lives, the global economy, and public health. India has been hit particularly hard, suffering economic disruption, unemployment, and a decline in GDP due to COVID-19. The country's healthcare system struggled with resource shortages, limited

Corresponding Author: Megha Sharma
Email: meghasharma15aug@gmail.com

hospital space, and personnel shortages (Sridhar, 2023, Dutta *et al.*, 2021). The pandemic has also triggered social and psychological issues, including increased domestic violence, mental health challenges, and gender inequality (Sardar *et al.*, 2020). Numerous previous studies have identified that social inequalities can facilitate the spread of diseases (Ahmed *et al.*, 2020). Poor living conditions (Pereira and Oliveira, 2020), population density (Rocklöv and Sjödin, 2020), inadequate access to healthcare, and a large proportion of susceptible population, such as the older and those with existing medical conditions (Dutta *et al.*, 2021), are all factors that make any region vulnerable to the spread of the virus. Temperature has also been associated with COVID-19 severity, with similar findings in China (Chen *et al.*, 2020), Indonesia (Tosepu *et al.*, 2020), Turkey (Chung *et al.*, 2021), and the USA (Bashir *et al.*, 2020). Additional risk factors like the prevalence of slums within cities (Sridhar, 2023), smoking habits, and many more contribute to an increased risk of transmission and disparities in access to prevention and treatment measures.

Spatial models have emerged as valuable tools for determining the relationships between the spread of infectious diseases and associated risk factors, incorporating the spatial dimension. Spatial methods are employed to model particular variables at diverse geographical locations, allowing us to address the diversity caused by regional differences (known as spatial heterogeneity) within the data. One effective method for identifying spatial heterogeneity is the Geographically Weighted Regression (GWR) model, which is highly effective in accurately estimating parameters when analyzing COVID-19 data (Sarkar *et al.*, 2021, Ramírez-Aldana *et al.*, 2020, Appiah-Otoo and Kursah, 2022, Adekunle *et al.*, 2020). The GWR model helps illustrate how the association between independent and dependent variables varies across distinct locations within the study area. However, a challenge arises when the risk factors examined within each local model exhibit linear relationships, which is referred to as local multicollinearity. This multicollinearity issue obstructs the precision of parameter estimates and makes it difficult to distinguish the individual effects of these variables.

In the context of addressing the challenge posed by multicollinearity in data, various alternative methodologies have arisen as effective solutions. One such prominent technique is ridge regression, initially proposed by Hoerl and Kennard in 1970 (Hoerl and Kennard, 1970), which has become widely adopted for mitigating the issues associated with multicollinearity. This shrinkage technique incorporates penalty terms into the regression framework to shrink the coefficients, resulting in more stable parameter estimates and mitigate the effect of multicollinearity. Ridge regression introduces a positive bias into the parameter estimation process, effectively guiding the coefficients towards zero. Although this approach yeilds baised results, it reduces variance. Recognizing the potential benefits of combining different methodologies, researchers have explored various approaches, such as combining ridge regression with the Liu estimator (Kejian, 1993) or integrating ridge regression with principal component regression (Baye and Parker, 1984, Chandra and Sarkar, 2016), among others. Additionally, in 1996, Tibshirani introduced a novel technique that has gained extensive attention. The technique combines the advantages of ridge regression with variable selection method, known as the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). LASSO leverages the LARS algorithm (Least Angle Regression) to shrink estimated coefficients towards zero and selectively sets less significant variables to precisely zero. The resulting model is notably interpretable, retaining only the most meaningful predictors relevant to the outcome variable.

Similar to their role in classical regression models, Ridge and LASSO techniques serve to mitigate the impact of multicollinearity in spatial context. Likewise, in the context of GWR model, tackling local multicollinearity is achievable through the incorporation of ridge regression and the LASSO method. Specifically, the utilization of ridge regression within the GWR framework is referred as GWRR, while the integration of the LASSO method with GWR as GWL (Wheeler, 2007). In this study, GWRR and GWL models were applied to investigate spatial heterogeneity and address multicollinearity concerns during the analysis of the COVID-19 pandemic across 626 districts in India.

2. Data

2.1. Data collection

This study included districts from all states and union territories in India, except six states: Assam, Delhi, Goa, Manipur, Telangana, and Sikkim, with no available COVID-19 updates at the district level in the state bulletin. This study took into account district boundaries as of 2019. We extracted district-level data on daily confirmed cases of COVID-19 and associated deaths in India from the website www.covidindia.org. This public domain collects data through state bulletins and official handles. They halted the operation after 18 months of daily updates. As a result, this study limits the availability of data until October 2021. So far, several variables have impacted COVID-19 spread during these pandemic outbreaks, from which some of the essential independent variables that may have affected COVID-19 spread in Indian districts have been selected. Table 1 lists these independent variables, their descriptions containing the reason behind taking these variables into our study, and the sources from which they were obtained.

The data were sourced from various sources, with the primary contributor being National Family Health Survey (NFHS-5). It is initiated by the Ministry of Health & Family Welfare (MoHFW), Government of India, plays a vital role in assessing health conditions in India. This extensive health survey is periodically conducted nationwide, offering health-related indicators at the district, state, and national levels. NFHS-5 was conducted in India during the time period of 2019-2021.

2.2. Data preparations and cleaning

India has undergone several surges of COVID-19 since the onset of the pandemic. Specifically, India encountered two distinct waves of COVID-19 between December 2019 and October 2021. These waves occurred during the periods of March 2020 to December 2020 and January 2021 to October 2021, respectively. This research studied the cumulative confirmed cases (CCC) and cumulative deaths ($deaths$) during the first and second waves at the district level in India. To facilitate the analysis, initial data preparation with a data cleaning procedure aimed at addressing issues like incomplete and duplicate entries by cross-referencing data from various sources. Given the multi-sourced nature of the data, a pivotal step involves data merging, wherein information from diverse origins is consolidated to establish a unified reference point. As the data on risk factors and COVID-19 incidences originates from distinct sources, it is imperative to standardize the data before any analysis can take place. Standardization, in this context, entails bringing different variables onto a common

Table 1: Lists of independent variables, their descriptions and justification and the sources

Abre-viation	Indicators	Assumptions/ Justifications	Data sources
V1	Population below age 15 years	Older population have higher risk of death after infected.	National Family Health Survey (NFHS-5) (2019-21) (District factsheet)
V2	Population living in households with electricity	The environment in which people live plays a significant role in the transmission of COVID-19. Factors such as overcrowding, sanitation and hand hygiene all contribute to susceptibility and should not be overlooked.	
V3	Population living in households that use an improved sanitation facility		
V4	Households using clean fuel for cooking		
V5	Households with any usual member covered under a health insurance	Accessible healthcare systems, affordability, capacity, and health security are vital for managing epidemics and promoting treatment-seeking.	
V6	Women who are literate	Women's literacy empowers them with knowledge, enabling them to understand COVID-19 prevention, access reliable information, and make informed decisions.	
V7	Educated women with 10+ years of schooling		
V8	Proportion of women undernourished	Undernourished and obesity weakened immunity and elevate risk due to underlying health problems when facing COVID-19.	
V9	Proportion of women obese		
V10	Tobacco use among those 15+	Smoking or tobacco or any kind of alcohol being exposed in any form can reduce the risk of COVID-19 infection (WHO 2020).	
V11	Alcohol use among those 15+		
V12	Population Density	High population density and urban areas posing a higher risk for the spread of the highly contagious SARS-CoV-2 virus.	Office of the Registrar General of India
V13	Proportion of urban population		
V14	Health Center [Sub center +PHCs+ CHCs]	Higher population per healthcare institution lower resilience in dealing with COVID-19.	Rural Health statistics
V15	Hypertension among Adults	Blood Sugar Level and Hypertension among Adults (age 15+) may regulate the severity of COVID-19 cases.	(NFHS-5)
V16	Adults' blood sugar levels (Age 15+)		
V17	Average temprature	The severity of COVID-19 associated with temperature and relative humidity	NASA open data portal
V18	Relative Humidity		
V19	Proportion of poor population	Studies have shown that areas with high poverty rates tend to have higher rates of COVID-19 infections.	Global Data Lab

scale to enable comparisons across variable types. These steps were undertaken prior to the transformation of the data into district-level counts and its merging with India's district administrative boundary shape file from the DIV-GIS database using ArcGIS Desktop 10.7.

2.3. Data description

This study involves 19 dependent variables and two independent variables (*CCC* and *deaths*), with data collected from diverse sources across 626 districts in India, resulting in around 13,000 observations—an extensive dataset for analysis. Emphasizing spatial analysis as the foundation, we prioritize reviewing the data before applying statistical methods.

During the first and second phases of the pandemic, specific Indian districts, including Bangalore, Mysuru, Belagavi, Pune, Mumbai, Thane, Nagpur, Ernakulam, Malappuram, Nashik, Kollam, Kolkata, Chennai, Coimbatore, Chittoor, and others in Kerala, Tamil Nadu, Andhra Pradesh, West Bengal, witnessed elevated COVID-19 cases and deaths. Geographical variations were evident, with northern and central states like Lucknow, Varanasi, Kanpur, Jaipur, Jodhpur, Ludhiana, and Jalandhar heavily affected, while areas like Hathras, Mahoba, Burhanpur, Agar Malwa, Mandla, and Baranala reported fewer cases. Central and northeastern regions generally had lower confirmed cases and deaths in both waves.

According to the data, higher population density is observed in Bihar, West Bengal, and Kerala, with 29 districts among the top 10%. On average, 4.24% of the population in these districts is aged 65 and above. Notably, Maharashtra, Kerala, Karnataka, Goa, and Punjab display a significant prevalence of districts with an aging population. Specifically, 15 out of Maharashtra's 36 districts and 9 out of Kerala's 14 districts rank in the top 10% for the percentage of elderly population. On average, 20.19% of households in Indian districts lack water supply within their premises. The data reveal pronounced water supply challenges in numerous districts of Odisha, Madhya Pradesh, and Rajasthan. Noteworthy is that 12 out of 14 districts in Kerala and 21 out of 30 districts in Tamil Nadu are in the highest quartile (>27%) for the proportion of women grappling with obesity. Additionally, 10 districts in Andhra Pradesh and 5 in Maharashtra fall into this category. Kerala, Goa, Tamil Nadu, and Andhra Pradesh also exhibit a significant presence of districts with the highest percentages (>7.5%) of the population facing elevated blood sugar levels. The data highlight certain districts in Rajasthan, such as Jaisalmer and Barmer, known for extremely high temperatures. Districts in the northern plains, including parts of Uttar Pradesh, Bihar, and Haryana, may also experience high temperatures. Gujarat, Maharashtra, and certain parts of Kerala might encounter high humidity levels. Alcohol and tobacco consumption is notably high in districts of northeastern states, Punjab, Goa, and select districts in Rajasthan.

According to the National Family Health Survey (NFHS-5), about 41% of India's total population has at least one member enrolled in health insurance or a health scheme. Rajasthan and Andhra Pradesh lead with the highest proportions of households covered (88% and 80%, respectively), while the Andaman and Nicobar Islands and Jammu and Kashmir show the lowest coverage, each below 15%.

2.4. Visualization and exploration

By using visualization techniques, patterns and discrepancies in the data are identified. The most widely used approach for visualizing this type of data is through choropleth maps that employ quantile breaks. These maps use various colors to depict the intensity of variables of interest in each geographic region. Such maps have been included in the study to present the spatial distribution of COVID cases and deaths in further sections. Exploration of spatial data includes cluster analysis to identify whether observed spatial patterns are random, using either nonspecific (global) or specific (local) techniques. Moran's I statistic, a global technique, is employed to ascertain cluster presence across the entire study area. Moran's I computes global spatial autocorrelation among observations and ranges from -1 to 1. Negative values indicate dispersion (clustering of dissimilar values), positive values indicate clustering (clustering of similar values), and values near zero suggest absolute spatial randomness, implying no autocorrelation. However, because Moran's I statistic is incapable of providing precise information on cluster locations, the LISA (local indicators of spatial association) tool was utilized to calculate local spatial autocorrelation. This method describes significant correlations at specific locations as local spatial clusters (hot spots) or correlations between observations and neighboring observations Anselin (1995). The next section is about the models and estimators considered in this study.

3. The models and estimators

3.1. Geographical weighted regression (GWR)

The GWR model estimated local interactions between the dependent and independent variables by fitting a regression model to each feature (spatial unit) in the dataset (Brunsdon *et al.* (1998)). The GWR model for each feature is

$$y_i = \beta_{i0} + \sum_{j=1}^{m} X_{ij}\beta_{ij} + \epsilon_i, i = 1, 2, ..., n. \tag{1}$$

where y_i represents the dependent variable at a specific location i, β_{i0} stands for the intercept parameter at that same location i, β_{ij} symbolizes the local regression coefficient pertaining to the j^{th} explanatory variable at location i, X_{ij} signifies the value of the j^{th} explanatory variable at location i, and ϵ_i corresponds to the random error observed at location i. The parameters estimates for each independent variable at i^{th} location is given by

$$\hat{\beta(i)} = (X^T W(i) X)^{-1} X^T W(i) y \tag{2}$$

where $\hat{\beta(i)}$ is $m \times 1$ vector of parameter estimates, $W(i)$ is spatial weight matrix calculated by the exponential kernel function which is defined as

$$w_k(i) = \begin{cases} \left[1 - \left(\frac{d_{ik}}{bw}\right)^2\right]^2, & \text{if } k \in \{N_i\} \\ 0, & \text{if } k \notin \{N_i\} \end{cases} \tag{3}$$

where d_{ik} is the distance between feature location i & k with bandwidth bw derived from the Euclidean distance between observation locations and neighboring points, this measure

ensures that the region remains influenced by proximate neighbors within this radius. The set N_i includes observations within this N^{th} nearest neighbor distance. Weights are zero for observations beyond this range, except for observation i which gets a weight of 1. Kernel function assigns higher weights to observations that are closer to the calibration location i. To fit the GWR model, the kernel bandwidth is estimated through cross-validation (CV) using all feature locations, followed by weight calculation using (3). CV function is outlined as

$$CV(bw) = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(bw)]^2 \tag{4}$$

where, $\hat{y}_{\neq i}(bw)$ is the estimated value of y achieved by excluding the data point at the i^{th} location during prediction. The bandwidth bw will be derived through an iterative procedure aiming to minimize the CV score.

3.2. Addressing multicollinearity: diagnosis and remediation

Collinearity's presence among independent variables can diminish the precision of coefficients (Wheeler and Tiefelsdorf (2005)). There are valuable diagnostic tools designed to uncover collinearity issues that might disrupt the interpretation of estimated regression coefficients. These diagnostic methods are derived from conventional regression techniques. Approaches for identifying collinearity among independent variables comprise metrics like variance inflation factors (VIF) and condition indices. Moreover, Ridge regression and LASSO are frequently employed methods for mitigating the multicollinearity.

Ridge regression

Ridge regression was uniquely formulated to alleviate the impacts of collinearity through the imposition of penalties on the magnitudes of regression coefficients. This strategy diminishes the impact of variables with comparatively low variance within the model. The parameter for ridge regression is determined by minimizing the sum of squared errors, introducing constraints that compel coefficients to approach zero (Hoerl and Kennard, 1970). More precisely, the ridge estimator coefficient is derived by minimizing the equation

$$\hat{\beta}_R = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_i \right)^2 \tag{5}$$

with $\sum_{j=1}^{m} (\beta_i)^2 \leq \rho$, where ρ is a control shrinkage amount. Then parameter estimates is obtained by

$$\hat{\beta(i)}_R = (X^T X + CI)^{-1} X^T y \tag{6}$$

where I is an identity matrices and C represents positive coefficient bias.

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is the regularization and penalization technique which shrinks the regression coefficients towards zero, also puts least significant variable coefficients to zero. This leads to a simplified and interpretative model, retaining only the significant predictors for the outcome variable (Tibshirani, 1996). The coefficients of Lasso parameters cannot be

directly calculated through closed-form equations, unlike Ridge regression. Instead, they are determined using quadratic programming techniques. LASSO is defined as follows as

$$\hat{\beta} = argmin \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 . \tag{7}$$

where $\sum_{i=j}^{p} |\hat{\beta}_j| \leq tp$ (threshold parameter). It is established that tp is a parameter governing the level of shrinkage in LASSO coefficient estimation, where $tp \geq 0$.

3.3. Geographic Weighted Ridge Regression (GWRR)

GWRR is a modified technique within the domain of spatial regression, in which GWR model combine with the ridge regression (Wheeler, 2009). Estimator of parameters of GWRR model at i^{th} location is obtained by

$$\hat{\beta(i)}_{GWRR} = (X^T W(i) X + CI)^{-1} X^T W(i) y. \tag{8}$$

The process involves predicting bandwidth values to form a weighted matrix, minimizing bias using CV, and iteratively determining the coefficient value C for each bandwidth. These results are then applied to estimate spatial model with ridge regression coefficients.

3.4. Geographically Weighted LASSO (GWL)

LASSO's application within a GWR model, later recognized as Geographically Weighted LASSO (GWL), addresses spatial variations and local multicollinearity. GWL offers unbiased coefficient estimates and enhances prediction accuracy (Wheeler, 2009). LASSO parameter estimation in GWL is executed concurrently, relying on a pre-established kernel bandwidth. During the GWL parameter estimation process, the shrinkage (s) value is determined prior to the final LASSO solution. Shrinkage parameter estimation in GWL's LASSO model is achieved through cross-validation (CV), resulting in a distinct shrinkage parameter for each geographical location.

4. Model selection criteria

Coefficient of determination (R^2) and root mean square error (RMSE) were used to compare the performances of various models. R^2 measures the goodness of fit; its values range from 0 to 1. Furthermore RMSE calculated how closely predicted values align with actual observations by measuring the average error magnitude. The model with lower RMSE value and higher value of R^2 better fits the observed data. In this study, analysis have been performed in R software version 4.3.1 using various packages such as sp, spgwr, spdep, gwrr, and spatstat.

5. Empirical findings

5.1. Visualization and exploration

This study employed choropleth maps using quantile breaks to visualize the total confirmed cases and total deaths during the pandemic outbreak, yielding successful results.

These maps use various colors to depict the intensity of variables of interest in geographic region. Referring to Figure 1, the districts that exhibited the highest numbers of confirmed COVID-19 cases and deaths were Bangaluru, Mysuru, Belagavi, and 13 other districts in Karnataka. Additionally, in Maharashtra, the districts of Pune, Mumbai, Thane, Nagpur, and 29 out of 35 districts stood out. Similar trends were observed in Kerala, Tamil Nadu, Andhra Pradesh, and West Bengal, particularly in districts such as Ernakulam, Malappuram, Nashik, Kollam, Kolkata, Chennai, Coimbatore, Chittoor, and their adjacent districts. These districts were among the most affected during the entire duration of the pandemic analyzed in this study. There were marked geographical distinctions among the northern and central states of India, with some districts like Lucknow, Varanasi, Kanpur, Jaipur, Jodhpur, Ludhiana and Jalandhar experiencing a high level of contagion while other areas like Hathras, Mahoba, Burahnpur, Agar Malwa, Mandla and Baranala and the locations around them having a much lesser effect. In contrast, the central and northeastern regions districts had the fewest confirmed cases and deaths in both waves. The global Moran's I statistic values for cumulative confirmed cases and deaths due to COVID-19 were significant for both waves (0.31, 0.43, and 0.27, 0.43, respectively, with p-value=0.0001[$<$ 0.05]), indicating strong spatial autocorrelation among Indian districts. Further, the LISA tool was employed to identify significant local clustering and detect non-clustered areas within the study that may be missed by global tests.

Using the LISA tool, the study found that the districts with the highest concentration of confirmed cases and deaths during both waves were the same, including Maharashtra, Kerala, Andhra Pradesh, West Bengal, and Karnataka. In contrast, the northern and central regions exhibited low clustering during the first wave, and the central region was also identified as having low clustering in the second wave (see Figure-2) and only a few districts fell into the high-low and low-high clusters.

5.2. Spatial modelling

The dataset encompassing all independent variables used in this study exhibits consistent values across both waves of COVID-19. With the aim of exploring the influence of these variables on the occurrences of COVID-19 cases and related fatalities, a comprehensive approach was adopted by examining the entire temporal span. The outcomes of the Global Moran's I test [Value of Global Morna's I= 0.42 for *CCC* and 0.45 for *deaths*] and the Breusch-Pagan test with p-value = 0.0001 $<$ 0.05 indicate that the data employed in this study exhibit noteworthy spatial heterogeneity.

5.2.1. Local multicollinearity

The presence of multicollinearity among independent variables can be ascertained by examining the VIF values of local observations and the condition index specific to that particular location. The summary of VIF values and condition indices is presented in the Table 2 as the GWR model incorporates all these independent variables to predict total confirmed cases and total number of deaths.

Upon referencing the table 2, it becomes evident that numerous locations exhibit VIF values and condition indices exceeding 30, indicating a significant level of concern regard-

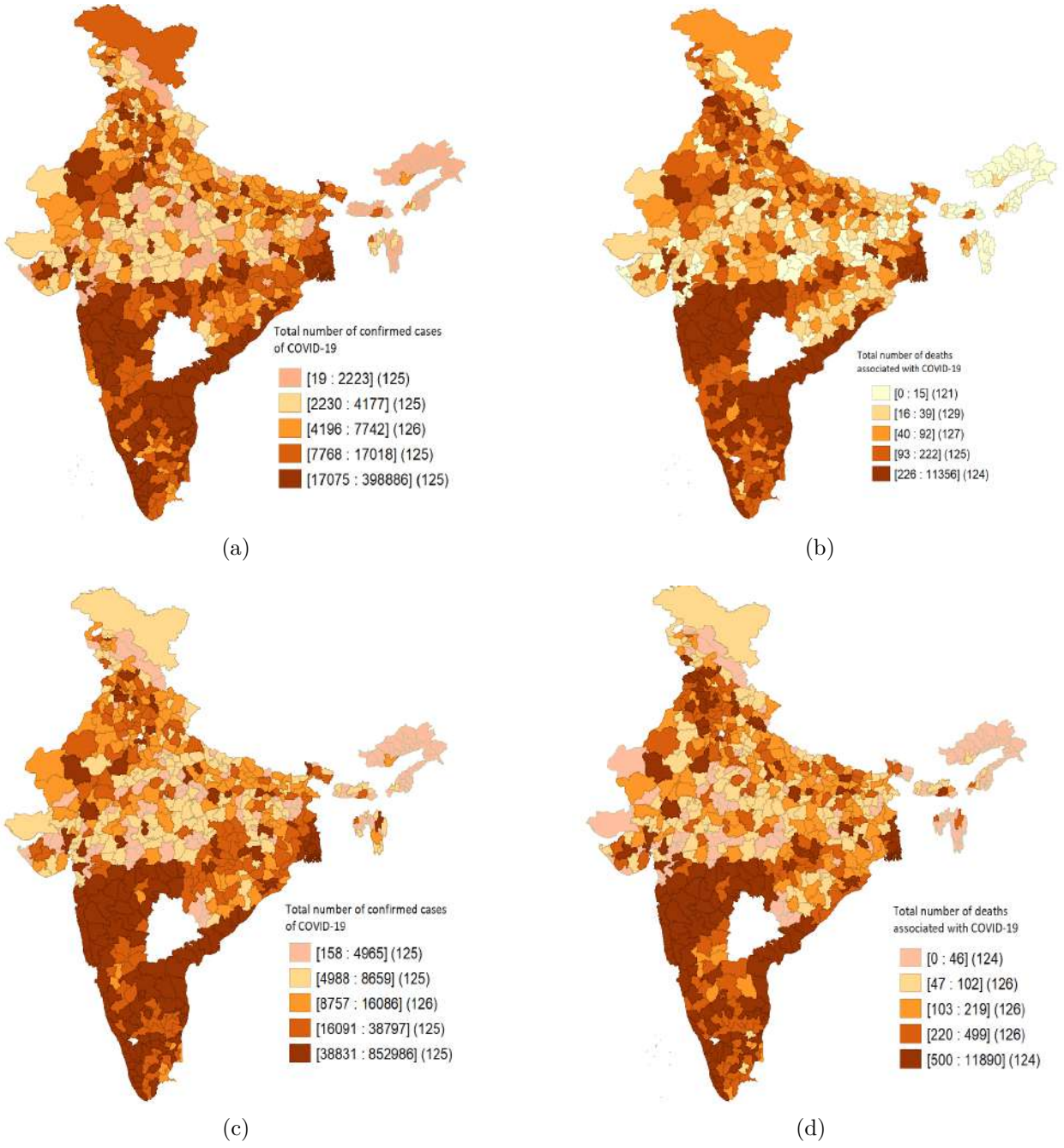


Figure 1: Quantitative spatial distribution of *Cumulative confirmed cases* (a, c) and *Total deaths* (b, d) in 1^{st} wave and 2^{nd} wave respectively in Indian districts

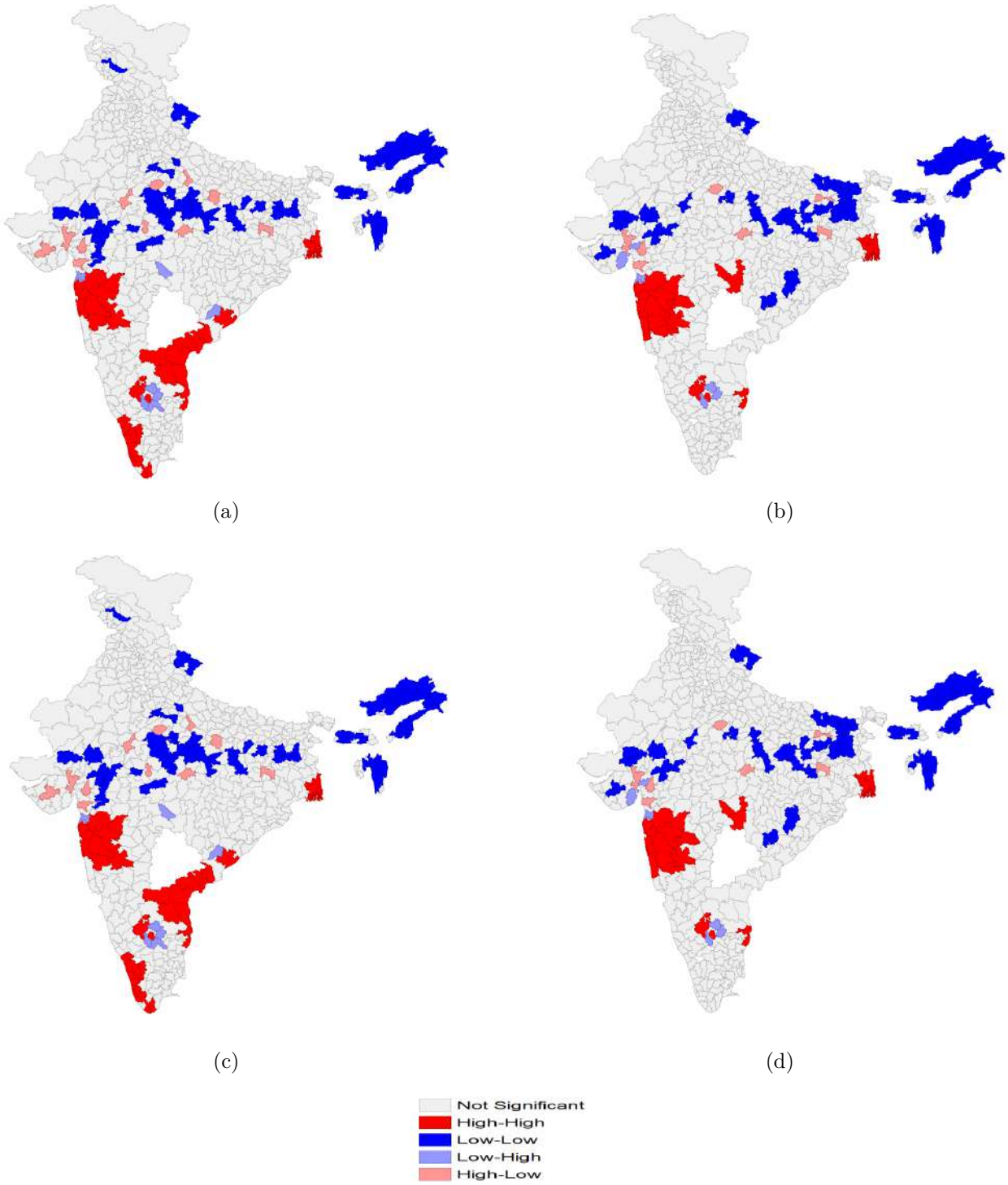


Figure 2: LISA clusters of *Cumulative confirmed cases* (a, c) and *Total deaths* (b, d) associated with COVID-19 in 1^{st} wave and 2^{nd} wave respectively in Indian districts

Table 2: Summary table of VIF and condition index at local level

	CCC			*deaths*		
VIF	Min.	Mean	Max.	min.	Mean	Max.
V1	3.049	3.23	3.416	3.037	3.954	4.955
V2	1.525	1.541	1.555	1.493	1.682	1.889
V3	2.516	2.657	2.87	2.208	3.056	5.138
V4	3.3	3.446	3.547	3.161	3.964	5.407
V5	1.339	1.419	1.452	1.327	1.828	2.244
V6	6.164	6.292	6.47	6.189	8.00	10.517
V7	5.055	5.194	5.462	4.759	6.672	9.71
V8	26.09	28.49	33.56	16.99	24.53	46.43
V9	19.36	20.12	22.43	14.28	21.31	31.89
V10	1.766	1.813	1.898	1.643	2.135	2.71
V11	1.688	1.814	2.155	1.18	1.877	3.338
V12	1.321	1.326	1.336	1.303	1.408	1.527
V13	2.509	2.551	2.593	2.302	2.63	3.096
V14	1.061	1.067	1.076	1.086	1.166	1.403
V15	25.89	28.06	32.71	17.13	23.41	42.63
V16	21.79	23.44	25.01	16.24	23.5	34.16
V17	1.829	2.008	2.336	1.275	3.074	7.281
V18	2.442	2.561	2.75	1.886	3.372	5.957
V19	2.546	2.71	2.906	1.876	3.298	4.901
CI	18.15	19.96	24.26	16.26	22.26	40.51

ing multicollinearity at particular location. This degree of multicollinearity contributes to heightened variability in the coefficient parameters, leading to less stable results. Furthermore, the existence of multicollinearity gives rise to an unstable model, a fact that becomes apparent through the modification of the classical GWR model. This modification involves the integration of multicollinearity mitigation techniques such as LASSO and ridge regression. The enhanced GWR model's effectiveness can be observed in the Table 3, where a comparison is made between the GWR model, the GWRR (Geographically Weighted Ridge Regression) model, and the GWL (Geographically Weighted LASSO Regression) model.

Table 3: Comparison table of modified and unmodified GWR model

		GWR	GWRR	GWL
CCC	$RMSE$	0.7921	0.5322	0.4630
	R^2	0.4558	0.7162	0.8371
	bw	4.9686	0.7724	0.0361
deaths	$RMSE$	0.6978	0.4597	0.4398
	R^2	0.534	0.8214	0.8625
	bw	3.1805	0.6197	0.3012

Upon evaluating the RMSE scores and R-squared values, it becomes apparent that the GWL model delivers the most accurate fit across the entire duration (refer to the Table 3). The GWL model is capable of explaining an average of 83% of the variation in cumulative

COVID-19 cases and 86% of the variation in COVID-19-related deaths across all districts of India, taking care of the challenge posed by multicollinearity. Furthermore, the GWL model provides a coherent interpretation for the disparities in confirmed COVID-19 cases and associated deaths among Indian districts. A comprehensive summary of coefficient estimates for all independent variables within the GWL model will be presented in the subsequent section.

5.2.2. GWL model summary

In GWL modelling, similar to how LASSO works, the importance of coefficients gradually decreases until they become zero due to shrinkage. When a coefficient reaches zero, it loses its influence on the outcomes of the model. Through an iteration process driven by cross-validation, GWL yields a bandwidth value of 0.09 for *CCC* and 0.89 for deaths. This bandwidth parameter, along with the associated shrinkage value, contributes to the delineation of GWL's specific parameters, all of which are detailed in the Table-4.

Referring to the summarized Table-4, it becomes evident that certain independent variables - such as Population age, households with electricity, the percentage of women who are obese, and relative humidity - possess either a zero or near-zero mean coefficient value across all regions in the context of modelling total confirmed cases. This observation signifies that these variables exert negligible influence on the incidence of COVID-19 cases. Similarly, when it comes to predicting the number of deaths, both relative humidity and household electrification also demonstrate insignificant effects. However, in contrast to COVID-19 cases, the prediction of COVID-19 deaths shows a positive association with individual age, underscoring the elevated risk of mortality among the elderly population subsequent to infection.

The cumulative confirmed COVID-19 cases in an Indian district are linked positively to factors such as the availability of sanitation facilities and healthcare services, the percentage of undernourished women, tobacco and alcohol consumption, population density, urbanization, average temperature, and the education level of women. Conversely, they are negatively associated with the number of people living in poverty. However, concerning the total number of COVID-19 related deaths, there is a negative correlation with the availability of sanitation facilities and health insurance coverage.

The data presented in the Table-4 indicates that the GWL model zeroes out coefficients for various factors in different locations, resulting in varied parameter magnitudes across regions. As a result, the GWL model generates distinct models with differing coefficients for various locations. To illustrate this, we have provided the model for the two most severely impacted districts (Pune and Bengaluru) in different zones.

$$y^*(TCC_{Pune}) = 4.55 - 0.0525V3^* + 0.115V4^* + 0.24V6^* + 0.13V7^* - 0.65V8^* + 0.11V9^* + 0.33V11^* + 0.33V13^* + 0.51V16^* + 0.20V18^*$$

$$y^*(Deaths_{Pune}) = 2.33 + 0.09V1^* - 0.001V3^* + 0.0086V4^* + 0.076V6^* - 0.047V8^* + 0.032V9^* + 0.75V11^* + 0.007V13^* + 0.078V16^* + 0.0091V17^*$$

Table 4: Summary statistics for GWL parameter estimates

	CCC					*deaths*				
	Intercept	V1	V2	V3	V4	Intercept	V1	V2	V3	V4
Min.	-2.361	-0.018	-0.016	-0.025	0.000	-2.516	-0.04321	-0.027	-0.147	0.000
1st Qu	-0.150	0.000	0.000	0.000	0.011	-0.141	0.000	0.000	0.000	0.110
median	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000	0.137
mean	0.039	0.001	0.000	0.002	0.036	0.0257	-0.003	0.000	-0.010	0.119
3rd Qu.	0.000	0.000	0.000	0.000	0.054	0.000	0.000	0.000	0.000	0.155
Max.	4.301	0.126	0.000	0.131	0.101	4.908	0.111	0.000	0.000	0.209
	V5	V6	V7	V8	V9	V5	V6	V7	V8	V9
Min.	-0.087	-0.054	0.000	-0.591	-0.514	-0.145	0.000	-0.016	-0.660	-0.413
1st Qu	0.000	0.000	0.075	0.000	0.000	0.000	0.000	0.000	0.000	0.000
median	0.000	0.000	0.104	0.000	0.000	0.000	0.036	0.016	0.000	0.000
mean	-0.002	-0.001	0.117	0.003	-0.014	-0.001	0.048	0.020	-0.015	-0.005
3rd Qu.	0.000	0.000	0.169	0.000	0.000	0.000	0.076	0.034	0.000	0.000
Max.	0.029	0.037	0.355	0.070	0.000	0.000	0.337	0.143	0.000	0.000
	V10	V11	V12	V13	V14	V10	V11	V12	V13	V14
Min.	-0.002	-0.038	0.000	0.000	0.000	0.000	0	0.000	0.024	0.000
1^{st} Qu.	0.000	0.000	0.030	0.110	0.000	0.000	0.000	0.143	0.080	0.000
median	0.000	0.000	0.094	0.125	0.000	0.000	0.000	0.199	0.090	0.000
mean	0.001	0.003	0.088	0.113	0.005	0.001	0.000	0.176	0.083	0.001
3^{rd} Qu.	0.000	0.000	0.136	0.135	0.000	0.000	0.000	0.245	0.101	0.000
Max.	0.089	0.062	0.246	0.152	0.089	0.116	0.101	0.344	0.110	0.029
	V15	V16	V17	V18	V19	V15	V16	V17	V18	V19
Min.	0.000	0.000	-0.003	-0.052	-0.119	0.000	0.000	-0.068	-0.182	-0.136
1^{st} Qu.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
median	0.027	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
mean	0.039	0.012	0.028	0.010	-0.003	0.006	0.006	0.024	0.000	-0.001
3^{rd} Qu.	0.058	0.000	0.032	0.000	0.000	0.000	0.000	0.045	0.000	0.000
Max.	0.602	0.489	0.221	0.222	0.000	0.520	0.440	0.147	0.113	0.000

$$y^*(TCC_{Bengaluru}) = 0.213 - 0.0034V6 + 0.029V12^* + 0.23V13^* - 0.65V14^* + 0.063V15^* + 0.031V16^*$$

$$y^*(Deaths_{Bengaluru}) = 1.84 + 0.004V1^* + 0.0076V8^* + 0.043V9^* + 0.0027V13^* + 0.032V16^* + 0.0038V17^*$$

6. Discussion

The current research implemented spatial analysis techniques to analyse the spatial distribution and clustering of COVID-19 in Indian districts. The data indicated a significant spatial heterogeneity in the distribution of COVID-19 across the country, with clusters of cases and deaths found to be almost identical for both waves with high intensity. The main reason for the lack of change in hotspots from the first to the second wave is attributed to the need to identify and monitor hotspots in the first wave properly. Further, the resurgence of cases has been linked to mass gatherings and non-adherence to safety protocols such as

wearing masks, social distancing, and handwashing. Significant clustering of COVID-19 cases was identified in specific districts of Maharashtra, Kerala, Andhra Pradesh, West Bengal, and Karnataka, forming clusters characterized by high numbers of COVID-19 cases and deaths. Conversely, districts in the northern and southern regions formed clusters with low COVID-19 cases and deaths. These findings imply that the risk of infection was not same across districts. The observed spatial autocorrelation suggests that the disease may spread from high-risk districts to neighbouring areas, underscoring the importance of coordinated efforts to control the spread of the disease across all districts. The findings of this study suggest that proper identification and monitoring of hotspots in the first wave could have enabled more effective management of COVID-19 cases in the second wave.

Spatial models have demonstrated their usefulness as tools for comprehending and examining pandemic behaviour. Nevertheless, the issue of multicollinearity often poses a challenge for these models. In the present study, it was observed that the independent variables utilized to identify risk factors exhibited a considerable degree of collinearity. In response to this concern, the ridge and LASSO techniques were initially employed on the spatial models. It was discovered that among the spatial models implemented in this research, the GWL model exhibited superior performance. By integrating spatially varying coefficients, the GWL model effectively captured localized fluctuations and heterogeneity in the association between the dependent and independent variables, while also addressing collinearity concerns among the independent variables. Although the GWL model generates different models and identifies significant independent variables for different locations, this study also determined the independent variables that, on average, influence COVID-19 cases and deaths across Indian districts.

The findings of the GWL model demonstrated a positive relationship between the high temperatures and the spread of the COVID-19 virus. This relationship is supported by epidemiological evidence indicating that an increase in ambient temperature can result in a higher transmission rate (Chen *et al.* (2020), Tosepu *et al.* (2020), Bashir *et al.* (2020)). The virus can endure in the air longer at higher temperatures and be more easily transmitted through droplets. Additionally, greater access to healthcare facilities was positively correlated with more accurate diagnosis and reporting of COVID-19 cases and deaths, which may explain the higher number of cases and deaths in these areas. Furthermore, areas with a high proportion of the population having alcohol and tobacco consumption, and high literacy rates among women were also positively associated. Smoking and drinking habits can weaken the immune system and make individuals more susceptible to the virus. High literacy rates among women could increase awareness of the virus and its symptoms, increase testing, more accurate diagnosis and reporting of cases, and increase transmission opportunities. Consistent with prior investigations, the proportion of the population residing in urban settings and the density of specific districts exerted anticipated effects on COVID-19 incidence and mortality within those particular areas. Conversely, certain variables like relative humidity, household access to electricity, and possession of health insurance exhibited negligible influence on COVID-19 patterns in Indian districts. Furthermore, no specific age group demonstrated disproportionate susceptibility to COVID-19; however, elderly individuals were identified as having an elevated risk of mortality attributed to the virus.

The GWL model introduces spatial variability in coefficients, capturing differences in various locations. The range of coefficients gives insight into how relationships between vari-

ables change across space. The GWL model's findings help understand the degree to which the identified risk factors account for differences in COVID-19 cases and deaths in diverse districts. For instance, districts like Mumbai, Chennai, Pune, Kolkata, Sagar, Jabalpur, Narshimpur, Raisen, Porbandar, Junagarh, and Somnath exhibit significant variation in COVID-19 outcomes (ranging from 80 to 86 percent). This highlights the strong impact of the identified risk factors in these areas. Conversely, the considered variables struggle to explain variations in certain districts, particularly in parts of Punjab (such as Bhatinda, Faridkot, Moga) and the northeastern region. Similarly, regions like Sirsa, Panchkula, and districts in Himachal Pradesh, JK, and Ladhak have limited explanatory capability. These anomalous ranges of coefficient estimate in these regions suggest that other unaccounted factors may play a more significant role in shaping COVID-19 outcomes.

The overall findings suggest that addressing multicollinearity in spatial models can significantly enhance their robustness and reliability. By mitigating the impact of collinearity among independent variables, researchers can obtain more accurate and trustworthy results. Consequently, this enables the identification of high-risk districts where targeted interventions can be implemented. Measures such as rigorous testing and contact tracing, targeted lockdowns, and intensified public health messaging can be strategically deployed to effectively control and mitigate the spread of the virus in these specific areas. However, limitations of the study include its reliance on reported case counts and its focus on only two waves of the pandemic due to data unavailability, which may not capture the full impact of the virus. Therefore, future research should address these shortcomings to develop more effective strategies for mitigating them.

7. Conclusion

This study aimed to employ spatial econometric modelling methods to enhance understanding of the spatial structures and associations among locations in India and to analyse the transmission patterns of COVID-19. By considering spatial proximity, the study assessed the impact of demographic, socioeconomic, climatic, and comorbidity on total COVID-19 cases and deaths across districts in India. Additionally, this study addressed the issue of multicollinearity in spatial models through the utilization of ridge and LASSO techniques. This approach successfully reduced interdependence among variables and improved the model's accuracy, allowing for the identification of key risk factors associated with the phenomenon under investigation. Significantly, the study brought to light the influence of distinct district factors on the occurrence of COVID-19. These factors encompass sanitation facilities, accessibility to healthcare, pre-existing medical conditions like high blood pressure and diabetes, women's educational levels, rates of tobacco and alcohol consumption, climatic conditions, and the presence of undernourished women. Moreover, the research established that older populations are at a heightened risk of mortality following infection with COVID-19. The findings of this study can inform the development of prevention strategies and strengthen public health capacities, particularly in regions where the healthcare system may be limited. However, it is worth noting that a limitation of the analysis was the lack of district-level data on deaths beyond October 2021 in India.

Acknowledgements

We would like to express my sincere gratitude to Banasthali Vidyapith for providing the necessary resources and support throughout the duration of this research. The academic environment and facilities at the university have played a crucial role in the successful completion of this study.

References

Adekunle, I. A., Onanuga, A. T., Akinola, O. O., and Ogunbanjo, O. W. (2020). Modelling spatial variations of coronavirus disease (COVID-19) in Africa. *Science of the Total Environment*, **729**, 138998.

Ahmed, F., Ahmed, N., Pissarides, C., and Stiglitz, J. (2020). Why inequality could spread Covid-19. *The Lancet Public Health*, **5**, e240.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*, **27**, 93–115.

Appiah-Otoo, I. and Kursah, M. B. (2022). Modelling spatial variations of novel coronavirus disease (Covid-19): evidence from a global perspective. *GeoJournal*, **87**, 3203–3217.

Bashir, M. F., Ma, B., Komal, B., Bashir, M. A., Tan, D., Bashir, M., *et al.* (2020). Correlation between climate indicators and Covid-19 pandemic in New York, USA. *Science of the Total Environment*, **728**, 138835.

Baye, M. R. and Parker, D. F. (1984). Combining ridge and principal component regression: a money demand illustration. *Communications in Statistics-Theory and Methods*, **13**, 197–205.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **47**, 431–443.

Chandra, S. and Sarkar, N. (2016). A restricted r-k class estimator in the mixed regression model with autocorrelated disturbances. *Statistical Papers*, **57**, 429–449.

Chen, B., Liang, H., Yuan, X., Hu, Y., Xu, M., Zhao, Y., Zhang, B., Tian, F., and Zhu, X. (2020). Roles of meteorological conditions in Covid-19 transmission on a worldwide scale. *MedRxiv*, **3**, 2020–03.

Chung, H. W., Apio, C., Goo, T., Heo, G., Han, K., Kim, T., Kim, H., Ko, Y., Lee, D., Lim, J., *et al.* (2021). Effects of government policies on the spread of Covid-19 worldwide. *Scientific Reports*, **11**, 20495.

Dutta, I., Basu, T., and Das, A. (2021). Spatial analysis of Covid-19 incidence and its determinants using spatial modeling: A study on India. *Environmental Challenges*, **4**, 100096.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Kejian, L. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, **22**, 393–402.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, **382**, 1199–1207.

Pereira, M. and Oliveira, A. M. (2020). Poverty and food insecurity may increase as the threat of Covid-19 spreads. *Public Health Nutrition*, **23**, 3236–3240.

Ramírez-Aldana, R., Gomez-Verjan, J. C., and Bello-Chavolla, O. Y. (2020). Spatial analysis of covid-19 spread in Iran: Insights into geographical and structural transmission determinants at a province level. *PLoS Neglected Tropical Diseases*, **14**, e0008875.

Rocklöv, J. and Sjödin, H. (2020). High population densities catalyse the spread of Covid-19. *Journal of Travel Medicine*, **27**, taaa038.

Sardar, S., Abdul-Khaliq, I., Ingar, A., Amaidia, H., and Mansour, N. (2020). Covid-19 lockdown: A protective measure or exacerbator of health inequalities? a comparison between the United Kingdom and India. a commentary on "the socio-economic implications of the coronavirus and covid-19 pandemic: A review. *International Journal of Surgery (London, England)*, **83**, 189.

Sarkar, S. K., Ekram, K. M. M., and Das, P. C. (2021). Spatial modeling of Covid-19 transmission in Bangladesh. *Spatial Information Research*, **29**, 715–726.

Sridhar, K. S. (2023). Urbanization and Covid-19 prevalence in India. *Regional Science Policy & Practice*, **15**, 493–505.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, 267–288.

Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P., *et al.* (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, **725**, 138436.

Wheeler, D. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, **7**, 161–187.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, **39**, 2464–2481.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A*, **41**, 722–742.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 161-180

Nonparametric Rectangular Prediction Regions for Setting Reference Regions in Laboratory Medicine

Michael Daniel Lucagbo[1,2] **and Thomas Mathew**[1]
[1]*Department of Mathematics & Statistics, University of Maryland Baltimore County, Baltimore, Maryland, USA*
[2]*School of Statistics, University of the Philippines Diliman, Quezon City, Philippines*

Received: 21 May 2024; Revised: 20 August 2024; Accepted: 22 August 2024

Abstract

The interpretation of clinical laboratory results of patients depends crucially on an established reference interval for each biochemical analyte. Often, the health status of a patient is decided based on the values of multiple analytes, and this calls for the use of a multivariate reference region, so that the possible cross-correlations among the analytes can be taken into consideration. If multivariate normality can be assumed, one of the recommendations in the laboratory medicine literature is to use an ellipsoidal prediction region as a reference region. However, an ellipsoidal region cannot detect if a particular analyte is within the normal range; a rectangular prediction region is necessary for this purpose. Under multivariate normality, rectangular prediction regions are available in the literature, and these can be used as reference regions for assessing the outlyingness of individual analytes. The present work is motivated by the need to construct such regions without making the multivariate normality assumption. Two approaches are pursued in our work: based on Box-Cox transformation of each marginal variable, and based on estimating each marginal density using a kernel density estimator. A non-parametric bootstrap is then employed for estimating the required prediction factors. Through simulations, it is noted that the resulting rectangular prediction regions meet the coverage probability requirements satisfactorily. The methodology can also be adopted for computing one sided prediction limits, or a combination of one-sided prediction limits for some variables, and two-sided prediction intervals for the rest. Algorithms are provided to compute the regions, and illustrative examples are also given.

Key words: Bootstrap; Box-Cox transformation; Kernel density estimator; One-sided prediction limits; Two-sided prediction intervals.

1. Introduction

Reference intervals are used in numerous medical applications such as the interpretation of blood tests, clinical urine tests, vital signs, and so forth. Due to the extensive

Corresponding Author: Thomas Mathew
Email: mathew@umbc.edu

applications of reference intervals in the field of laboratory medicine, Horn and Pesce (2005) have called it "the most widely used medical decision-making tool." A reference interval is defined as the interval that contains 95% of the "central measurements" for a reference population. Thus, the endpoints of a two-sided reference intervals are the 2.5th and 97.5th percentiles of the reference population. If only a one-sided reference limit is of interest, then the required reference limit is either the 95th percentile (for an upper reference limit), or the 5th percentile (for the lower reference limit).

Since the population percentiles are unknown in actual practice, reference ranges are typically constructed based on data from a random sample of individuals (reference subjects). The selection of reference subjects is obviously critically important, and the prevailing view is that the reference subjects should consist of healthy subjects. For example, Wellek (2011) mentions that a population suitable for establishing reference values should consist of people free of the disease condition one aims to detect. A naive way to construct reference intervals using the available data is to use estimated sample percentiles. Under such a scheme, the percentage of the population covered by the resulting reference interval will be different from 95%. As an alternative to this, a common approach is to compute a 95% prediction interval and use it as a reference interval. This approach has been recommended in practice; see the document by the National Committee for Clinical Laboratory Standards (2010) and the User's Guide by Horn and Pesce (2005). Another option, advocated by authors such as Liu et al. (2021) and Lucagbo and Mathew (2023) is to compute a 95% tolerance interval, which can be used to assess the uncertainty in the estimated reference intervals. In this study, we adopt the prediction interval criterion.

For complex diagnoses, such as for kidney function or liver function, several analytes are needed to properly assess the health status of a patient. For such scenarios, the use of separate univariate reference intervals is an inefficient way to proceed since such an approach disregards the cross-correlations among the analytes. Moreover, it increases the risk of false-positive diagnoses (Harris, 1981; Winkel et al., 1972). When multiple analytes are needed to assess the health status of individuals, a multivariate reference region (MRR), which accounts for the cross-correlations among analytes, is needed. Nonetheless, MRRs are not without shortcomings. The conventional approach to compute MRRs, especially under the assumption of multivariate normality, is to construct ellipsoidal regions. Unfortunately, ellipsoidal reference regions are difficult to interpret. Moreover, ellipsoidal reference regions tend to produce false negative results in the presence of only one or two extreme components (Albert and Harris, 1987; Strike, 1991). Finally, ellipsoidal regions are unable to detect component-wise outliers. In other words, whenever patients are diagnosed as non-healthy based on an MRR, no conclusion can be drawn on which specific analyte/s have caused the positive result. For this reason, Wellek (2011) notes that MRRs "have only a marginal role in the practice of clinical chemistry and laboratory medicine."

To address the above difficulties associated with ellipsoidal reference regions, this paper aims to derive rectangular reference regions, which are easily interpretable regions that can detect the outlyingness of specific analytes. In view of the fact that laboratory test results are typically skewed (or at least not normally distributed), we shall derive such regions under a nonparametric framework. Previous work on nonparametric reference regions includes that of Wellek (2011) and Young and Mathew (2020). The work of Wellek (2011) includes both parametric and nonparametric estimation of rectangular reference regions.

Rectangular prediction regions are derived in Young and Mathew (2020); however, fairly large sample sizes are required for meeting the coverage probability requirement.

In the present investigation, we aim to develop rectangular nonparametric prediction regions to be used as reference regions. The methodologies described in this study are based on either transforming the marginal data using a Box-Cox transformation, or estimating the marginal densities through kernel density estimation. The accuracy of these approaches will be assessed by reporting the relevant coverage probabilities. We investigate the performance of the proposed methodologies using sample sizes starting from $n = 50$.

1.1. Rectangular prediction regions

We now define the criterion to be used in obtaining the rectangular nonparametric reference region. Our goal is to find a rectangular reference region of the form (1)

$$[c_1, d_1] \times [c_2, d_2] \times \cdots \times [c_p, d_p]\,, \tag{1}$$

subject to the prediction region criterion in (2)

$$P\left(\bigcap_{i=1}^{p} \{X_i \in [c_i, d_i]\}\right) = 1 - \alpha. \tag{2}$$

It should be clear that the set of intervals $[c_i, d_i]$, $i = 1, 2, \cdots, p$, satisfying the above requirement is not unique. Nevertheless, it is to be expected that each marginal interval $[c_i, d_i]$, $i = 1, 2, \cdots, p$, can be appropriately specified if we know the marginal distributions. In the absence of any information on the marginal distributions, we shall explore two options. The first option is to apply separate Box-Cox transformations to each set of marginal data, so that each marginal distribution is approximately normal. We can now specify a common prediction factor on the transformed scale, which can be estimated via a nonparametric bootstrap subject to the requirement in (2). Details of this appear in Section 2. The second approach, described in Section 3, uses the kernel density estimate (KDE) of the marginal densities, which also leads to a common prediction factor. In the same section, we also extend the KDE idea in order to construct mixed-sided prediction regions. These are regions where some variables have an upper prediction limits and the rest have two-sided prediction intervals. Section 4 gives numerical results on estimated coverage probabilities in order to assess the accuracy of the proposed methodologies and illustrates the methodologies through a real-life example. Section 5 gives some brief concluding remarks.

2. Nonparametric prediction regions using the Box-Cox transformation

Our first strategy to deal with the problem of computing nonparametric rectangular prediction regions is to transform the marginal data so that it has a normal distribution, approximately. Ichihara and Boyd (2010) note that "Since almost all distributions of laboratory test results are non-Gaussian, it is essential to convert these to a Gaussian distribution." They investigate transformation to normality using the Box-Cox transformation (Box and Cox, 1964) and also a modified Box-Cox formula introduced by Ichihara and Kawai (1997). The International Federation of Clinical Chemisty (IFCC) Expert Committee on Reference Intervals has actually recommended the Box-Cox transformation to normality for the purpose of computing reference intervals (Solberg, 1987). In other words, the idea of using the

Box-Cox transformation in the context of computing reference intervals is already mentioned in the literature, but only in the univariate context.

We shall employ the Box-Cox transformation to develop rectangular reference regions in a nonparametric setup. The prediction factor is computed based on the transformed data, under the assumption that the data are approximately normal. Here we want to point out that we shall apply the Box-Cox transformation to the sample from each univariate marginal distribution. In other words, the transformation is univariate, not multivariate. As will be seen shortly, normality will not be fully utilized when we derive the prediction factor, since we will be employing a nonparametric bootstrap procedure. However, normality is perhaps necessary to justify the use of a common prediction factor.

Suppose that the data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ consist of a random sample coming from an unknown multivariate distribution with nonnegative support, where

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$$

is a p-variate vector, $i = 1, 2, \ldots, n$. For each fixed $j = 1, 2, \ldots, p$, $X_{1j}, X_{2j} \ldots, X_{nj}$, form a random sample from the univariate marginal distribution of the jth component. We assume that these random variables can be transformed as

$$Y_{ij} = g_j\left(X_{ij}\right) \tag{3}$$

so that their distribution is approximately normal. The transformation could be different for the different components. Once such a transformation has been identified, we can then construct prediction regions for the transformed data Y_{ij}, $i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$. Since the transformed data are assumed to be approximately normal, we restrict the two-sided prediction region to be of the symmetric form

$$\bar{Y}_j \pm \kappa\sqrt{S_{y,jj}} \tag{4}$$

for $j = 1, 2, \ldots, p$, and our goal is to estimate κ. Here $\bar{\boldsymbol{Y}} = \left(\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_p\right)'$ is the sample mean vector, and $S_{y,jj}$ is the jth diagonal element of the sample covariance matrix among the transformed sample values $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$, $i = 1, 2, \ldots, n$.

Introduced by Box and Cox (1964), the Box-Cox transformation, as it has come to be called in the statistical literature, is a well-known method to transform skewed data to normality. For a random variable X that assumes positive values, the Box-Cox transformed quantity, say Y, takes the form

$$Y = \begin{cases} \frac{X^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases} \tag{5}$$

where $\lambda > 1$ for negatively skewed data and $\lambda < 1$ for positively skewed data. The value of λ is to be estimated using the data on X. In this study, λ is estimated through maximum likelihood.

Once a λ_j has been estimated based on $X_{1j}, X_{2j}, \ldots, X_{nj}$, for each fixed $j = 1, 2, \ldots, p$, we shall choose the form of $g_j(\cdot)$ in (3) to be

$$Y_{ij} = g_j\left(X_{ij}\right) = \begin{cases} X_{ij}^{\lambda_j}, & \lambda_j \neq 0 \\ \log X_{ij}, & \lambda_j = 0, \end{cases} \tag{6}$$

instead of (5). Since X_{ij} assumes positive values, if $\lambda_j \neq 0$ it follows that Y_{ij} in (6) is always positive. Therefore, the back-transformed value Y_{ij}^{1/λ_j} is always defined. This is not always the case with (5), where the back-transformed value is $(\lambda Y + 1)^{1/\lambda}$, which can be undefined if $\lambda Y + 1 < 0$. For this reason, once the value of λ_j is identified, the power transformation in (6) will be adopted. The next goal is to estimate κ to form prediction regions of the form (4).

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ be the future observation to be predicted, and let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)' = \left(X_1^{\lambda_1}, X_2^{\lambda_2}, \ldots, X_p^{\lambda_p}\right)'$ be its transformed version. We shall find the value of κ that satisfies

$$P\left(Y_j \in \bar{Y}_j \pm \kappa\sqrt{S_{y,jj}} \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha.$$

That is,

$$P\left(\left|\frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}}\right| \leq \kappa \quad \forall\, j = 1, 2, \ldots, p\right) = 1 - \alpha.$$

Equivalently,

$$P\left(\max_{1\leq j\leq p}\left|\frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}}\right| \leq \kappa\right) = 1 - \alpha. \tag{7}$$

The choice of a common κ is justified in view of the approximate normality of the marginal components of the $\boldsymbol{Y}$. The statement (7) facilitates the estimation of κ via a nonparametric bootstrap, by sampling with replacement from the collection $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$. Since the λs are unknown parameters, they are also estimated in each bootstrap sample. Algorithm 1 gives the procedure to estimate κ; we shall denote the estimate by k.

2.1. Remarks on back-transforming the data

Caution must be taken in Step 10 of Algorithm 1, where the prediction region is transformed back to the original scale. First of all, since for $\lambda_j \neq 0$, $Y_{ij} = X_{ij}^{\lambda_j}$ where X_{ij}s are positive, the prediction regions in the Y_{ij} scale should contain only nonnegative limits. In some instances, however, the lower limit of an interval $\bar{Y}_j - k\sqrt{S_{jj,y}}$ could be negative. In such a case, we recommend that the lower limit just be changed to 0.

It is possible for $\hat{\lambda}_j$ to be negative. Whenever this happens concurrently with a negative lower limit, then the quantity

$$\left(\bar{Y}_j - k\sqrt{S_{jj,y}}\right)^{1/\hat{\lambda}_j} \tag{10}$$

is undefined, even when the lower limit is changed to 0. The case where both lower limit and λ_j are negative occurs rarely in the simulations, but it occurs more often when the sample size is small than when large, presumably because λ_j cannot be estimated accurately from a small sample. Such a phenomenon never occurred in the simulations included in this paper, but the authors have seen it occur when the sample size is small (such as when $n = 30$). Nonetheless, such an occurrence is still highly unlikely when p is small (*e.g.*, 2 or 3). Since

Algorithm 1 Nonparametric prediction regions using the Box-Cox transformation

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_n$ be the data and write $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$.

2. For each $j = 1, 2, \ldots, p$, estimate the Box-Cox transformation parameter λ_j for the observations $X_{1j}, X_{2j}, \ldots, X_{nj}$. Let $\hat{\lambda}_j$ be the estimated value of λ_j.

3. For each $\hat{\lambda}_j$ in Step 2, compute

$$Y_{ij} = \begin{cases} X_{ij}^{\hat{\lambda}_j}, & \hat{\lambda}_j \neq 0 \\ \log X_{ij}, & \hat{\lambda}_j = 0 \end{cases}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p.$$

   Define $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})'$, $i = 1, 2, \ldots, n$.

4. Take B random samples with replacement of size $n+1$ from the collection $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, call these $\boldsymbol{X}_{1b}^*, \boldsymbol{X}_{2b}^*, \ldots \boldsymbol{X}_{nb}^*, \boldsymbol{X}_b^*$, where $b = 1, 2, \ldots, B$. Write $\boldsymbol{X}_b^* = \left(X_{1b}^*, X_{2b}^*, \ldots, X_{pb}^*\right)'$ and $\boldsymbol{X}_{ib}^* = \left(X_{i1b}^*, X_{i2b}^*, \ldots, X_{ipb}^*\right)'$, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$.

5. For each bootstrap sample in Step 4, estimate the transformation parameter for the observations in the jth column of the data matrix, and denote this estimate by $\hat{\lambda}_{jb}^*$.

6. For each $\hat{\lambda}_{jb}^*$ in Step 5, compute

$$Y_{ijb}^* = \begin{cases} \left(X_{ijb}^*\right)^{\hat{\lambda}_{jb}^*}, & \hat{\lambda}_{jb}^* \neq 0 \\ \log X_{ijb}^*, & \hat{\lambda}_{jb}^* = 0 \end{cases} \quad \text{and} \quad Y_{jb}^* = \begin{cases} \left(X_{jb}^*\right)^{\hat{\lambda}_{jb}^*}, & \hat{\lambda}_{jb}^* \neq 0 \\ \log X_{jb}^*, & \hat{\lambda}_{jb}^* = 0 \end{cases},$$

   where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, $b = 1, 2, \ldots, B$. Write $\boldsymbol{Y}_{ib}^* = \left(Y_{i1b}^*, Y_{i2b}^*, \ldots, Y_{ipb}^*\right)'$ and $\boldsymbol{Y}_b^* = \left(Y_{1b}^*, Y_{2b}^*, \ldots, Y_{pb}^*\right)'$.

7. Compute $k_b^* = \max\limits_{1 \leq j \leq p} \left| \frac{Y_{jb}^* - \bar{Y}_{jb}^*}{\sqrt{S_{b,jj}^*}} \right|$, $b = 1, 2, \ldots, B$, where $\bar{Y}_{jb}^*$ is the jth element in the sample mean of $\boldsymbol{Y}_{1b}^*, \boldsymbol{Y}_{2b}^*, \ldots, \boldsymbol{Y}_{nb}^*$; Y_{jb}^* is the jth component of $\boldsymbol{Y}_b^*$; and $S_{b,jj}^*$ is the jth diagonal element in the sample covariance matrix of $\boldsymbol{Y}_{1b}^*, \boldsymbol{Y}_{2b}^*, \ldots \boldsymbol{Y}_{nb}^*$.

8. Compute k as the $(1-\alpha)$-quantile of $k_1^*, k_2^*, \ldots, k_B^*$.

9. The prediction region for the transformed data is given by

$$\left[\bar{Y}_1 \pm k\sqrt{S_{y,11}}\right] \times \left[\bar{Y}_2 \pm k\sqrt{S_{y,22}}\right] \times \cdots \times \left[\bar{Y}_p \pm k\sqrt{S_{y,pp}}\right], \tag{8}$$

   where $\bar{Y}_j$ and $S_{y,jj}$ are respectively the jth element and jth diagonal element in the sample mean vector and sample covariance matrix of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$.

10. Finally, the prediction region for the original data is given by

$$\left[\bar{Y}_1 \pm k\sqrt{S_{y,11}}\right]^{1/\hat{\lambda}_1} \times \left[\bar{Y}_2 \pm k\sqrt{S_{y,22}}\right]^{1/\hat{\lambda}_2} \times \cdots \times \left[\bar{Y}_p \pm k\sqrt{S_{y,pp}}\right]^{1/\hat{\lambda}_p}, \tag{9}$$

    where we define $[a, b]^{1/q} = \left[a^{1/q}, b^{1/q}\right]$ if $q > 0$, $\left[b^{1/q}, a^{1/q}\right]$ if $q < 0$, and $\left[e^a, e^b\right]$ if $q = 0$.

(10) is undefined when we run into this situation, we can address this problem by redefining the interval corresponding to the particular component of $\boldsymbol{X}$, say the jth. For example, it is reasonable to redefine it as

$$\left[\left(\bar{Y}_j + k\sqrt{S_{jj,y}}\right)^{1/\hat{\lambda}_j}, \infty\right).$$

We therefore end up with a prediction region that is a mix of one and two-sided intervals (called a mixed-sided reference region in Section 5). It is not clear how we can compare expected volumes in such a scenario.

The reference region given in (9) can also exhibit erratic behaviors, for example, when at least one of $\bar{Y}_j - k\sqrt{S_{jj,y}}$ and $\bar{Y}_j + k\sqrt{S_{jj,y}}$ is very close to 0 and $\hat{\lambda}_j < 0$. Ideally, having a large sample size is the best remedy to avoid the above undesirable behaviors of reference regions. When this is not possible (for example, due to cost considerations), we would like to recommend the following: instead of back-transforming to the original scale, do a "forward-transform" of the future observation to see if it falls inside the reference range in the transformed scale. That is, whenever $\hat{\lambda}_j \neq 0$, consider the transformation

$$(X_1, X_2, \ldots, X_p)' \longmapsto \left(X_1^{\hat{\lambda}_1}, X_2^{\hat{\lambda}_2}, \ldots, X_p^{\hat{\lambda}_p}\right)'$$

and when $\hat{\lambda}_j = 0$, use $X_j \longmapsto \log X_j$, and then use the region in (8) as the reference region, instead of (9). We emphasize that the limits in (8) are always defined.

2.2. One-sided prediction regions

Modifications of Algorithm 1 that are necessary to compute one-sided regions are straightforward whenever we do not run into problems involving the sign of λ_j or the lower limit, or when we choose to adopt the reference region in the transformed scale. To compute a one-sided upper prediction region, we first estimate the prediction factor κ that satisfies

$$\begin{aligned} P\left(Y_j \leq \bar{Y}_j + \kappa\sqrt{S_{y,jj}} \quad \forall\, j = 1, 2, \ldots, p\right) &= 1 - \alpha \\ \iff P\left(\max_{1\leq j\leq p} \frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}} \leq \kappa\right) &= 1 - \alpha. \end{aligned} \tag{11}$$

Condition (11) implies that the modification is to be done in Step 7 of Algorithm 1, in which the quantity k_b will be redefined as

$$k_b = \max_{1\leq j\leq p} \frac{Y_{jb}^* - \bar{Y}_{jb}^*}{\sqrt{S_{jj,b}^*}}.$$

For a $(1-\alpha)$-one-sided upper prediction region, we can take k to be the $(1-\alpha)$-quantile of $k_1, k_2, \ldots, k_B$, and the prediction region in the $\boldsymbol{Y}$-scale is given by

$$\left(-\infty, \bar{Y}_1 + k\sqrt{S_{11,y}}\right] \times \cdots \times \left(-\infty, \bar{Y}_p + k\sqrt{S_{pp,y}}\right].$$

If $\hat{\lambda}_j > 0$ for all $j = 1, \ldots, p$, the prediction region in the $\boldsymbol{X}$-scale can then be defined as

$$\left(-\infty, \left(\bar{Y}_1 + k\sqrt{S_{11,y}}\right)^{1/\hat{\lambda}_1}\right] \times \cdots \times \left(-\infty, \left(\bar{Y}_p + k\sqrt{S_{pp,y}}\right)^{1/\hat{\lambda}_p}\right].$$

When $\hat{\lambda}_j < 0$, the corresponding univariate reference limit simply becomes a lower limit instead of an upper limit. Finally, if $\hat{\lambda}_j = 0$, then the corresponding interval becomes

$$\left(-\infty, \exp\left(\bar{Y}_j + k\sqrt{S_{jj,y}}\right)\right].$$

Similarly, for a $(1-\alpha)$-one-sided lower prediction region, we note that the prediction factor should satisfy

$$\begin{aligned} P\left(\bar{Y}_j + \kappa\sqrt{S_{y,jj}} \leq Y_j \quad \forall\, j = 1, 2, \ldots, p\right) &= 1 - \alpha \\ \iff P\left(\min_{1\leq j\leq p} \frac{Y_j - \bar{Y}_j}{\sqrt{S_{y,jj}}} < \kappa\right) &= \alpha. \end{aligned}$$

Thus we change the definition of k_b in Step 7 of Algorithm 1 to be

$$k_b = \min_{1\leq j\leq p} \frac{Y^*_{jb} - \bar{Y}^*_{jb}}{\sqrt{S^*_{jj,b}}},$$

and then we take the estimated prediction factor k to be the α-quantile of $k_1, k_2 \ldots, k_B$. The prediction region in the $\boldsymbol{Y}$-scale is given by

$$\left[\bar{Y}_1 + k\sqrt{S_{11,y}}, \infty\right) \times \cdots \times \left[\bar{Y}_p + k\sqrt{S_{pp,y}}, \infty\right).$$

If $\hat{\lambda}_j > 0$ for all $j = 1, 2, \ldots, p$, then in the $\boldsymbol{X}$-scale the $(1-\alpha)$-one-sided lower prediction region is given by

$$\left[\left(\bar{Y}_1 + k\sqrt{S_{11,y}}\right)^{1/\hat{\lambda}_1}, \infty\right) \times \cdots \times \left[\left(\bar{Y}_p + k\sqrt{S_{pp,y}}\right)^{1/\hat{\lambda}_p}, \infty\right),$$

and we deal with a zero or negative $\hat{\lambda}_j$ analogously.

3. Nonparametric prediction regions using kernel density estimation

We shall now explore an alternative approach to construct nonparametric rectangular prediction regions. The approach consists of obtaining a kernel density estimate of the unknown probability density function, and then use the probability integral transform based on the estimated density function to derive a rectangular prediction region. Such an approach provides us with a justification for using common prediction limits for each marginal component in the transformed scale, quite analogous to the use of a common prediction factor κ in the transformed scale in Section 2. An inverse transformation can then be used to obtain the required prediction limits in the original scale. We shall now present the details.

3.1. One-sided upper and lower prediction regions

We shall first present our approach for computing upper prediction limits; the case of lower prediction limits can be handled similarly. The case of two-sided prediction regions will be explained later. Suppose we want to compute upper one-sided prediction limits for the components of the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$, using the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the distribution of $\boldsymbol{X}$. Thus we have to use the data to estimate $\gamma_1, \gamma_2, \ldots, \gamma_p$ that satisfy

$$P\left(X_1 \leq \gamma_1, X_2 \leq \gamma_2, \ldots, X_p \leq \gamma_p\right) = 1 - \alpha. \tag{12}$$

If we can find one-to-one transformations $Y_j = g_j\left(X_j\right), j = 1, 2, \ldots, p$, so that $Y_1, Y_2, \ldots, Y_p$ are identically distributed random variables, then it makes sense to have a common upper limit ζ that satisfies

$$P\left(Y_1 \leq \zeta, Y_2 \leq \zeta, \ldots, Y_p \leq \zeta\right) = P\left(\max_{1 \leq j \leq p} Y_j \leq \zeta\right) = 1 - \alpha. \tag{13}$$

If the distribution functions $F_j\left(x\right)$ of all the X_js, $j = 1, 2, \ldots, p$, were completely known, then an obvious transformation that one can use is $Y_j = g_j\left(X_j\right) = F_j\left(X_j\right), j = 1, 2, \ldots, p$. Clearly, the transformed random variables $F_j\left(X_j\right), j = 1, 2, \ldots, p$, are identically distributed as $U\left(0, 1\right)$ random variables. However, since the $F_j\left(X_j\right), j = 1, 2, \ldots, p$ are unknown, the idea is to estimate them marginally using *kernel density estimation* (KDE). Call the estimated distribution functions $\hat{F}_j, j = 1, 2, \ldots, p$, and let z be an estimate of ζ satisfying (13), where $Y_j = \hat{F}_j$, $j = 1, 2, \ldots, p$. We can now obtain estimates of the upper limits γ_j, $j = 1, 2, \ldots, p$, satisfying (12) as $\hat{\gamma}_j = c_j = \hat{F}_j^{-1}\left(z\right), j = 1, 2, \ldots, p$. Obviously, since the upper limits so computed are estimates obtained from the data, we do not expect (12) to hold exactly. We shall explore this shortly based on numerical results.

Kernel density estimation is a nonparametric statistical method used to estimate an unknown PDF or CDF. Whenever we have a random sample $X_1, X_2, \ldots, X_n$ from a continuous univariate distribution with an unknown density function $f\left(x\right)$, the kernel density estimate of $f\left(x\right)$, say $\hat{f}\left(x\right)$, is given by:

$$\hat{f}\left(x\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

where h denotes the bandwidth, which can be thought of as a smoothing parameter. The *kernel function* $K\left(\cdot\right)$ satisfies $K\left(\cdot\right) \geq 0$ and

$$\int_{-\infty}^{\infty} K\left(t\right) dt = 1.$$

In this study, we shall use the Gaussian kernel $K\left(t\right) = \phi\left(t\right)$, where $\phi\left(\cdot\right)$ is the standard normal density function. Our interest is in estimating the CDF, say $F\left(\cdot\right)$. The corresponding

estimate, say $\hat{F}(t)$, is given by:

$$\hat{F}(t) = \int_{-\infty}^{t} \hat{f}(u)\,du = \int_{-\infty}^{t} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{u-X_i}{h}\right)du$$
$$= \int_{-\infty}^{t} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\phi\left(\frac{u-X_i}{h}\right)du = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{t-X_i}{h}\right), \tag{14}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. Moreover, our choice for the bandwidth h will be *Silverman's Rule of Thumb*, given by

$$h = 0.9\min(S, IQR/1.34)\,n^{-1/5}. \tag{15}$$

This bandwidth appears to be the preferred choice whenever $K(\cdot)$ is chosen to be the Gaussian kernel (Silverman, 1986).

We point out that since $\hat{F}_j$ is a one-to-one function, $\hat{F}_j^{-1}(z)$ always exists and hence the resulting region based on KDE always has defined limits, unlike the case for Box-Cox transformation. Algorithm 2 gives the steps necessary to compute nonparametric one-sided upper and lower prediction limits using KDE. In order to understand Step 5 in the algorithm, we recall that when we have a random sample of univariate observations $X_1, X_2, \ldots, X_n$, and we want to construct a $100(1-\alpha)\%$ nonparametric upper prediction limit for a future observation, the upper prediction limit is given by the rth order statistic $X_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$. Similarly, to construct a $100(1-\alpha)\%$ lower prediction limit for a future observation, the required limit is given by $X_{(r)}$, where $r = \lfloor \alpha(n+1) \rfloor$ (Meeker et al., 2017).

3.2. Two-sided prediction regions

We shall now develop the methodology to compute nonparametric two-sided prediction regions using kernel density estimation. In constrast to (12), we need to estimate $\gamma_{11}, \gamma_{21}, \ldots, \gamma_{p1}$ and $\gamma_{12}, \gamma_{22}, \ldots, \gamma_{p2}$ that satisfy the condition:

$$P(\gamma_{11} \le X_1 \le \gamma_{12}, \gamma_{21} \le X_2 \le \gamma_{22}, \ldots, \gamma_{p1} \le X_p \le \gamma_{p2}) = 1-\alpha. \tag{16}$$

Similar to the development of one-sided prediction regions, if the distribution functions $F_j(x)$, $j = 1, 2, \ldots, p$, were completely known, then we can use the transformation $Y_j = F_j(X_j)$ and find the common upper and lower prediction limits ζ_1 and ζ_2 satisfying

$$P(\zeta_1 \le Y_j \le \zeta_2, j = 1, 2, \ldots, p) = 1-\alpha. \tag{17}$$

Since Y_j, $j = 1, 2, \ldots, p$ are all $U(0,1)$ random variables, and since the $U(0,1)$ distribution is symmetric, it makes sense to set $\zeta_1 = 1-\zeta_2$. Thus, write (17) as

$$P(1-\zeta_2 \le Y_j \le \zeta_2, j = 1, 2, \ldots, p) = P(\max\{Y_j, 1-Y_j\} \le \zeta_2, j = 1, 2, \ldots, p) \tag{18}$$
$$= P\left(\max_{1\le j\le p} \max\{Y_j, 1-Y_j\} \le \zeta_2\right) = 1-\alpha. \tag{19}$$

As in the one-sided case, since $F_j(x), j = 1, 2, \ldots, p$, are unknown, we estimate them via KDE. Let $\hat{F}_j$, $j = 1, 2, \ldots, p$, be the estimated CDFs and let z be the estimate of ζ_2 that

Algorithm 2 Nonparametric one-sided upper (lower) prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$ is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate the distribution function of the jth component using KDE (see (14)).The data used to estimate F_j are $X_{1j}, X_{2j}, \ldots, X_{nj}$. Call the estimated CDF $\hat{F}_j$.
3. Compute $Y_{ij} = \hat{F}_j(X_{ij})$ for each X_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
4. Compute $z_i = \max\limits_{1 \le j \le p} Y_{ij}$ $\left(z_i = \min\limits_{1 \le j \le p} Y_{ij}\right)$, for each $i = 1, 2, \ldots, n$.
5. Compute the nonparametric $(1-\alpha)$ upper (lower) prediction limit of the $z_1, z_2, \ldots, z_n$. Denote this upper (lower) limit by z; thus $z = z_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$ $(r = \lfloor \alpha(n+1) \rfloor)$.
6. Now compute for $c_j = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$.
7. The $(1-\alpha)$-nonparametric upper (lower) prediction region is given by

$$(-\infty, c_1] \times (-\infty, c_2] \times \cdots \times (-\infty, c_p] \qquad ([c_1, \infty) \times [c_2, \infty) \times \cdots \times [c_p, \infty)).$$

satisfies (19). We can estimate the prediction limits in (16) as $\hat{\gamma}_{j1} = \hat{F}_j^{-1}(1-z)$ and $\hat{\gamma}_{j2} = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$. Algorithm 3 gives the procedure to compute the nonparametric two-sided prediction region using KDE. Step 6 of Algorithm 3 is motivated by the fact that ζ_2 has been expressed in (19) as the $(1-\alpha)$-quantile of the random variable $\max\limits_{1 \le j \le p} \max\{Y_j, 1 - Y_j\}$.

Here we would like to make an important remark concerning Step 6 in Algorithm 2. The computation of the order statistic-based nonparametric upper prediction limit in Step 6 requires the independence of $z_1, z_2, \ldots, z_n$. However, these quantities are not independent since the $\hat{F}_j$s are not independent. In formulating the algorithm, we have simply ignored this. The estimated coverage probabilities that we shall shortly report will indicate the effect of ignoring the lack of independence among $z_1, z_2, \ldots, z_n$.

3.3. Mixed-sided nonparametric prediction regions using kernel density estimation

In many applications, we are interested in prediction regions that are a combination of one-sided and two-sided intervals, since some variables may require two-sided reference limits while others are appropriately bounded by one-sided reference limits. We shall refer to such regions as *mixed-sided prediction regions*. For example, we may be interested in finding the region $[c_1, d_1] \times (-\infty, d_2]$ such that

$$P(c_1 \le X_1 \le d_1, X_2 \le d_2) = 1 - \alpha.$$

Algorithm 3 Nonparametric two-sided prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$, is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate the distribution function of the jth component using KDE (see (14)).The data used to estimate F_j are $X_{1j}, X_{2j}, \ldots, X_{nj}$. Call the estimated CDF $\hat{F}_j$.
3. Compute $Y_{ij} = \hat{F}_j(X_{ij})$ for each X_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
4. Compute $U_{ij} = \max\{Y_{ij}, 1 - Y_{ij}\}$ for each Y_{ij}, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$.
5. Compute $z_i = \max\limits_{1 \le j \le p} U_{ij}$, for each $i = 1, 2, \ldots, n$.
6. Compute the nonparametric $(1 - \alpha)$ upper prediction limit of the $z_1, z_2, \ldots, z_n$. Denote this upper limit by z; thus $z = z_{(r)}$, where $r = \lceil (1 - \alpha)(n + 1) \rceil$.
7. Compute $c_j = \hat{F}_j^{-1}(1 - z)$ and $d_j = \hat{F}_j^{-1}(z)$, $j = 1, 2, \ldots, p$.
8. The $(1 - \alpha)$-nonparametric two-sided prediction region is given by

$$[c_1, d_1] \times [c_2, d_2] \times \cdots \times [c_p, d_p].$$

We now take up the problem of computing mixed-sided nonparametric prediction regions. Suppose our data consists of the random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, where each $\boldsymbol{X}_i$ is p-variate. Moreover, let $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ be the observation that we wish to predict and assume that it has the same distribution as the $\boldsymbol{X}_i$s and is independent of them. Without loss of generality, we develop a procedure to compute two-sided prediction limits for the first p_1 components of $\boldsymbol{X}$ and upper prediction limits for the remaining $p - p_1$ components. In doing so, we use a KDE-based approach since this approach generally shows superior performance over the Box-Cox transformation-based approach, as we have seen in Section 4.

Let $F_j(\cdot)$ be the CDF of X_j, $j = 1, 2, \ldots, p$. If we can find scalar quantities u, u', and v, all three being functions of $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, that satisfy

$$\begin{aligned} P_{\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n} \Big(& F_j^{-1}(u') \le X_j \le F_j^{-1}(u), \forall j = 1, \ldots, p_1 \\ & \text{and } X_j \le F_j^{-1}(v), \forall j = p_1 + 1, \ldots, p \Big) = 1 - \alpha, \end{aligned} \tag{20}$$

then the region

$$\left[F_1^{-1}(u'), F_1^{-1}(u)\right] \times \cdots \times \left[F_{p_1}^{-1}(u'), F_{p_1}^{-1}(u)\right] \times \left(-\infty, F_{p_1+1}^{-1}(v)\right] \times \cdots \times \left(-\infty, F_p^{-1}(v)\right] \tag{21}$$

is a $(1 - \alpha)$-mixed-sided nonparametric prediction region for $\boldsymbol{X}$. The condition in (20) is equivalent to

$$\begin{aligned} & P_{\boldsymbol{X}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n} (u' \le F_j(X_j) \le u, \forall j = 1, \ldots, p_1 \\ & \text{and } F_j(X_j) \le v, \forall j = p_1 + 1, \ldots, p) = 1 - \alpha. \end{aligned} \tag{22}$$

Since each $F_j(X_j)$ follows a $U(0,1)$ distribution, we can choose $u' = 1-u$. Furthermore, since infinitely many possible values of u and v can satisfy (22), we shall impose a constraint on u and v so as to arrive at a unique solution. The constraint to be imposed is that the marginal probabilities in (22) should be equal. This amounts to choosing u and v such that $v = 2u-1$. We can see this by observing that if U and V are $U(0,1)$ random variables, then imposing the condition $P(1-u \le U \le u) = P(V \le v)$ implies $v = 2u-1$. Substituting these expressions for u' and v, (22) becomes

$$\begin{aligned} &P_{\boldsymbol{X},\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(1-u \le F_j(X_j) \le u, \forall j = 1,2,\ldots,p_1 \\ &\qquad \text{and } F_j(X_j) \le 2u-1, \forall j = p_1+1,\ldots,p) = 1-\alpha. \end{aligned} \tag{23}$$

Since

$$\begin{aligned} &1-u \le F_j(X_j) \le u, \forall j = 1,2,\ldots,p_1 \\ &\iff \max\left\{\max_{1\le j\le p_1} F_j(X_j), \max_{1\le j\le p_1}(1-F_j(X_j))\right\} \le u, \end{aligned} \tag{24}$$

$$F_j(X_j) \le 2u-1, \forall j = p_1+1,\ldots,p \iff \max_{p_1+1\le j\le p}\left\{\frac{1+F_j(X_j)}{2}\right\} \le u \tag{25}$$

then we can write (23) as

$$P_{\boldsymbol{X},\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n}(\max\{h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), h_3(\boldsymbol{X})\} \le u) = 1-\alpha, \tag{26}$$

where

$$\begin{aligned} h_1(\boldsymbol{X}) &= \max_{1\le j\le p_1} F_j(X_j) \\ h_2(\boldsymbol{X}) &= \max_{1\le j\le p_1}(1-F_j(X_j)) \\ h_3(\boldsymbol{X}) &= \max_{p_1+1\le j\le p}\left\{\frac{1+F_j(X_j)}{2}\right\}. \end{aligned}$$

From (26) we can conclude that u is a $(1-\alpha)$-upper prediction limit of

$$\max\{h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), h_3(\boldsymbol{X})\}.$$

Since the distribution functions $F_j(\cdot)$ are unknown, we estimate them using KDE. Algorithm 4 gives the steps to compute the mixed-sided nonparametric prediction region using KDE.

3.4. KDE with logarithmic transformation

Studies such as Geenens and Wang (2016) and Jones et al. (2018) suggest that whenever the density is supported on the set of positive real numbers, we should first apply a logarithmic transformation on the observations before estimating the density function. Geenens and Wang (2016) argue that the KDE approach to estimate the density of a positive random variable is inadequate due to the boundary bias problem and the fact that such a density might have a long right tail. Charpentier and Flachaire (2014) also mention that

Algorithm 4 Mixed-sided nonparametric prediction regions based on KDE

1. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be the random sample, where each $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$ is a $(p \times 1)$ column vector of measurements from the ith subject.
2. For each $j = 1, 2, \ldots, p$, estimate $\hat{F}_j$ via KDE using $X_{1j}, X_{2j}, \ldots, X_{nj}$.
3. Compute $Y_{ij} = \hat{F}_j(X_{ij})$ for all $j = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$.
4. Compute $u_i = \max\left\{\max_{1\le j\le p_1} Y_{ij}, \max_{1\le j\le p_1}(1 - Y_{ij}), \max_{p_1+1\le j\le p}\left(\frac{1+Y_{ij}}{2}\right)\right\}$, for each $i = 1, 2, \ldots, n$.
5. Compute the nonparametric $(1-\alpha)$-upper prediction limit of the $u_1, u_2, \ldots, u_n$. Denote this upper limit by u; thus $u = u_{(r)}$, where $r = \lceil (1-\alpha)(n+1) \rceil$.
6. Compute $c_j = \hat{F}_j^{-1}(1-u)$ and $d_j = \hat{F}_j^{-1}(u)$, $j = 1, 2, \ldots, p_1$; and $d_j = \hat{F}_j^{-1}(2u-1)$, $j = p_1 + 1, \ldots, p$.
7. The $(1-\alpha)$-mixed-sided nonparametric prediction region is given by $[c_1, d_1] \times \cdots \times [c_{p_1}, d_{p_1}] \times (-\infty, d_{p_1+1}] \times \cdots \times (-\infty, d_p]$.

doing a preliminary logarithmic transformation before applying KDE can provide a better fit for heavy-tailed densities.

To apply this idea to the proposed KDE-based procedure, we can modify Algorithm 2 (or Algorithm 3 for the two-sided case) by taking the logarithm component-wise of each $\boldsymbol{X}_i$, $i = 1, 2, \ldots, n$ in Step 1 before proceeding to the other steps, and then exponentiating each limit in Step 6 of Algorithm 2 (or Step 7 Algorithm 3) to get the reference limits in the original scale. We shall refer to this procedure as the *KDE with log transform* procedure.

4. Numerical results and an example

In order to evaluate the performance of the proposed procedures to construct prediction regions in the nonparametric case, simulations will be carried out to estimate coverage probabilities and expected volumes for data generated from a multivariate lognormal distribution with mean vector in the logarithmic scale of $\mathbf{0}$, and covariance matrix in the logarithmic scale $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_p + \rho \mathbf{1}_p \mathbf{1}_p'$ where $\rho = 0.5$, and $\mathbf{1}_p$ is the $(p \times 1)$ column vector of 1s and $\boldsymbol{I}_p$ is the $(p \times p)$ identity matrix. We use the R package **compositions** of van den Boogaart and Tolosana-Delgado (2008) to generate samples from the multivariate lognormal distribution. We examine the performance for sample sizes $n = 50$, 100, and 200 and refer to these as small, moderate and large sample sizes. We also use dimensions $p = 2$ and 3 since most applications of MRRs involve only at most three analytes. For the Box-Cox transformation-based procedure, we use $B = 500$ bootstrap samples to estimate the prediction factor. For the KDE-based procedure, the inverse function in Step 6 of Algorithm 2 and Step 7 in Algorithm 3 and all other occurrences of the inverse function in this study are computed using the R package **GoFKernel** of Pavia (2015). The coverage probabilities are

based on 5000 simulated samples, and the results are given in Table 1.

From the numerical results in Table 1, we can see that the coverage probabilities of the proposed methodologies are generally close to the nominal level of 0.95, even for a sample of size $n = 50$. Furthermore, it seems that the KDE-based procedure is slightly more accurate than the Box-Cox transformation-based procedure. It is worth comparing our sample sizes with those of Young and Mathew (2020), who also propose nonparametric reference regions. Young and Mathew (2020) examine the performance of their procedure only for sample sizes 300 and 1000. We note that for these dimensions, the coverage probabilities for $n = 100$ in Table 1 are already comparable to Young and Mathew's results for $n = 300$.

Table 2 gives the expected volumes obtained from the proposed methodologies. Table 2 shows that for both the Box-Cox transformation-based procedure and the KDE-based procedure, the expected volume decreases with the sample size. We can see that the KDE-based prediction regions have smaller expected volumes than the Box-Cox transformation-based prediction regions. This implies that the KDE-based procedure results in better precision in estimating the prediction region. On the basis of the results in Tables 1 and 2, the KDE-based procedure has better overall performance than the Box-Cox transformation-based procedure. We note that in computing the expected volume for the Box-Cox transformation-based procedure, we replaced any negative lower limit with zero except when $\hat{\lambda}_j = 0$, in which case a negative lower limit is kept negative. In Table 3 we present the results of the proposed KDE-based one-sided lower and upper prediction regions. The results show accurate coverage, even for small sample sizes.

Table 1: Estimated coverage probabilities of the nonparametric rectangular prediction regions based on Box-Cox transformation and KDE for nominal level = 0.95

	Box-Cox		KDE	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	0.9344	0.9396	0.9582	0.9414
$n = 100$	0.9408	0.9396	0.9472	0.9428
$n = 200$	0.9480	0.9398	0.9428	0.9460

Table 2: Expected volumes of the nonparametric two-sided prediction regions based on Box-Cox transformation and KDE for nominal level = 0.95

	Box-Cox		KDE	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	116.06	2,476.42	95.41	1061.11
$n = 100$	96.67	1,582.25	60.11	823.90
$n = 200$	89.76	1,342.86	53.02	660.93

Table 3: Estimated coverage probabilities of the nonparametric one-sided lower and upper prediction regions based on KDE for nominal level = 0.95

	Lower		Upper	
	$p = 2$	$p = 3$	$p = 2$	$p = 3$
$n = 50$	0.9588	0.9594	0.9518	0.9512
$n = 100$	0.9430	0.9436	0.9526	0.9526
$n = 200$	0.9476	0.9452	0.9464	0.9488

4.1. Comparison of nonparametric procedures when sampling from a highly skewed distribution

We now compare the performances of the KDE-based procedures (both with and without a preliminary log transformation) and the Box-Cox transformation approach to compute prediction regions when we sample from a highly skewed distribution. In the simulations, we generate the data from a gamma distribution with density function given in (27)

$$f(x) = \frac{1}{\lambda^{\eta}\Gamma(\eta)} x^{\eta-1} e^{-x/\lambda}, \quad x \geq 0, \tag{27}$$

with shape parameter $\eta = 0.04$ and scale parameter $\lambda = 1$. This distribution has skewness $2/\sqrt{\eta} = 10$. Table 4 shows the estimated coverage probabilities. We can see that Box-Cox transformation-based procedure results in estimated coverage probabilities very close to 0.95. On the other hand, the usual KDE procedure on the original data is too conservative. While in the previous results, we have seen that the KDE-based procedure outperforms the Box-Cox transformation-based procedure, Table 4 suggests that the Box-Cox-based procedure is more robust to highly skewed distributions, and the KDE-based procedure breaks down under such extreme skewness. Nonetheless, the KDE with log transform procedure rectifies the coverage.

Table 4: Estimated coverage probabilities of the Box-Cox transformation-based and the KDE-based two-sided prediction regions under highly skewed distributions

	Box-Cox transformation	KDE	KDE with log transform
$n = 50$	0.9424	0.9934	0.9668
$n = 100$	0.9500	0.9898	0.9514
$n = 200$	0.9490	0.9936	0.9564

4.2. An example: assessment of liver function

To apply the proposed procedure to compute nonparametric rectangular prediction regions, we use the liver function data from Appendix 4.2 of Harris and Boyd (1995). The measurements are from single blood specimens taken from 596 male medical students during the years 1987-1991 at the University of Virginia. Among the measurements taken from each subject are two liver enzymes: alanine transaminase (ALT) in U/L and aspartate transaminase (AST) in U/L. After the removal of three outliers, the summary statistics are given in Table 5.

Table 5: Summary statistics for measurements on ALT and AST taken from the liver function data of Harris and Boyd (1995)

Analyte	Mean	Median	S.D.	Skewness
ALT	26.97	23.00	17.83	3.63
AST	23.66	22.00	9.51	1.80

Figure 1 shows the density plots for these two analytes. Clearly, both analytes are skewed to the right. Table 5 above also shows that the sample coefficient of skewness is positive. Thus, we use our proposed procedures to compute nonparametric prediction regions. The resulting MRRs using both the Box-Cox transformation and KDE-based approaches are given in Table 6. According to Mayo Clinic (2020), the normal levels for ALT and AST are, respectively, 7-55 and 8-48. Therefore, while the lower limits of the MRR for our proposed procedures agree closely with the lower limits of the reference intervals used in practice, the upper limits are quite different. We hasten to say that these enzymes can be erratically large, in some conditions they can be in the 1000s range (eMedicine Health, 2020). Figure 1 also shows that there are several outlying measurements for ALT, and this could be a factor leading to the unexpectedly high upper reference limit for ALT.

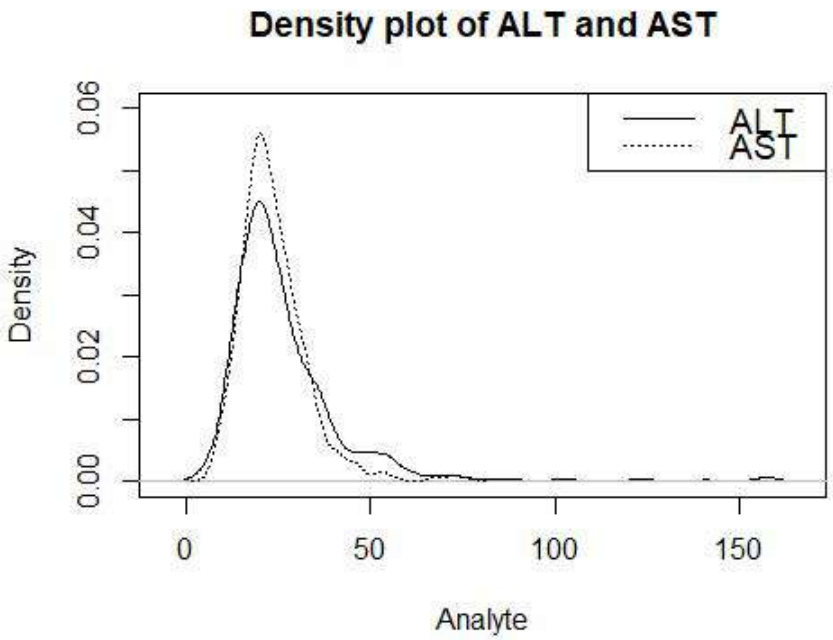


Figure 1: Density plot of ALT and AST

Table 6: MRR for liver function data computed as a two-sided prediction region using Box-Cox transformation and KDE

Analyte	Box-Cox	KDE
ALT	7.2-84.8	7.1-79.1
AST	9.0-54.8	8.8-52.0

4.3. Numerical results on mixed-sided nonparametric prediction regions

We shall now evaluate the performance of our proposed procedure to compute mixed-sided nonparametric prediction regions using KDE, described in Section 3. We generate data from the same distribution used in previous subsections. That is, we estimate coverage probabilities for data generated from a multivariate lognormal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = (1-\rho)\,\boldsymbol{I}_p + \rho \mathbf{1}_p \mathbf{1}_p'$, where $\rho = 0.5$ on the logarithmic scale. The 95% prediction regions will be computed based on 5000 simulated samples. Moreover, we use sample sizes $n = 50$, 100, 200, and $(p, p_1) = (2, 1)$ and $(3, 2)$. Table 7 shows the results. It appears that a sample of size $n = 50$ is sufficient for the proposed methodology to yield accurate results.

Table 7: Estimated coverage probabilities of the mixed-sided nonparametric prediction regions based on KDE for nominal level = 0.95

	$p = 2, p_1 = 1$	$p = 3, p_1 = 2$
$n = 50$	0.9548	0.9448
$n = 100$	0.9510	0.9450
$n = 200$	0.9510	0.9502

5. Discussion

The problem of constructing multivariate reference regions has received proper attention in the literature only recently, except the computation of traditional ellipsoidal prediction regions under the multivariate normality assumption. There are two difficulties associated with the latter region; first, the multivariate normality assumption is not always valid and second, ellipsoidal regions are not appropriate for deciding which among several analytes are outside the normal range. The nonparametric rectangular regions that we have constructed address both of these issues satisfactorily. A different construction of nonparametric rectangular prediction regions is described in Young and Mathew (2020); however, the resulting region exhibits satisfactory coverage probabilities only under relatively large sample sizes. While our work is focused on the computation of rectangular prediction regions only, an important issue is whether a prediction region is appropriate for the purpose for which a reference region is to be used. Some of the recent literature has emphasized tolerance regions, and rectangular tolerance regions are indeed available in the parametric setup of multivariate normality, and in a nonparametric scenario; see Lucagbo and Mathew (2023) and Young and Mathew (2020). Here we do want to note that some laboratory medicine experts have pointed out the role of prediction intervals and regions; see Horn and Pesce (2005), National Committee for Clinical Laboratory Standards (2010), and Trost (2006). In particular, while discussing ellipsoidal regions, Trost (2006, p. 38) notes that "Reference intervals referred to in this document are arguably the closest to prediction intervals since we want exactly 95% of the future observations from reference individuals to fall inside the bounds". We shall not further consider the issue of what criterion is appropriate for the construction of a reference region; this clearly requires input from experts in laboratory medicine.

In our work we have employed two approaches for computing a nonparametric rectangular prediction regions: using the Box-Cox transformation and using kernel density estimation. Estimated coverage probabilities lead us to the conclusion that both approaches are

satisfactory. Based on estimated coverage probabilities and expected volumes, our overall recommendation is the solution based on kernel density estimates. A problem of considerable interest in the context of reference regions is the computation of such regions that are covariate dependent, perhaps using a multivariate regression model. We hope to address this problem in the near future.

Acknowledgements

The authors are grateful to a reviewer for several constructive suggestions that resulted in clarification of some of the ideas and better organization of the results.

References

Albert, A., and Harris, E. K. (1987). *Multivariate Interpretation of Clinical Lab- oratory Data*. Marcel Dekker, Inc., New York.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-252.

Charpentier, A. and Flachaire, E. (2015). Log-transform kernel density estimation of income distribution. (AMSE Working Paper 2015 No. 06). Marseille, France: Aix-Marseille School of Economics.

eMedicineHealth. (2020). Retrieved from: https://www.emedicinehealth.com/liver_blood_tests/article_em.htm#what_should_i_know_about_liver_blood_tests_why_are_they_used. Retrieved April 11, 2020.

Geenens, G. and Wang, C. (2016). Local-likelihood transformation kernel density estimation for positive. arXiv:1602.04862.

Harris, E. K. (1981). Statistical Aspects of Reference Values in Clinical Pathology. In *Progress in Clinical Pathology*, Vol. 8, Mario Stefanini and Ellis S. Benson, eds. Grune & Stratton, New York, pp. 45-66.

Harris, E. K. and Boyd, J. C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. Marcel Dekker, Inc., New York.

Horn, P. S. and Pesce, A. J. (2005). *Reference Intervals: A Users Guide*. American Association for Clinical Chemistry Press, Washington, D. C.

Ichihara, K. and Boyd, J. C. (2010). An appraisal of statistical procedures used in derivation of reference intervals. *Clinical Chemistry and Laboratory Medicine*, **48**, 1537-1551.

Ichihara, K. and Kawai, T. (1996). Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28- P,1992): trial to select reference individuals by results of screening tests and application of maximal likelihood method. *Journal of Clinical Laboratory Analysis*, **10**, 110-117.

Jones, A. T., Nguyen, H. D., and McLachlan, G. J. (2018). Positive data kernel density estimation via the logKDE package for R. arXiv:1804.08365v2 [stat.CO].

Liu, W., Bretz, F., and Cortina-Borja, M. (2021). Reference range: Which statistical intervals to use? *Statistical Methods in Medical Research*, **30**, 523-534.

Lucagbo, M. D., Mathew, T., and Young, D. S. (2023). Rectangular multivariate normal prediction regions for setting reference regions in laboratory medicine, *Journal of Biopharmaceutical Statistics*, **33**, 191-209.

Mayo Clinic. (2020). Retrieved from: https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595. Retrieved April 11, 2020.

Meeker, W. Q., Hahn, G. J., and and Escobar, L. A. (2017). *Statistical Intervals: A Guide for Practitioners and Researchers*, Second Edition. John Wiley & Sons.

National Committee for Clinical Laboratory Standards. (2010). *EP28-A3C: Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline*, - Third Edition. Clinical and Laboratory Standards Institute, Wayne, PA.

Pavia, J. M. (2015). Testing goodness-of-fit with the kernel density estimator: GoFKernel. *Journal of Statistical Software*, Code Snippets, **66**, 1-27.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Solberg, H. E. (1987). Approved recommendation on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *Journal of Clinical Chemistry and Clinical Biochemistry*, **25**, 645-656.

Strike, P. W. (1991). *Statistical Methods in Laboratory Medicine*, Butterworth- Heinemann, Oxford, UK.

Trost, D. C. (2006). Multivariate probability-based detection of drug-induced hepatic signals. *Toxicological Reviews*, **25**, 37-54.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2008). compositions: A unified R package to analyze compositional data. *Computers and Geosciences*, **34**, 320-338.

Wellek, S. (2011). On easily interpretable multivariate reference regions of rectangular shape. *Biometrical Journal*, **53**, 491-511.

Winkel, P., Gaede, P., and Lyngbye, J. (1976). Method for monitoring plasma progesterone concentrations in pregnancy. *Clinical Chemistry*, **22**, 422-428.

Young, D. S. and Mathew, T. (2020). Nonparametric hyperrectangular tolerance and prediction regions for setting multivariate reference regions in laboratory medicine. *Statistical Methods in Medical Research*, **29**, 3569-3585.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 181-195

Two-stage Adaptive Cluster Sampling: A Prediction Approach

Sanghamitra Pal[1] **and Dipika Patra**[2]
[1]*Department of Statistics, West Bengal State University, North 24 Parganas, West Bengal - 700126*
[2]*Department of Statistics, Seth Anandram Jaipuria College, Kolkata, West Bengal -700005*

Received: 05 August 2024; Revised: 16 August 2024; Accepted: 20 August 2024

Abstract

Adaptive Cluster Sampling (ACS) due to Thompson's (1990) is a useful tool to survey rare and clustered population. Salehi and Seber (1997) described a two-stage ACS design that used simple random sampling without replacement (SRSWOR) of primary units and then the ACS of secondary units within each of the selected primary unit. Two variations on this design were proposed in their paper depending on whether networks in secondary units are allowed to cross primary unit boundaries or not.

In executing the adaptive sampling design, it is observed that the collection of information from all neighbouring rare units becomes challenging due to various hazards. Pal and Patra (2021) duly addressed the issue and proposed predictors of the population total considering appropriate superpopulation models with suitable assumptions in single stage ACS. The current work is an attempt to find predictors for two-stage ACS under same situation. To illustrate the findings, a numerical example has been carried out.

Key words: Adaptive cluster sampling; Horvitz-Thompson estimator; Prediction approach; Superpopulation; Two-stage designs; Unequal probability sampling.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Let $U = (1, 2, \ldots, N)$ be a finite population and $y = (y_1, y_2, \ldots, y_N)$ be the variable of interest bearing rarity and clustered characteristics. It is challenging to survey such population through any traditional sampling methods due to the absence of such units in sample with enough number. Thus, the estimation procedures related to the traditional sampling methods such as simple random sampling, stratified sampling may underestimate the population parameters. Thompson's (1990) adaptive cluster sampling reduces the effort to adapt enough number of rare units in the sample and increases the precision. This design has been recently gaining attention because of its greater efficiency. Thompson (1991a) introduced the idea of primary units and secondary units in ACS. The design was further extended by

Corresponding Author: Dipika Patra
Email: dipika.patra1988@gmail.com

Thompson (1991b, 1992). Chaudhuri (2000), Pal and Patra (2021, 2023); Patra and Pal (2023) developed this design under unequal probability sampling designs. A monograph of Seber and Salehi (2013) covers many advents of this design.

In estimating $\tau = \sum_{i=1}^{n} y_i$ by ACS design, an initial sample s of size n is drawn by a probability sampling design and the y-values are observed. Wherever the observed unit satisfies the pre-considered condition of rarity say, $y_i > c$, the uniquely defined neighbouring units (for example - South, North, East, and West) are observed for further detection of rarity. Now, if some of them are found to meet the rarity condition, their neighbouring units are also observed and such procedure continues until a unit is detected with no rarity. It is worth noting that the neighbourhood relation is symmetric. Now, to proceed further in details, one need to know few related terminologies like cluster, edge units, network etc. All neighbouring units corresponding to an initial sampling unit form a cluster. Edge unit is the neighbouring unit that does not satisfy the rarity condition. Thus, each cluster is bounded by edge units. Eliminating all edge units from a cluster, the remaining units that meet the pre-considered rarity condition belong to the network of that particular initial sampling unit. It is also noteworthy, if a unit in s does not satisfy the rarity condition, its network consists of that unit only.

Salehi and Seber (1997), Rocco (2008) and many others, strengthened the literature of Two-stage ACS design. In their proposal, a sample of primary units (PSU) is selected first by simple random sampling without replacement (SRSWOR). Then, an initial sample is taken from secondary units within each selected primary unit, to carry out the ACS design. Surveyors then have two possibilities to stop the adaptively adding procedure of secondary units. Either they can stop at the boundary of PSU (non-overlapping scheme) or allow overlapping into neighbouring PSUs (overlapping scheme). However, in execution stage, surveyors may be unable to observe all the neighbouring secondary units due to hazardous conditions. This deficiency was highlighted in Pal and Patra (2021) for single stage ACS design under unequal probability sampling. Appropriate superpopulation models were adopted there to employ Royall (1970) prediction approach. Implementation of Pal and Patra (2021) approach in two-stage ACS design becomes critical for overlapping scheme. Thus, some modifications are needed following Royall (1976), Valliant *et al.* (2000). In this paper, the main contribution is to develop prediction approach for two-stage ACS-overlapping scheme.

Section 2 elaborately describes the estimation procedure of two-stage ACS design. The next section describes how a superpopulation approach can be used in two-stage ACS to predict the population total or mean in presence of various hazards. Suitable predictors and mean square errors (MSEs) are derived in Section 4. Section 5 illustrates our contribution with a numerical example. Finally, it is concluded in Section 6.

2. Two-stage ACS

Suppose the population U of size N can be partitioned into M primary units of sizes $N_i, i = 1, 2, \ldots, M$ and y_{ij} denotes the y-value of the j^{th} $(j = 1, 2, \ldots, N_i)$ secondary unit of the i^{th} primary unit. Also let $\tau = \sum_{j=1}^{N_i} y_{ij}$ be the sum of the y-value in the i^{th} primary unit and $\tau = \sum_{i=1}^{M} \tau_i$ be the population total. A rarity condition is defined as $y_{ij} > c$ and neighbouring units might be observed only if this rarity condition is satisfied for a given unit.

According to Salehi and Seber (1997)s two stage ACS design, at first, simple random sample (SRS) of size m is drawn from a M primary stage units (PSU). Next, an initial sample s_i of size n_i $(i = 1, 2, \ldots, m)$ is drawn from secondary stage units (SSU) of i^{th} selected PSUs by SRS, such that $n = \sum_{i=1}^{n} n_i$ - total initial sample size. Then, the neighbourhoods may be added adaptively to build up a cluster as well as network.

Now in two-stage ACS, two design-based situations arise. In the first-design, the clusters are truncated at selected PSU's boundaries so that each PSU can be treated separately and it is termed as Non-overlapping scheme. The other one, called overlapping scheme, ignores the PSU boundary so that total population units N can be partitioned into distinct networks. We narrate these two schemes below in details, in the subsections 2.1 - 2.2, with Horvitz-Thompson estimation procedure only. However, Salehi and Seber (1997) described the estimation procedures for Hansen and Hurwitz (1943) and Horvitz and Thompson (1952) both.

2.1. Non-overlapping scheme

In the non-overlapping scheme, the modified Horvitz-Thompson estimator for the population mean $(\mu = \frac{\tau}{N})$ is

$$\widehat{\mu}_{HT}^{N} = \frac{1}{N}(M \sum_{i=1}^{m} \frac{\widehat{\tau}_i}{m} \quad)$$

where $\widehat{\tau}_i = \sum_{k=1}^{K_i} y_{ik}^{*}(\frac{I_{ik}}{\alpha_{ik}})$ is the unbiased estimate of i^{th} primary units total having variance $var\,(\widehat{\tau}_i) = \sum_{r=1}^{K_i} \sum_{s=1}^{K_i} y_{ir}^{*} y_{is}^{*}(\frac{\alpha_{irs} - \alpha_{ir}\alpha_{is}}{\alpha_{ir}\alpha_{is}})$.

To the above equations, K_i denotes the number of networks of the i^{th} primary unit and $\alpha_{ik} = 1 - \frac{\binom{N_i - m_{ik}}{n_i}}{\binom{N_i}{n_i}}$ is the probability that the initial sample of unit in i^{th} primary unit intersect the network k. Also, $\alpha_{ikk^T} = \alpha_{ik} + \alpha_{ik^T} - \left(1 - \frac{\binom{N_i - m_{ik} - m_{ikT}}{n_i}}{\binom{N_i}{n_i}}\right)$ is the probability that the initial sample of unit in i^{th} primary unit intersect both the networks k and k^T. The sum of y-value associate with the network k is denoted here by y_{ik}^{*}.

The variance estimator of $\widehat{\mu}_{HT}^{N}$ is

$$V(\widehat{\mu}_{HT}^{N}) = \frac{1}{N^2} M(M-m)\frac{\sigma_M^2}{m} + \frac{1}{N^2}\frac{M}{m}\sum_{i=1}^{M} var(\widehat{\tau}_i)$$

taking $\sigma_M^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\tau_i - \overline{\tau})^2$ and $\overline{\tau} = \frac{1}{M}\sum_{i=1}^{M}\tau_i$.

An unbiased estimate of $V(\widehat{\mu}_{HT}^{N})$ is

$$v(\widehat{\mu}_{HT}^{N}) = \frac{1}{N^2} M(M-m)\frac{s_M^2}{m} + \frac{1}{N^2}\frac{M}{m}\sum_{i=1}^{m} \widehat{var(\widehat{\tau}_i)}$$

where $\widehat{var(\widehat{\tau}_i)} = \sum_{r=1}^{\chi_i} \sum_{s=1}^{\chi_i} y_{ir}^{*} y_{is}^{*}(\frac{\alpha_{irs} - \alpha_{ir}\alpha_{is}}{\alpha_{irs}\alpha_{ir}\alpha_{is}})$.

Here, χ_i denotes the number of distinct networks intersected in the i^{th} primary unit.

2.2. Overlapping scheme

Here, all population units can be partitioned into K number of distinct networks, ignoring the PSU boundaries.

Thus, the modified Horvitz-Thompson estimator is

$$\widehat{\mu}_{HT}^{O} = \frac{1}{N}\left(\sum_{k=1}^{K}\frac{y_k^* J_k}{\alpha_k}\right) \quad .$$

In the above equation, J_k is the indicator function with the value 1 or 0 if the initial sample of size $n = \sum_{i=1}^{m} n_i$ intersects network k or not and y_k^* is the sum of $y-$values for the network k. Salehi and Seber (1997) derived the variance of $\widehat{\mu}_{HT}^{O}$ $\left(V\left(\widehat{\mu}_{HT}^{O}\right)\right)$ and an unbiased variance estimate $\left(v\left(\widehat{\mu}_{HT}^{O}\right)\right)$ as follows,

$$V\left(\widehat{\mu}_{HT}^{O}\right) = \frac{1}{N^2}\sum_{k=1}^{K}\sum_{k^T=1}^{K}\frac{y_k^* y_{k^T}^* (\alpha_{kk^T} - \alpha_k \alpha_{k^T})}{\alpha_k \alpha_{k^T}}$$

$$v\left(\widehat{\mu}_{HT}^{O}\right) = \frac{1}{N^2}\sum_{k=1}^{\chi}\sum_{k^T=1}^{\chi}\frac{y_k^* y_{k^T}^* (\alpha_{kk^T} - \alpha_k \alpha_{k^T})}{\alpha_{kk^T}\alpha_k \alpha_{k^T}} \quad .$$

Here, χ denotes the number of distinct networks in the sample and α_k is the inclusion probability for the network k and α_{kk^T} is the probability that the initial sample intersects both networks k and k^T. In order to evaluate $V\left(\widehat{\mu}_{HT}^{O}\right)$ and $v\left(\widehat{\mu}_{HT}^{O}\right)$, one needs to know the expressions for α_{kk^T} and α_k which are derived in the Appendix of Salehi and Seber (1997). Here, we have just written the formulas.

$$\begin{aligned}\alpha_k &= P\left[J_k = 1\right] \\ &= \sum_{i\in B_k}\frac{m}{M}\left(1 - \frac{\binom{N_i - m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right) - \sum_{i}\sum_{i^T<i}\frac{m(m-1)}{M(M-1)}\left(1 - \frac{\binom{N_i - m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right)\left(1 - \frac{\binom{N_{i^T} - m_{i^T k}}{n_i}}{\binom{N_{i^T}}{n_{i^T}}}\right) + \ldots \\ &\quad + (-1)^{g_k+1}\frac{m(m-1)\ldots(m-g_k+1)}{M(M-1)\ldots(M-g_k+1)}\prod_{i\in B_k}\left(1 - \frac{\binom{N_i - m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right)\end{aligned}$$

and

$$\alpha_{kk^T} = P\left[J_k = 1,\ J_{k^T} = 1\right]$$

where B_k is the set of PSUs intersected by the network k having g_k number of elements and m_{ik} is the number of units of network k located in i^{th} PSU.

3. Prediction approach in two stage sampling scheme

A finite population problem can be formulated as prediction problem and can be solved using Bayesian approach. A more classical superpopulation approach is also possible using Royall (1976)s theorem of best linear unbiased estimator.

Suppose, the objective is to estimate the population total

$$\tau = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{M} \tau_i$$

by two-stage design which can be expressed as

$$\tau = \sum_{i \in s} \sum_{j \in s_i} y_{ij} + \sum_{i \in s} \sum_{j \in s_i^c} y_{ij} + \sum_{i \in s^c} \sum_{j=1}^{N_i} y_{ij}. \tag{1}$$

Here, s is the PSU sample of size m and s^c is the set of PSU units not in s. Similarly, s_i is the SSU sample of $i^{th}(i \in s)$ PSU and s_i^c is the complementary of s_i.

In the above expression, it is obvious that the first term is known from the sample. However the second and third terms are unknown and it should be estimated.

The prediction approach of finite population theory considers the total τ is a realization of a random vector T. For a given sample,

$$T = \sum_{i \in s} \sum_{j \in s_i} y_{ij} + Z \tag{2}$$

with $Z = \sum_{i \in s} \sum_{j \in s_i^c} y_{ij} + \sum_{i \in s^c} \sum_{j=1}^{N_i} y_{ij}$.

Now, expressing T as (2), the problem of estimating T is equivalent to the prediction of Z.

Mathematically,

$$\widehat{T} = \sum_{i \in s} \sum_{j \in s_i} y_{ij} + \widehat{Z} \tag{3}$$

clarifies the matter.

The following probability model is adopted here to establish the relationship among N random variable Y_{ij} ; $i = 1, 2 \ldots . M \quad j = 1, 2 \ldots N_i$:

$$\begin{cases} E\left(Y_{ij}\right) = \theta \\ Cov\ \left(Y_{ij},\ Y_{lm}\right) & = \sigma_i^2, \quad i = l, j = m \\ & = \rho_i \sigma_i^2, i = l, j \neq m \\ & = 0, i \neq l \end{cases} \tag{4}$$

It is assumed here that the random variables within cluster i have common mean θ_i and variance σ_i^{T2} and covariance $\rho_i^T \sigma_i^{T2}$ and the $\{\theta_i^T\}$ are the realizations of uncorrelated random variables with common mean θ and variance φ^2. Then the model (4) applies with $\sigma_i^2 = \varphi^2 + \sigma_i^{T2}$ and $\rho_i = \frac{\varphi^2 + \rho_i^T \sigma_i^{T2}}{\varphi^2 + \sigma_i^{T2}}$.

Royall (1976) suggested an optimal (BLU) estimator in such case and this can be expressed as

$$\widehat{T}^* = \sum_{i \in s} \sum_{j \in s_i} y_{ij} + \sum_{i \in s} \left(N_i - n_i\right) \left[\omega_i \overline{y}_{si} + \left(1 - \omega_i\right) \widehat{\theta}\right] + \sum_{i \notin s} N_i \widehat{\theta} \tag{5}$$

where $\omega_i = \frac{\rho_i n_i}{(1-\rho_i+n_i\rho_i)}$) and $\widehat{\theta} = \sum_{i\in s}\theta_i\overline{y}_{si}$ is the weighted average of sample means with weights $\theta_i = \left[\frac{n_i\sigma_i^2}{(1-\rho_i+n_i\rho_i)}\right] / [\sum_{i\in s}\frac{n_i\sigma_i^2}{(1-\rho_i+n_i\rho_i)}]$.

Here, in $\widehat{T}^*$, non-sampled units in sample cluster i can be estimated by $\omega_i\overline{y}_{si} + (1-\omega_i)\,\widehat{\theta}$ and all the units in non-sampled clusters are estimated by $\widehat{\theta}$.

This $\widehat{T}^*$ further can be written as $\widehat{T}^* = \sum_{i\in s}(1+g_i)n_i\overline{y}_{si}$ taking $\sum_{i\in U}N_i = N$ and $\sum_{i\in s}n_i = n$, and $g_i = [\omega_i\frac{(N_i-n_i)}{n_i} + \{N-n-\sum_{i\in s}\omega_i\,(N_i-n_i)\}\frac{\theta_i}{n_i}]$.

The error variance of $\widehat{T}^*$ can be written as

$$
\begin{aligned}
Var\left(\widehat{T}^*-T\right) &= Var\left(\sum_{i\in s}g_in_i\overline{y}_{si} - \sum_{i\in s}\sum_{j\in s_i^c}y_{ij} - \sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right)\\
&= Var\left(\sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right) + Var\left(\sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right) + Var\left(\sum_{i\in s}g_in_i\overline{y}_{si}\right) + 2cov\left(\sum_{i\in s}\sum_{j\in s_i^c}y_{ij}, \sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right)\\
&\qquad - 2cov\left(\sum_{i\in s}g_in_i\overline{y}_{si}, \sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right) - 2cov\left(\sum_{i\in s}g_in_i\overline{y}_{si},\ \sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right)\\
&= Var\left(\sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right) + Var\left(\sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right) + Var\left(\sum_{i\in s}g_in_i\overline{y}_{si}\right) - 2cov\left(\sum_{i\in s}g_in_i\overline{y}_{si},\ \sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right)\\
&= v + Var\left(\sum_{i\in s}g_in_i\overline{y}_{si}\right) - 2cov\left(\sum_{i\in s}g_in_i\overline{y}_{si},\ \sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right)\\
&= v + \left(\sum_{i\in s}\rho_i\sigma_i^2n_i^2g_i^2 + \sum_{i\in s}(1-\rho_i)\,\sigma_i^2n_ig_i^2\right) - 2\sum_{i\in s}\rho_i\sigma_i^2g_in_i(N_i-n_i)\\
&= v - \sum_{i\in s}\rho_i\sigma_i^2(N_i-n_i)^2 + \sum_{i\in s}\rho_i\sigma_i^2[n_ig_i-(N_i-n_i)]^2 + \sum_{i\in s}(1-\rho_i)\sigma_i^2n_ig_i^2
\end{aligned}
$$

where

$$
\begin{aligned}
v &= Var\left(\sum_{i\in s}\sum_{j\in s_i^c}y_{ij}\right) + Var\left(\sum_{i\notin s}\sum_{j=1}^{N_i}y_{ij}\right)\\
&= \sum_{i\in s}(N_i-n_i)\,\sigma_i^2[1-\rho_i+(N_i-n_i)\,\rho_i] + \sum_{i\notin s}N_i\sigma_i^2[1-\rho_i+N_i\rho_i].
\end{aligned}
$$

4. Proposed predictors for two-stage ACS

Simple random sampling without replacement scheme is frequently used in ACS design to draw an initial sample. Chaudhuri (2000) clarified that any sampling method admitting an unbiased estimator for a population total may be extended to adaptive sampling design yielding unbiased estimator. This work insisted us to select PSUs adapting an unequal probability sampling, say PPSWOR instead of SRSWOR in case of Two-stage ACS design, as discussed in Section 2.

4.1. Non-overlapping scheme

Therefore, an unbiased estimator of population total $\tau = \sum_{i=1}^{M} \tau_i = \sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij}$ is

$$e = \sum_{i=1}^{m} \frac{\widehat{\tau}_i}{\pi_i} \tag{6}$$

taking $\widehat{\tau}_i$ as the estimate of i^{th} PSU total. This

$$\widehat{\tau}_i = \sum_{k=1}^{\chi_i} \frac{y_{ik}^*}{\alpha_{ik}}, \tag{7}$$

if networks are truncated at selected PSU (Non-overlapping scheme). Here, $y_{ik}^* = \sum_{j \in A(i,k)} y_{ij}$ is the sum of the $y-$values present in $A(i,k)$, the k^{th} network of i^{th} PSU. This network $A(i,k)$ can be partitioned into two parts captured $A_c(i,k)$ and uncaptured $A_{uc}(i,k)$. Obviously, $A(i,k) = A_c(i,k) \cup A_{uc}(i,k)$.

It is obvious that $E(e) = E_1E_2(e) = E_1\left(\sum_{i=1}^{m} \frac{\tau_i}{\pi_i}\right) = \sum_{i=1}^{M} \tau_i = \tau$. E_1 denotes here the expectation due to first stage unit selection and E_2 , the expectation due to second stage.

The variance of e can be written as,

$$V(e) = E_1V_2(e) + V_1E_2(e)$$

where $E_1(V_2(e)) = E_1(V_2\left(\sum_{i=1}^{m} \frac{1}{\pi_i} \sum_{k=1}^{\chi_i} \frac{y_{ik}^*}{\alpha_{ik}}\right)) = E_1(\sum_{i=1}^{m} \frac{1}{\pi_i^2} V_2\left(\sum_{k=1}^{\chi_i} \frac{y_{ik}^*}{\alpha_{ik}}\right))$

$$= E_1\left(\sum_{i=1}^{m} \frac{1}{\pi_i^2}\left(\sum_{k=1}^{K_i}\sum_{k^T=1}^{K_i} y_{ik}^* y_{ik^T}^* \left(\frac{\alpha_{ikk^T} - \alpha_{ik}\alpha_{ik^T}}{\alpha_{ik}\alpha_{ik^T}}\right)\right)\right)$$

$$= \sum_{i=1}^{M} \frac{1}{\pi_i}\left(\sum_{k=1}^{K_i}\sum_{k^T=1}^{K_i} y_{ik}^* y_{ik^T}^* \left(\frac{\alpha_{ikk^T} - \alpha_{ik}\alpha_{ik^T}}{\alpha_{ik}\alpha_{ik^T}}\right)\right)$$

and $V_1(E_2(e)) = V_1\left(\sum_{i=1}^{m} \frac{\tau_i}{\pi_i}\right) = \sum_{i<j} \sum_{=1}^{M} (\pi_i\pi_j - \pi_{ij})(\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j})^2$.

To compute unbiased estimate of $V(e)$, let assume

$$v_1(e) = \sum_{i=1}^{m} \frac{1}{\pi_i^2}\left(\sum_{k=1}^{\chi_i}\sum_{k^T=1}^{\chi_i} y_{ik}^* y_{ik^T}^* \left(\frac{\alpha_{ikk^T} - \alpha_{ik}\alpha_{ik^T}}{\alpha_{ikk^T}\alpha_{ik}\alpha_{ik^T}}\right)\right) + \sum_{i<j}\sum_{=1}^{m}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)(\frac{\widehat{\tau}_i}{\pi_i} - \frac{\widehat{\tau}_j}{\pi_j})^2$$

Therefore, $E\left(v_1(e)\right) = E_1E_2\left(v_1(e)\right)$

$$= E_1\left(\sum_{i=1}^{m}\frac{1}{\pi_i^2}\left(\sum_{k=1}^{K_i}\sum_{k^T=1}^{K_i} y_{ik}^* y_{ik^T}^*\left(\frac{\alpha_{ikk^T}-\alpha_{ik}\alpha_{ik^T}}{\alpha_{ik}\alpha_{ik^T}}\right)\right)\right)$$
$$+ E_1\left(\sum_{i<j}\sum_{=1}^{m}\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\left(\frac{\tau_i}{\pi_i}-\frac{\tau_j}{\pi_j}\right)^2+\frac{V_2\left(\widehat{\tau}_i\right)}{\pi_i^2}+\frac{V_2\left(\widehat{\tau}_j\right)}{\pi_j^2}\right)\right)$$
$$= \sum_{i=1}^{m}\frac{1}{\pi_i}\left(\sum_{k=1}^{K_i}\sum_{k^T=1}^{K_i} y_{ik}^* y_{ik^T}^*\left(\frac{\alpha_{ikk^T}-\alpha_{ik}\alpha_{ik^T}}{\alpha_{ik}\alpha_{ik^T}}\right)\right)+\sum_{i<j}\sum_{=1}^{M}(\pi_i\pi_j-\pi_{ij})\left(\frac{\tau_i}{\pi_i}-\frac{\tau_j}{\pi_j}\right)^2$$
$$+\sum_{i<j}\sum_{=1}^{M}(\pi_i\pi_j-\pi_{ij})\left(\frac{V_2\left(\widehat{\tau}_i\right)}{\pi_i^2}+\frac{V_2\left(\widehat{\tau}_j\right)}{\pi_j^2}\right)$$
$$= V(e)+\sum_{i<j}\sum_{=1}^{M}(\pi_i\pi_j-\pi_{ij})\left(\frac{V_2\left(\widehat{\tau}_i\right)}{\pi_i^2}+\frac{V_2\left(\widehat{\tau}_j\right)}{\pi_j^2}\right)$$

where $V_2\left(\widehat{\tau}_i\right) = \sum_{k=1}^{K_i}\sum_{k^T=1}^{K_i} y_{ik}^* y_{ik^T}^*\left(\frac{\alpha_{ikk^T}-\alpha_{ik}\alpha_{ik^T}}{\alpha_{ik}\alpha_{ik^T}}\right)$.

Thus,

$$v(e) = v_1(e) - \sum_{i<j}\sum_{=1}^{m}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}\left(\frac{\widehat{V_2\left(\widehat{\tau}_i\right)}}{\pi_i^2}+\frac{\widehat{V_2\left(\widehat{\tau}_j\right)}}{\pi_j^2}\right) \tag{8}$$

is an unbiased estimator of $V(e)$ where $\widehat{V_2\left(\widehat{\tau}_i\right)} = \sum_{k=1}^{\chi_i}\sum_{k^T=1}^{\chi_i} y_{ik}^* y_{ik^T}^*\left(\frac{\alpha_{ikk^T}-\alpha_{ik}\alpha_{ik^T}}{\alpha_{ikk^T}\alpha_{ik}\alpha_{ik^T}}\right)$.

Now, in case surveyors are unable to gather information from all units belonging to a network, then mathematically it can be express as

$$y_{ik}^* = \sum_{j\in A(i,k)} y_{ij} = \sum_{j\in A_c(i,k)} y_{ij} + \sum_{j\in A_{uc}(i,k)} y_{ij}.$$

Undoubtedly, second term of this expression is unknown and can be predicted easily following Section 3.1 and 3.2 of Pal and Patra (2021). We avoid here the unnecessary repetition.

However, the complication arises if the surveyor decided to ignore PSU boundaries for network construction. Below we describe prediction steps in this case, in details.

4.2. Overlapping scheme

In this case, we need to consider the distinct networks included in two-stage. Thus, an unbiased estimator of the population total τ may be written as

$$e^* = \sum_{k=1}^{\chi}\frac{y_k^*}{\alpha_k^*} \tag{9}$$

where $\alpha_k^* = \sum_{i\in B_k}\pi_i\left(1-\frac{\binom{N_i-m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right) - \sum_i\sum_{i^T<i}\pi_{ii^T}\left(1-\frac{\binom{N_i-m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right)\left(1-\frac{\binom{N_{i^T}-m_{i^Tk}}{n_i}}{\binom{N_{i^T}}{n_{i^T}}}\right)+\ldots+(-1)^{g_k+1}\pi_{ii^T\ldots..l}\prod_{i\in B_k}\left(1-\frac{\binom{N_i-m_{ik}}{n_i}}{\binom{N_i}{n_i}}\right)$ and $\alpha_{kk^T}^* = P\left[J_k=1,\ J_{k^T}=1\right]$. Here B_k, with g_k number of elements, is the set of those primary units intersected by k^{th} network.

The variance is $V(e^*) = \left(\sum_{k=1}^{\chi}\sum_{k^T=1}^{\chi} y_k^* y_{k^T}^* \left(\frac{\alpha_{kk^T}^* - \alpha_k^*\alpha_{k^T}^*}{\alpha_k^*\alpha_{k^T}^*}\right)\right)$ and an unbiased estimate of variance is $v(e^*) = \left(\sum_{k=1}^{\chi}\sum_{k^T=1}^{\chi} y_k^* y_{k^T}^* \left(\frac{\alpha_{kk^T}^* - \alpha_k^*\alpha_{k^T}^*}{\alpha_{kk^T}^*\alpha_k^*\alpha_{k^T}^*}\right)\right)$. However, the computation of α_k^* and $\alpha_{kk^T}^*$ are not an easy task. Thus, a modification is needed.

We take Chaudhuri (2000)s approach here to propose an unbiased estimator of τ as

$$e^{T*} = \sum_{i=1}^{m}\frac{\widehat{\tau}_i}{\pi_i} = \sum_{i=1}^{m}\frac{1}{\pi_i}\left(\frac{N_i}{n_i}\sum_{j=1}^{n_i} t_{ij}\right) \tag{10}$$

where $t_{ij} = \frac{1}{d_{ij}}\sum_{i=1}^{M}\sum_{j\in A(i,j)} y_{ij}$ is the average of $y-values$ of the units belong to the network $A(i,j)$, ignoring the PSU boundaries. It is much easier to compute than the previous one (equation 9).

Taking expectation, we get

$$\begin{aligned}
E\left(e^{T*}\right) &= E_2E_1\left(\sum_{i=1}^{m}\frac{1}{\pi_i}\left(\frac{N_i}{n_i}\sum_{j=1}^{n_i} t_{ij}\right)\right) = E_2(\sum_{i=1}^{M}\frac{1}{\pi_i}(\frac{N_i}{n_i}\sum_{j=1}^{n_i} t_{ij})\pi_i) \\
&= E_2\left(\sum_{i=1}^{M}\frac{N_i}{n_i}\sum_{j=1}^{n_i} t_{ij}\right) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N_i} t_{ij} = \sum_{i=1}^{M}\sum_{j=1}^{N_i} y_{ij} \quad \text{(see Thompson's (1990) and Chaudhuri (2000))} \\
&= \tau, \text{ the population total.}
\end{aligned}$$

Table 1: Two-stage ACS structure for population

PSU	SSU	Networks of SSU	Cardinality of Networks	Statistic based on SSU
1	$y_{11}, y_{12}, \ldots, y_{1N_1}$	$A(1;1),\ A(1;2)\ldots A(1,N_1)$	$d_{11}, d_{12}, \ldots, d_{1N_1}$	$t_{11}, t_{12}, \ldots, t_{1N_1}$
2	$y_{21}, y_{22}, \ldots, y_{2N_2}$	$A(2;1),\ A(2;2)\ldots A(2,N_2)$	$d_{21}, d_{22}, \ldots, d_{2N_2}$	$t_{21}, t_{22}, \ldots, t_{2N_2}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	$y_{i1}, y_{i2}, \ldots, y_{iN_i}$	$A(i;1),\ A(i;2)\ldots, A(i,N_i)$	$d_{i1}, d_{i2}, \ldots, d_{iN_i}$	$t_{i1}, t_{i2}, \ldots, t_{iN_i}$
M	$y_{M1}, y_{M2}, \ldots, y_{MN_M}$	$A(M;1),\ A(M;2)\ldots, A(M,N_M)$	$d_{M1}, d_{M2}, \ldots, d_{MN_M}$	$t_{M1}, t_{M2}, \ldots, t_{MN_M}$

The variance can be written as

$V\left(e^{T*}\right) = E_2V_1\left(e^{T*}\right) + V_2E_1\left(e^{T*}\right) = \sum_{i<j}\sum_{=1}^{M}(\pi_i\pi_j - \pi_{ij})(\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j})^2 + \sum_{i=1}^{M}\frac{N_i^2}{n_i}(1-f_i)S_i^2$

where $S_i^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(t_{ij}-\bar{t}_i)^2$ and $\bar{t}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} t_{ij}$

An unbiased estimator of $V\left(e^{T*}\right)$ is

$$v\left(e^{T*}\right) = \sum_{i<j}\sum_{=1}^{m}(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}})(\frac{\widehat{\tau}_i}{\pi_i} - \frac{\widehat{\tau}_j}{\pi_j})^2 + \sum_{i=1}^{m}\frac{N_i^2}{n_i}(1-f_i)\,s_i^2 \tag{11}$$

where s_i^2 is an unbiased estimator of S_i^2.

Let $A(i;j)$ can be written as $A_c(i;j) \cup A_{uc}(i;j)$ where $A_c(i;j)$ is the observed units and $A_{uc}(i;j)$ is a set of unobserved units of the network of j^{th} unit of i^{th} PSU. Also, let the cardinality of each set is known and it is possible due to satellite imagery or previous records.

Then, $t_{ij} = \frac{1}{d_{ij}} \sum_{i=1}^{M} \sum_{j \in A(i,j)} y_{ij}$ may be treated as

$$t_{ij} = \frac{1}{d_{ij}} \left[\left(\sum_{i \in s} \sum_{j \in A_c(i;j)} y_{ij} + \sum_{i \notin s} \sum_{j \in A_c(i;j)} y_{ij} \right) + \left(\sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij} \right) + \left(\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij} \right) \right] \quad (12)$$

$$= \frac{1}{d_{ij}} [(\sum_{i \in s} \text{sum of observed units from } i^{th} \text{ PSU} + \sum_{i \notin s} \text{sum of observed units from } i^{th} \text{ PSU})$$

$$+ \sum_{i \in s} \text{sum of unobserved from } i^{th} \text{ PSU} + \sum_{i \notin s} \text{sum of unobserved from } i^{th} \text{ PSU}]$$

for a network $A(i;j)$

Similarly, $d_{ij}-$ cardinality of the network $A(i;j)$ can be partitioned as

$$d_{ij} = \left(\sum_{i \in s} d_{i1}(ij) + \sum_{i \notin s} d_{i2}(ij) \right) + \sum_{i \in s} d_{i3}(ij) + \sum_{i \notin s} d_{i4}(ij) \quad (13)$$

$$= (d_1(ij) + d_2(ij)) + d_3(ij) + d_4(ij) \quad (14)$$

where $d_{i1}(ij)$ is the number of observed units belongs to $i^{th}(i \in s)$ PSU from $A(i;j)$ network and $d_{i2}(ij)$ is the number of observed units belongs to $i^{th}(i \notin s)$ PSU but from the network $A(i;j)$. Similarly, $d_{i3}(ij)$ and $d_{i4}(ij)$ are the numbers of unobserved units belongs to sampled and non-sampled PSU from $A(i;j)$ network, respectively.

Thus to estimate t_{ij}, we need to predict the terms $\sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij}$ and $\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}$.

Now, following Royall (1976)s prediction approach and adopting the model (4) with restrictions $\rho_i = \rho$ and $\sigma_i^2 = \sigma^2$, we get

$$\begin{cases} E(Y_{ij}) = \delta \\ Cov\ (Y_{ij},\ Y_{lm}) = \sigma^2, & i = l, j = m \\ \qquad\qquad\quad = \rho\sigma^2, & i = l, j \neq m \\ \qquad\qquad\quad = 0, & i \neq l \end{cases} \quad (15)$$

With the model (15), we may get

$$est \left(\sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij} \right) = \sum_{i \in s} d_{i3}(ij) \left[w_i \overline{y}_{si} + (1 - w_i) \widehat{\delta} \right] \qquad \ldots using\ (5)$$

where $w_i = \frac{\rho d_{i1}(ij)}{1-\rho-\rho d_{i1}(ij)}$ and $\widehat{\delta} = \sum_{i \in s} \delta_i \overline{y}_{si}$ with $\delta_i = \frac{[\frac{d_{i1}(ij)}{(1-\rho-\rho d_{i1}(ij))}]}{[\sum_{i \in s} \frac{d_{i1}(ij)}{(1-\rho-\rho d_{i1}(ij))}]}$.

Here, $\overline{y}_{si} = \frac{1}{d_{i1}(ij)} \sum_{j \in A_c(i;j)} y_{ij} \quad \forall i \in s$ is the average of those observed units from a sampled PSU i, belongs to the network $A_c(i;j)$.

It is noteworthy that ρ is generally unknown to us. It can be estimated by analysis of variance (ANOVA) technique (see Valliant *et al.* (2000) chapter 8), if prior information is not given.

With the assumption $\overline{\overline{y}}_* = \frac{\sum_{i \in s} d_{i1}(ij)\overline{y}_{si}}{\sum_{i \in s} d_{i1}(ij)} = \frac{\sum_{i \in s} d_{i1}(ij)\overline{y}_{si}}{d_1(ij)}$, sum of squares of the ANOVA is derived based on the following relation:

$$\begin{aligned}
\sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{\overline{y}}_*)^2 &= \sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si} + \overline{y}_{si} - \overline{\overline{y}}_*)^2 \\
&= \sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si})^2 + \sum_{i \in s} \sum_{j \in A_c(i;j)} (\overline{y}_{si} - \overline{\overline{y}}_*)^2 + 2 \sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si})(\overline{y}_{si} - \overline{\overline{y}}_*) \\
&= \sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si})^2 + \sum_{i \in s} \sum_{j \in A_c(i;j)} (\overline{y}_{si} - \overline{\overline{y}}_*)^2 \\
&= \sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si})^2 + \sum_{i \in s} d_{i1}(ij)(\overline{y}_{si} - \overline{\overline{y}}_*)^2
\end{aligned}$$

Table 2: ANOVA table for a sample taken by two-stage adaptive cluster sampling

Source	Sum of squares	Degrees of freedom	Expected mean squares
Between Clusters	$\sum_{i \in s} d_{i1}(ij)(\overline{y}_{si} - \overline{\overline{y}}_*)^2$	$m^* - 1$	$\sigma^2(1-\rho) + \frac{\rho\sigma^2}{m^*-1}\{d_1(ij) - \sum_{i \in s} \frac{d_{i1}^2(ij)}{d_1(ij)}\}$
Within Clusters	$\sum_{i \in s} \sum_{j \in A_c(i;j)} (y_{ij} - \overline{y}_{si})^2$	$d_1(ij) - m^*$	$\sigma^2(1-\rho)$

m^* =Number of sampled PSUs in the cluster

Now, for the term $\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}$,

$$est\left(\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right) = \sum_{i \notin s} d_{i4}(i;j)\widehat{\delta} \quad \ldots using\ (5)$$

Thus, our suggested optimal (BLU) predictor is

$$\widehat{t_{ij}} = \frac{1}{d_{ij}} \left[\left(\sum_{i \in s} \sum_{j \in A_c(i;j)} y_{ij} + \sum_{i \notin s} \sum_{j \in A_c(i;j)} y_{ij} \right) + \sum_{i \in s} d_{i3}(ij) \left[w_i \overline{y}_{si} + (1 - w_i)\, \widehat{\delta} \right] + \sum_{i \notin s} d_{i4}(ij)\widehat{\delta} \right] \tag{16}$$

and the above can be written as

$$\widehat{t_{ij}} = \left(\frac{1}{d_{ij}} \sum_{i \in s} \sum_{j \in A_c(i;j)} y_{ij} \right) + \frac{1}{d_{ij}} \sum_{i \in s} (1 + g_i^*) d_{i1}(ij) \overline{y}_{si} \tag{17}$$

where

$$g_i^* = \left\{ \frac{d_{i3}(ij)}{d_{i1}(ij)} w_i + \frac{\delta_i}{d_{i1}(ij)} \sum_{i \in s} d_{i3}(ij)\,(1 - w_i) + \frac{\delta_i}{d_{i1}(ij)} d_4(ij) \right\}.$$

Now, the error variance of $\widehat{t_{ij}}$ can be derived as below.

$$\begin{aligned} MSE\left(\widehat{t_{ij}}\right) &= Var\left(\widehat{t_{ij}} - t_{ij}\right) \qquad (18)\\ &= Var\left(\frac{1}{d_{ij}} \sum_{i \in s} g_i^* d_{i1}(ij) \overline{y}_{si} - \frac{1}{d_{ij}} \sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij} - \frac{1}{d_{ij}} \sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right)\\ &= \frac{1}{d_{ij}^2} \left\{ Var\left(\sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right) + Var\left(\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right) + Var\left(\sum_{i \in s} g_i^* d_{i1}(ij) \overline{y}_{si}\right) \right\}\\ &\qquad - \frac{1}{d_{ij}^2} \left\{ 2\ cov\left(\sum_{i \in s} g_i^* d_{i1}(ij) \overline{y}_{si}\ ,\ \sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right) \right\}\\ &= \frac{1}{d_{ij}^2} \left[v^* + \left(\sum_{i \in s} \rho \sigma^2 d_{i1}^2(ij) g_i^{*2} + \sum_{i \in s} (1 - \rho) \sigma^2 d_{i1}(ij) g_i^{*2} \right) - 2 \sum_{i \in s} \rho \sigma^2 g_i^* d_{i1}(ij) d_{i3}(ij) \right]\\ &= \frac{1}{d_{ij}^2} \left[v^* - \rho \sigma^2 \sum_{i \in s} d_{i3}^2(ij) + \rho \sigma^2 \sum_{i \in s} (d_{i1}(ij) g_i^* - d_{i3}(ij))^2 + \sum_{i \in s} (1 - \rho) \sigma^2 d_{i1}(ij) g_i^{*2} \right] \end{aligned}$$

where $v^* = Var\left(\sum_{i \in s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right) + Var\left(\sum_{i \notin s} \sum_{j \in A_{uc}(i;j)} y_{ij}\right)$

$= \sum_{i \in s} d_{i3}(ij) \sigma^2 (1 - \rho + d_{i3}(ij) \rho) + \sum_{i \notin s} d_{i4}(ij) \sigma^2 (1 - \rho + d_{i4}(ij) \rho).$

Thus,

$$\widehat{e}^{T*} = \sum_{i=1}^{m} \frac{\widehat{\tau}_i^*}{\pi_i} = \sum_{i=1}^{m} \frac{1}{\pi_i} \left(\frac{N_i}{n_i} \sum_{j=1}^{n_i} \widehat{t_{ij}} \right) \qquad (19)$$

becomes our final estimator of population total with variance estimator

$$v\left(\widehat{e}^{T*}\right) = \sum_{i<j}^{m} \sum_{=1} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{\widehat{\tau}_i^*}{\pi_i} - \frac{\widehat{\tau}_j^*}{\pi_j}\right)^2 + \sum_{i=1}^{m} \frac{N_i^2}{n_i} (1 - f_i)\, s_i^{*2} + \sum_{i=1}^{m} \sum_{j=1}^{n_i} MSE(\widehat{t_{ij}}) \qquad (20)$$

where $s_i^{*2} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\widehat{t_{ij}} - \widehat{\bar{t}}_i)^2$ and $\widehat{\bar{t}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \widehat{t_{ij}}$.

Note that $\widehat{t_{ij}} = t_{ij}$ if values of all units in the network $A(i, j)$is known.

5. Numerical Example

To illustrate our proposed methodology in prediction approach for Two-stage ACS-overlapping scheme numerically, we consider here Population 1 - the point-objects population of Thompson's (1990) which is further reproduced in Rocco (2008)-page-319 as Figure 1. The population contains $N = 400$ units and it is partitioned into $M = 20$ primary units each of $N_i = 20$ $(\forall i = 1, 2, \ldots, M)$ secondary units. From Population 1, it can be seen that very few units having $y-$ values greater than 0 and the population total is $\tau = \sum_{i=1}^{20} \sum_{j=1}^{20} y_{ij} = 190$. Now, we assume the rarity condition for ACS design is $y > 0$.

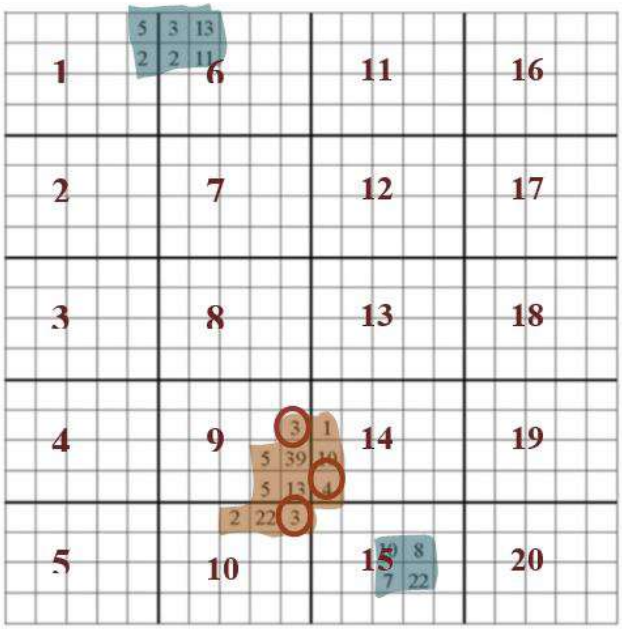


Figure 1: Two-stage ACS data

Table 3 represents the information gathered from a sample with six PSUs each with five SSUs. This sample is selected by Lahiri (1951)-Midzuno (1952)-Sen (1953) sampling scheme for first stage units and SRSWOR for second stage units. With this sampled data, we mainly illustrate the proposed methodology numerically for overlapping scheme, step-by-step. Here, $m = 6$ and $n_i = 5 \ \forall \ i = 1, 2, \ldots, m$. The 1^{st} column of Table 3 shows the selected PSUs. Inclusion probabilities of the selected PSUs are computed for Lahiri-Midzuno-Sen scheme and mentioned in 2^{nd} column. Also, $y - values$ of selected SSUs are shown in the table (4^{th} column) along with the network size (5^{th} column), if two-stage ACS-overlapping scheme is performed. It can be seen that in this case there are only 2 SSUs having non-zero $y - value$ and based on these SSUs, we can capture more rare units through ACS design. In this way, we found a network of 11 units ignoring the PSU boundaries, of which some are unobserved. Here the unobserved units are marked by red circle.

Table 3: Sampled data

Selected PSU	**Inclusion Probability** (π_i)	**Selected SSU(j) for a particular PSU(i)**	**y-values** (y_{ij})	**Cardinality of Network**(d_{ij})
1	0.323	11,17,14, 07,10	0,**5**,0,0,0	01,**06**,01,01,01
5	0.309	05,09,01,13,18	0,0,0,0,0	01,01,01,01,01
9	0.325	04,11,17,20,03	0,0,0,**13,**0	01,01,01,**11,**01
10	0.270	02,11,10,18,12	0,0,0,0,0	01,01,01,01,01
12	0.268	19,16,08,05,10	0,0,0,0,0	01,01,01,01,01
17	0.276	05,03,18,10,15	0,0,0,0,0	01,01,01,01,01

Now, let us consider the network of **20^{th} SSU** of **9^{th} PSU**s ($A(9;20)$-*orange shaded area*) which can be treated as $A_c(9;20) \cup A_{uc}(9;20)$. The set of observed units, $A_c(9;20)$ contains 15^{th}, 16^{th}, 19^{th}, 20^{th} units from **9^{th} PSU** and 9^{th}, 13^{th} units from **10^{th} PSU** and

also 2^{nd} , 3^{rd} units from **14^{th} PSU**. The set of unobserved units, $A_{uc}(9;20)$ contains 18^{th} unit from **9^{th} PSU**, 17^{th} unit from **10^{th} PSU** and 4^{th} unit from **14^{th} PSU**.

Thus, the cardinality of the network $A(9;20)$ can be partitioned as $11 = (4+2) + (1+1) + (1+1) + 1$, according to equations (13) and (14).

Now from equation (12) we get,

$\sum_{i\in s}\sum_{j\in A_c(i,j)} y_{ij} = (5+5+39+13) + (2+22\ ) = 62+24 = 86$ and

$\sum_{i\notin s}\sum_{j\in A_c(i,j)} y_{ij} = (1+10) = 11.$

However, $\sum_{i\in s}\sum_{j\in A_{uc}(i,j)} y_{ij}$ and $\sum_{i\notin s}\sum_{j\in A_{uc}(i,j)} y_{ij}$ are unknown to us and can be predicted through equation (4.2.7) and ANOVA with ρ and σ^2 , two unknown again.

Now, to predict ρ and σ^2, let us first compute sum of squares for between cluster (SSB) and within cluster (SSW).

Here, $SSB = \sum_{i\in s} d_{i1}(ij)(\overline{y}_{si} - \overline{\overline{y}}_{*})^2 = 4(\frac{62}{4} - \frac{86}{6})^2 + 2(\frac{24}{2} - \frac{86}{6})^2 = 16.33$ and $SSW = \sum_{i\in s}\sum_{j\in A_c(i;j)} (y_{ij} - \overline{y}_{si})^2 = 979$ and $\rho = -0.538,\ \sigma^2 = 159.103.$

It is noteworthy that under model (16), ρ can be negative however there is a lower bound. In this case the lower bound is -0.599. To get better idea of this, readers may consider Valliant *et al.* (2000, page 261). The above mentioned two unknown sums can be predicted by $\sum_{i\in s} d_{i3}(ij)\left[w_i\overline{y}_{si} + (1-w_i)\,\widehat{\delta}\right]$, $\sum_{i\notin s} d_{i4}(ij)\widehat{\delta}$ respectively and the predicted values are 28.103, 14.051. The values of w_i and $\widehat{\delta}$ are computed as per given formulas in Section 4. Thus, $\widehat{t_{ij}} = \frac{1}{11}(86 + 11 + 28.103 + 14.051) = 12.65$. Note that, the actual t_{ij} is 9.727 if all units of this network ($A(9;20)$) are observed.

Therefore, based on the sampled data (see Table 3) the final estimate of population total is $229.9957 \approx 230$ (using equation 19) and the estimated variance is 16074.43 (using equation 20). It is worth noting that if all units from the sampled networks are observed, then the estimated population total and estimated variance are 194.0203 and 10722.75 respectively. In other words, if all units from a sampled network are observed, one may get better result. It is obvious condition. However, these two situations are incomparable.

6. Conclusion

Two-stage sampling has several advantages over ordinary single stage (one-stage) sampling. In application of two-stage sampling in ACS, we add many units stopping at the PSU boundary or crossing across the PSU boundary. It is quite obvious that the surveyors may be unable to gather information from one or more rare units. Under such a situation, prediction approach under linear regression model considering correlation structure within network in two-stage ACS is satisfactory. Thus, to achieve a practical solution in two-stage ACS, we have employed Royalls prediction approach. In practice, it brings a novelty in prediction of the population total involving rare units under two stage sampling.

References

Chaudhuri, A. (2000). Network and adaptive sampling with unequal probabilities. *Calcutta Statistical Association Bulletin*, **50**, 237–253.

Hansen, M. M. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14**, 333–362.

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

Lahiri, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of International Statistical Institute*, **3**, 133–140.

Midzuno, H. (1952). On the sampling system with probability proportional to the sum of the sizes. *Annals of Institute of Statistical Mathematics*, **3**, 99–107.

Pal, S. and Patra, D. (2021). A prediction approach in adaptive sampling. *Metron*, **79**, 93–108.

Pal, S. and Patra, D. (2023). Modifications on re-scaling bootstrap for adaptive sampling. *Communications in Statistics - Simulation and Computation*, **52**, 1607–1620.

Patra, D. and Pal, S. (2023). Application of kalman filtering with bayesian formulation in adaptive sampling. *Communications in Statistics - Simulation and Computation*, **1**. https://doi.org/10.1080/03610918.2023.2265084.

Rocco, E. (2008). Two-stage restricted adaptive cluster sampling. *Metron*, **66**, 313–327.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377–389.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of American Statistical Association*, **71**, 657–664.

Salehi, M. M. and Seber, G. A. (1997). Adaptive cluster sampling with networks selected without replacement. *Biometrika*, **84**, 209–219.

Seber, G. A. and Salehi, M. M. (2013). *Adaptive Sampling Designs: Inference for Sparse and Clustered Populations*. Springer.

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

Thompson, S. K. (1991a). Adaptive cluster sampling: designs with primary and secondary units. *Biometrics*, **47**, 1103–1115.

Thompson, S. K. (1991b). Stratified adaptive cluster sampling. *Biometrika*, **78**, 389–397.

Thompson, S. K. (1992). *Sampling*. John Wiley & Sons.

Thompson's, S. K. (1990). Adaptive cluster sampling. *Journal of American Statistical Association*, **85**, 1050–1059.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley New York.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26–28 February 2024; pp 197–206

Predicting Stunting in Under-Five Children in Low Socio-Demographic Index States of India: A Machine Learning Approach

Mukesh Vishwakarma[1], **Gargi Tyagi**[1], **and Pawan Kumar Dubey**[2]
[1]*Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan.*
[2]*Monitoring & Evaluation, India Health Action Trust, Lucknow, Uttar Pradesh.*

Received: 10 August 2024; Revised: 01 September 2024; Accepted: 06 September 2024

Abstract

Stunting, the most prevalent form of child malnutrition, is characterized by a lack of height relative to age in children. Globally, 5.2 million children under five dies, with catastrophic stunting rates in central and southern Asia. Pakistan, India, and Afghanistan have the highest malnutrition rates in South Asia. India has about 270 million poor individuals and one-third of malnourished children. The National Family Health Survey (NFHS-5), 2019-21 found that 35.5% of Indian children under five are stunted. Factors influencing nutritional status include social, economic, educational, and maternal health issues. Despite efforts, India struggles to significantly curb child undernutrition, especially in states like Uttar Pradesh, Bihar, Rajasthan, Madhya Pradesh, Chhattisgarh, and Jharkhand, classified as low SDI states based on their Socio-Demographic Index (SDI) in the Global Burden of Disease study for 2019. The objective of this study is to train and evaluate machine learning (ML) classification algorithms on the National Family Health Survey (NFHS-5), 2019-21 dataset for predicting stunting among children under five years of age in states with a low socio-demographic index (LSDI). The machine learning models applied in this study include logistic regression (LR), random forest (RF), support vector classification (SVC), decision tree classifier (DTC), and gradient boosting classifier (GBC) algorithms. The performance of the ML algorithms are evaluated and compared using accuracy, recall, precision, F1-score, receiver operating curve (ROC) and recall curves on test dataset and 5-fold cross validation dataset. Important features of childhood stunting are also identified using Random Forest algorithm. It is observed that out of 82,158 children, 39% were stunted. Among the algorithms applied, the GBC algorithm achieved the highest accuracy in predicting stunting, with 65.5% on the testing data and 65 % $\pm$ 7% on the 5-fold cross validation data. In LSDI states of India, social structure and mother education are found to be major predictors of stunting in children under five, according to the random forest model for features importance. These results can aid in the swift diagnosis of stunting and the prompt development of preventive measures.

Key words: Stunting; Malnutrition; Child development; Healthy growth; Machine learning.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Mukesh Vishwakarma
Email: vishwakarma.m96@gmail.com

1. Introduction

Stunting is one of the most serious health and welfare issues across the world with more than 149 million children, which accounts for 21% of all children under the age of five, suffer from stunted growth. Moreover, the majority, 91%, of these children reside in low- and middle-income countries (LMICs) like India (UNICEF, 2020). Stunting is a condition that occurs when children suffer from prolonged inadequate nutrition and is defined as weight-for-height < -2 SD (standard deviation) in the WHO Growth Standard median (WHO, 2019).

Between 2005–06 and 2015–16, India has made a modest decrease in the prevalence of stunting and underweight in children under five years, but the progress is insufficient compared to its economic growth. Although there was a moderate decline in child undernutrition during this period, over one-third of children under five years old remain stunted (Jose *et al.*, 2018). This situation of stunting in India is quite evident in highly populated regions, namely Uttar Pradesh, Bihar, Rajasthan, and Madhya Pradesh *etc.* These states also come at lower end of Socio demographic Index (SDI) paradigm. The Socio-demographic Index (SDI) is a composite index of development that is significantly associated with health impact. It represents the geometric mean of the indices ranging from 0 to 1 for mean education among individuals aged 15 or older (EDU15+), total fertility rate under 25 (TFU25), and lag distributed income (LDI) per capita. A location with an SDI of 0 has a theoretical minimal level of health-related development, whereas a location with an SDI of 1 has a theoretical maximum level (Global Burden of Disease Collaborative Network, 2020). The SDI quantiles are utilized for classification. Based on their SDI, the states Uttar Pradesh, Bihar, Rajasthan, Madhya Pradesh, Chhattisgarh, and Jharkhand fall into the category of LSDI states.

Over the years, classical statistical models have been utilised to discover characteristics that are autonomously linked to stunting in children under the age of five (Mzumara *et al.*, 2018; Rakotomanana *et al.*, 2017; Das and Gulshan, 2017). However, these methods are not reliable in instances where the number of covariates exceeds the number of observations and when there is multicollinearity among variables. In addition, these models adhere to stringent assumptions regarding the data and the method by which the data is generated. These assumptions include the distribution of errors and the linearity of parameters with linear predictors. However, it is important to note that these assumptions may not be valid in real-world scenarios (Rajula *et al.*, 2020). Machine learning approaches surpass traditional models by addressing the analytical difficulties associated with a large number of covariates and multicollinearity. They also require fewer assumptions and can handle high dimensional data, resulting in a more adaptable relationship between predictor and outcome variables (Iniesta *et al.*, 2016). These techniques have been utilised to forecast malnutrition by employing various datasets (Shahriar *et al.*, 2019; Jin *et al.*, 2020; Markos *et al.*, 2014; Talukder and Ahammed, 2020). Moreover, machine learning techniques have demonstrated their superiority over traditional statistical methods in solving categorisation difficulties.

In this study, we focused on training, evaluating, and selecting the optimal machine learning classifier to predict stunting in children under five years old in low sociodemographic index (LSDI) states of India. Utilizing data from the NFHS-5 (2019-21), we also aimed to identify key variables that contribute to stunting. This model is intended to lay the

groundwork for creating an intelligent system for diagnosing or predicting stunting. The identified predictors of stunting will be prioritized in the development of interventions aimed at preventing stunting among children under five in LSDI states of India.

2. Materials and methods

2.1. Data source

The data used in this study is obtained from the Children's Recode (KR) dataset of the National Family Health Survey (NFHS-5), conducted from 2019 to 2021. This dataset includes one record for every child born to interviewed women within the five years preceding the survey consisting 232920 children. Unit-level data is accessible through the Demographic Health Survey (DHS) data repository, and requests for access can be made via the DHS Program website (www.dhsprogram.com/data/). The unit of analysis in this research encompasses 424 districts from LSDI states of India, including a total of 104692 children under the age of five years at the time of the survey.

2.2. Data pre-processing

The target variable in this study was stunting, defined according to the WHO standard as a height-for-age Z-score (HAZ) of less than -2 standard deviations (SD) (WHO, 2019). Various socioeconomic, demographic, and environmental factors were considered as features in the analysis. Missing instances for each variable included in the study were excluded from the analysis. After removing the missing observations, we were left with 82,158 cases. Further, to prepare categorical features for machine learning, the traditional one-hot encoding technique was utilized, in which multicategorical variables are transformed into several binary feature vectors. The continuous variables in the study were standardized.

2.3. Feature selection

Random Forest (RF) feature selection was employed to identify the most significant features. As suggested by Talukder and Ahammed (2020) for the application of RF feature selection in constructing predictive models for malnutrition. This method assigns an importance score to each feature, and those with scores below the average were excluded from the model. Figure 2 illustrates the importance scores associated with each feature.

2.4. Model training

The data was divided into a training dataset including 70% of the total data, and a testing dataset comprising the remaining 30%. We have employed five commonly used machine learning classifiers, namely, logistic regression (LR), random forest (RF), decision tree (DT), support vector classifier (SVC), and gradient boosting (Géron, 2022). The algorithms are implemented using scikit-learn library in Python (Pedregosa *et al.*, 2011).

2.5. Evaluation of model performance

The model's performance on the test set was evaluated using metrics like as accuracy, precision, recall, area under the curve (AUC), and F1 score.

Suppose an output variable has two classes, namely, positive and negative classes. A confusion matrix is a square matrix that includes the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (PN) values. These values can be used to calculate the one-dimensional performance measures (Luque *et al.*, 2019).

Accuracy: It is the ratio of events that have been correctly classified to the total number of cases in the dataset.

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

Precision: It measures the ratio of true positive correct predictions by the classifier out of all positive predictions

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

Recall: It measures the percentage of real positive predictions out of all actual positive instances in the dataset.

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

F_1 **score:** The F_1 score is computed by taking the harmonic mean of the precision and recall.

$$F_1 \text{ score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Area under the Curve (AUC): The area under the curve (AUC) is the are under the receiver operating characteristic (ROC) curve. ROC curve is a graphical representation of the true positive rate (TPR) vs the false positive rate (FPR) at different thresholds. The AUC provides an assessment of the overall model performance at every potential classification threshold. It measures the degree of separability between positive and negative classes. The value of AUC lies between 0 and 1. The higher the value, the better a model can distinguish between positive and negative classes.

3. Results

Table 1 shows that the prevalence of stunting in rural areas is 40.15%, which is higher compared to 32% in urban areas. Children whose mothers have no education exhibit the highest prevalence of stunting at 46.9%, followed by those with only primary education at 42.9%. Households lacking improved sanitation facilities and those using unclean cooking fuel have stunting prevalences of 45.3% and 41.6%, respectively. The prevalence of stunting is highest among Muslim children (40.7%), followed by Hindu children (38.3%) and those of other religions (37.1%). In Scheduled Castes, the prevalence is 40.5%. The average age of the mothers having stunted child is 27 years, while the average weight of the stunted children in the study is 10 kg.

Various machine learning models have been applied to predict the likelihood of stunting and the predictive performance of each classifier on the test and 5-fold cross-validation datasets. The results have been presented in Table 2.

Table 1: Prevalence of stunting in children under 5 in low sociodemographic index states of India by characteristics; NFHS 2019-2021

Variable	Category	Frequency / Mean	Percentage (%) / SD
Place of residence	Urban	4933	32
	Rural	26735	40.1
Highest educational level	No education	11291	46.9
	Primary	4966	42.9
	Secondary	13003	35.5
	Higher	2409	24.5
Sanitation facility (Improved)	No	12128	45.3
	Yes	19541	35.3
Clean Cooking fuel	No	21816	41.6
	Yes	9852	33.2
Religion	Hindu	26964	38.3
	Muslim	4199	40.7
	Others	505	37.1
Caste	SC	8274	43.2
	ST	4086	40.8
	OBC	15484	37.8
	Other	3825	31.7
Media Exposure	No	13991	45.4
	Yes	17678	34.4
Wealth index	Poorest	12796	47.3
	Poorer	8095	40.4
	Middle	5077	36
	Richer	3485	29.8
	Richest	2216	23.9
Mother Anemia	No	11803	36.7
	Yes	19865	39.8
Birth Order	1	9558	34.6
	2	9629	37.2
	3	6159	41.6
	4 or more	6322	45.9
Sex of child	Male	16822	39.4
	Female	14846	37.6
Skilled Birth Attendant	No	5179	45.4
	Yes	26490	37.4
Institutional Births	No	5811	47
	Yes	25857	37
Delivery by Caesarean Section	No	28849	39.8
	Yes	2819	29.1
Size of child at birth	No	5072	37.8
	Yes	3695	43.2
Birth Weight less than 2.5 kg	No	25927	37.4
	Yes	5742	44.7
Infectious diseases in past 2 weeks	No	24944	38.4
	Yes	6725	39.3
Child immediately put on chest after the birth?	No	5473	37.3
	Yes	26196	38.8
Distance to health facility is a big problem	No	22449	37.6
	Yes	9220	41.1
Mother Age (in years)		27	5
Respondent's height (in cm)		149.9	5.99
Mother BMI		20.72	3.35
Mother haemoglobin level (g/dl)		11.24	1.58
Child's weight (in kg)		10.08	2.86

Table 2: Summary of classification model performance in predicting stunting

Model		Accuracy	Precision	Recall	f1-score	AUC
Logistic Regression	5-folds cross-validation	0.62±0.10	0.29±0.05	0.55±0.03	0.38±0.017	0.64±0.31
	Test data	0.633	0.298	0.560	0.389	0.658
Random Forest Classifier	5-folds cross-validation	0.65±0.10	0.39±0.12	0.58±0.14	0.47±0.12	0.68±0.11
	Test data	0.658	0.394	0.598	0.475	0.690
Support Vector Classifier	5-folds cross-validation	0.65±0.05	0.32±0.12	0.60±0.13	0.42±0.11	0.68±0.16
	Test data	0.657	0.332	0.617	0.432	0.682
Gradient Boosting Classifier	5-folds cross-validation	0.65±0.07	0.40±0.10	0.59±0.13	0.49±0.11	0.70±0.06
	Test data	0.665	0.412	0.608	0.492	0.702
Decision Tree	5-folds cross-validation	0.56±0.11	0.45±0.13	0.46±0.14	0.45±0.12	0.54±0.17
	Test data	0.575	0.473	0.459	0.466	0.557

It can be seen from the Table 2 that the Decision Tree classifier was the least accurate model on both the test and cross-validation sets, with accuracies of 57.5% and 56 ± 11%, respectively. In contrast, the Gradient Boosting Classifier was the most effective model, achieving accuracies of 66.5% on the test set and 65 ± 7% on the cross-validation set. The Gradient Boosting model also had the highest F1 score at 49.2%, while Logistic Regression had the lowest F1 score in both the testing and 5-fold cross-validation. The Random Forest and Support Vector Machine classifiers had accuracy scores of 65.8% and 65.7%, respectively, on the test dataset, with F1 scores of 47.5% and 43.2%, respectively (Table 2). The area under the curve (AUC) was highest in the Gradient Boosting model at 70%, followed by the Random Forest and Support Vector Classifier at 69% and 68%, respectively, and was the lowest in the Decision Tree model.

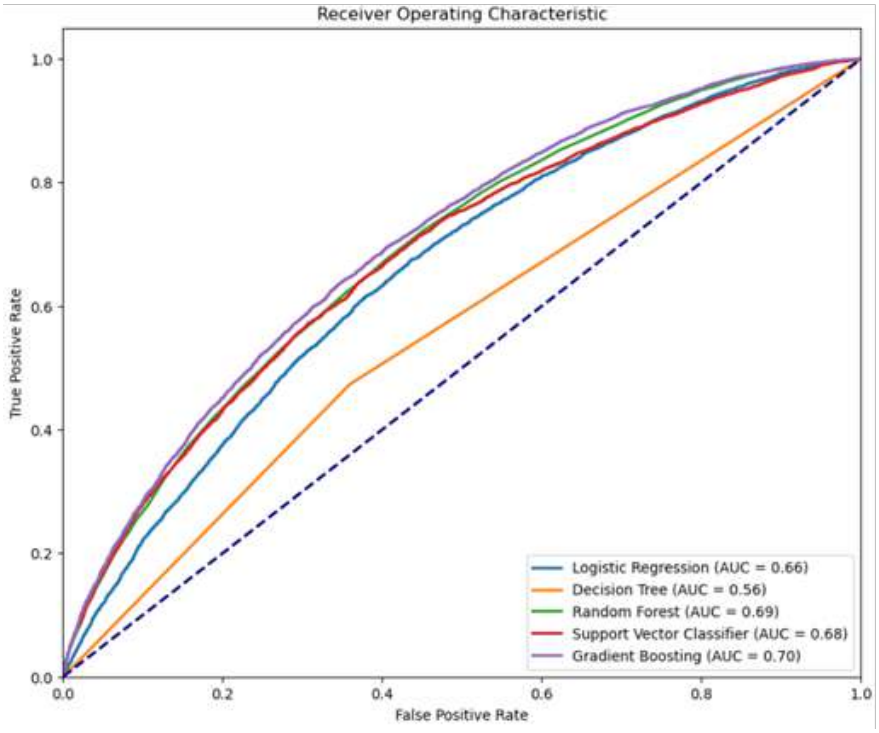


Figure 1: Receiver operating characteristic curve of stunting model

The ROC curve analysis revealed that the Gradient Boosting model achieved the highest Area Under the Curve (AUC) at 0.70, indicating superior performance in distinguishing between the positive and negative classes, followed by the Random Forest (AUC =

0.69) and Support Vector Classifier (AUC = 0.68). Logistic Regression and Decision Tree had lower AUC values of 0.66 and 0.56, respectively. The Decision Tree model performed the worst showing the lowest precision and recall values, indicating it was less effective in handling imbalanced datasets. These results underscore the robustness of the Gradient Boosting model in providing a balanced trade-off between precision and recall, making it the most reliable model among those evaluated (Figure 1).

Further, the importance scores of the determinants in the Random Forest model are presented in Figure 2.

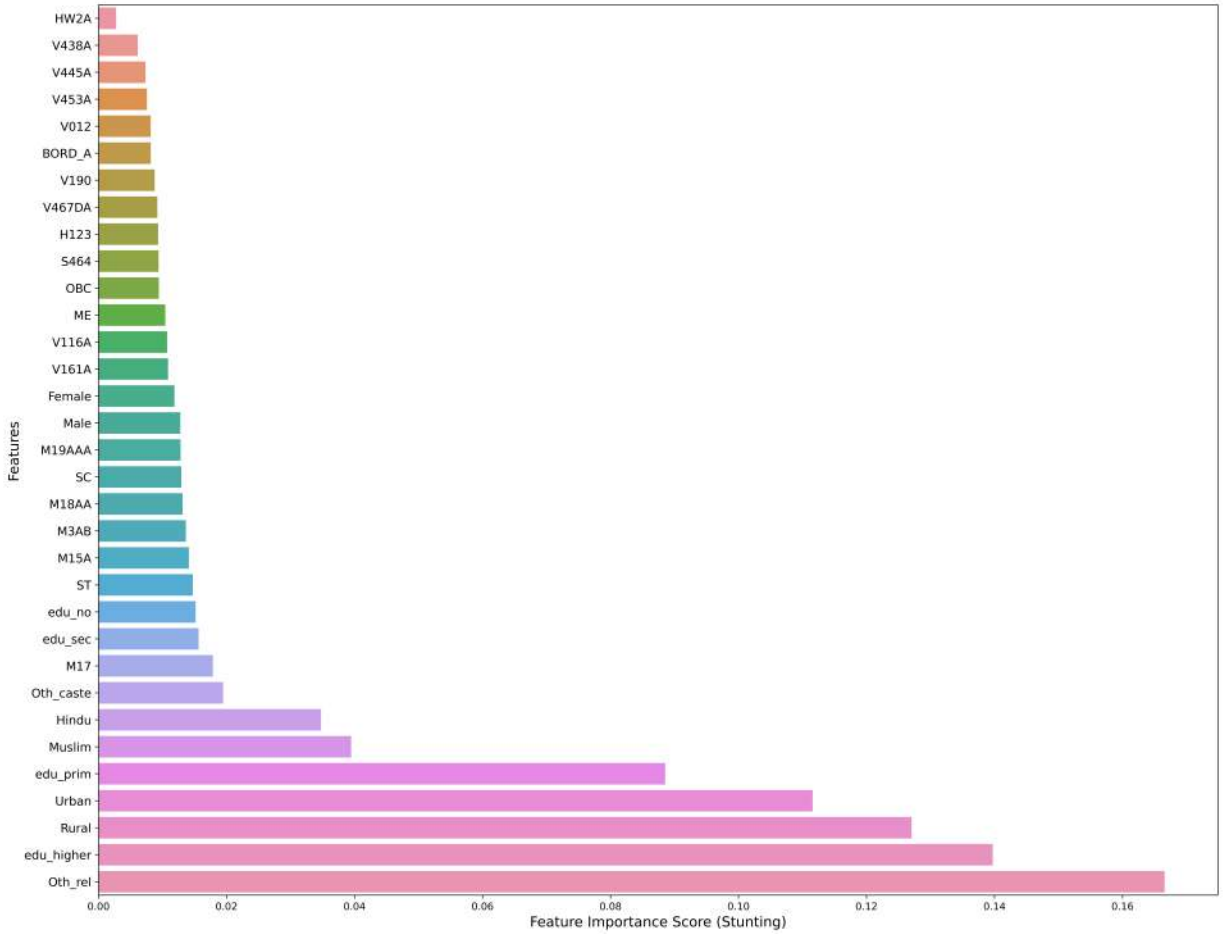


Figure 2: Features importance score

From the Figure 2, the key determinants of stunting can be identified, with the top variables being "Other" Religion (0.16), "higher education of mother" (0.13), and "Rural" place of residence (0.13). Other significant factors include "mother primary education" (0.10), and "Other caste" (0.07). These results highlight the critical roles of education level, religious affiliation, and place of residence in influencing stunting outcomes. Understanding these variables can guide targeted interventions to address and reduce stunting in affected populations.

4. Discussion

Our study found that the Gradient Boosting Classifier had the highest predictive accuracy and precision for stunting among children under five years in low socio-demographic index states of India, using the fifth round of NFHS data. The Gradient Boosting Classifier achieved the largest Area Under the Curve (AUC), suggesting its superior ability to distinguish between positive and negative classes. The Random Forest also performed well, although not as well as the Gradient Boosting Classifier.

In addition, our study has revealed significant factors for predicting stunting in children under the age of five in low socio-demographic index states of India. Some of the identified top features include religions other than Hindu and Muslim, mother higher education level, and a rural place of residence. Other significant factors include mother primary education and other castes (other than SC/ST and OBC). These were identified commonly as predictors of stunting in other studies also (Mzumara *et al.*, 2018; Shahriar *et al.*, 2019; Mediani, 2020; Bwalya *et al.*, 2015).

An important advantage of this study is its use of the NFHS dataset. The NFHS employs a robust sampling procedure, providing a reliable representation of the under-five population in both rural and urban areas of low socio-demographic index states in India. Additionally, the model's accuracy was assessed using test data, and 5-fold cross-validation was employed to prevent overfitting. However, there are still certain possible constraints that need to be considered. Consequently, the interpretability of the findings remains constrained. Furthermore, despite our efforts to incorporate a wide range of covariates, we were unable to eliminate the possibility of residual confounding resulting from unmeasured factors, such as the mother's height and weight. Additionally, certain details about the children were obtained from their mothers' memories, such as instances of diarrhoea and fever in the past two weeks. However, it is possible that there is a bias in these recollections.

5. Conclusion

In this study, several machine learning models have been employed to predict the prevalence of stunting in children under 5 years of age in LSDI states on India. The performance of the models was compared using various performance metrics. The results suggest that the Gradient Boosting Classifier model has highest predictive accuracy for stunting compared to the other applied models in this study. Feature importance is also studied using the random forest model. It suggests that social structure, mother education are the important predictors of stunting among the children under five in low socio demographic index states of India.

References

Bwalya, B. B., Lemba, M., Mapoma, C. C., and Mutombo, N. (2015). Factors associated with stunting among children aged 6-23 months in Zambian: evidence from the 2007 Zambia demographic and health survey. *International Journal of Advanced Nutritional and Health Science*, **3**, 116–31.

Das, S. and Gulshan, J. (2017). Different forms of malnutrition among under five children in Bangladesh: a cross sectional study on prevalence and determinants. *BMC Nutrition*, **3**, 1–12.

Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Global Burden of Disease Collaborative Network (2020). *Global Burden of Disease Study 2019 (GBD 2019) Socio-Demographic Index (SDI) 1950-2019*. Institute for Health Metrics and Evaluation, Seattle, USA.

Iniesta, R., Stahl, D., and McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, **46**, 2455–2465.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., and Qiang, B. (2020). RFRSF: Employee turnover prediction based on random forests and survival analysis. In Huang, Z., Beek, W., Wang, H., Zhou, R., and Zhang, Y., editors, *Web Information Systems Engineering – WISE 2020*, pages 503–515. Springer International Publishing, Cham.

Jose, S., Bheemeshwar, R. A., and Agrawal, M. (2018). Child undernutrition in India. *Economic and Political Weekly*, **53**, 63–70.

Luque, A., Carrasco, A., Martín, A., and de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, **91**, 216–231.

Markos, Z., Doyore, F., Yifiru, M., and Haidar, J. (2014). Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey. *Journal of Health & Medical Informatics*, **5**, 152.

Mediani, H. S. (2020). Predictors of stunting among children under five year of age in Indonesia: a scoping review. *Global Journal of Health Science*, **12**, 83–95.

Mzumara, B., Bwembya, P., Halwiindi, H., Mugode, R., and Banda, J. (2018). Factors associated with stunting among children below five years of age in Zambia: evidence from the 2014 Zambia demographic and health survey. *BMC Nutrition*, **4**, 51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, **12**, 2825–2830.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., and Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, **56**, 455.

Rakotomanana, H., Gates, G. E., Hildebrand, D., and Stoecker, B. J. (2017). Determinants of stunting in children under 5 years in Madagascar. *Maternal & Child Nutrition*, **13**, e12409.

Shahriar, M. M., Iqubal, M. S., Mitra, S., and Das, A. K. (2019). A Deep Learning Approach to Predict Malnutrition Status of 0-59 Month's Older Children in Bangladesh. In *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 145–149. IEEE.

Talukder, A. and Ahammed, B. (2020). Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*, **78**, 110861.

UNICEF (2020). *Nutrition, for Every Child UNICEF Nutrition Strategy 2020-2030*. UNICEF, New York.

WHO (2019). Nutrition Landscape Information System (NLIS) country profile indicators: interpretation guide. Technical report, World Health Organization.

Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 207-220

V. K. Gupta Endowment Award Lecture 2024: Modernizing Linear Mixed Model Prediction

J. Sunil Rao
Division of Biostatistics, University of Minnesota, Twin Cities, USA

Received: 30 April 2024; Revised: 18 August 2024; Accepted: 20 August 2024

Abstract

Mixed models have widespread appeal in many areas of statistical modeling from biostatistics to small area estimation. Here we review a variety of recent approaches for modernizing linear mixed model prediction including robust prediction via the observed best predictor (OBP) to prediction for new test data using a classified random effect, namely classified mixed model prediction (CMMP). Finally, a brief mention will be made to a proposal for using mixed model prediction to project outside of the range of the training data using classified mixed model projections.

Key words: Mixed model selection; GIC; Fence method; Small area estimation; Subareas.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

It is now well accepted that it is necessary to *borrow strength* from relevant domains and resources to increase the efficiency of direct estimates and that linear mixed models provide a pathway to do so. In this regard, the empirical best linear unbiased prediction, or EBLUP, method has had a dominant influence (*e.g.*, Rao 2003; Jiang and Lahiri 2006). The method utilizes a linear mixed effects model (*e.g.*, Jiang 2007) in order to borrow strength. The standard procedure of computing the EBLUP is the following. First one derives the best predictor (BP) of the mixed effects of interests, such as the small area means. Then, one replaces the vector of the fixed effects by its maximum likelihood estimator (MLE), assuming that the variance components are known (up to this stage one obtains the best linear unbiased predictor, or BLUP). Finally, one replaces the unknown variance components by their ML or REML estimators. It follows that the EBLUP is the BP, in which the unknown fixed parameters, including the fixed effects and variance components, are estimated either by ML or REML. The latter are known to be asymptotically optimal under estimation considerations (*e.g.*, Jiang 2007). However, in many cases, such as in SAE, the problem of main interest is prediction, rather than estimation. The implication is that the EBLUP may be regarded as a hybrid of optimal prediction (*i.e.*, BP) and optimal estimation (*e.g.*, ML). Nevertheless, if prediction is of main interest, it would be more natural to have a purely

Corresponding Author: J. Sunil Rao
Email: js-rao@umn.edu

predictive procedure, in which both the predictor and estimator are derived from predictive considerations.

2. A general framework for the observed best predictor (OBP)

First, consider a general mixed model prediction problem (*e.g.*, Robinson 1991). The assumed model is $y = X\beta + Zv + e$, where X, Z are known matrices; β is a vector of fixed effects; v, e are vectors random effects and errors, respectively, such that $v \sim N(0, G)$, $e \sim N(0, \Sigma)$, and v, e are uncorrelated. Suppose that the true underlying model is $y = \mu + Zv + e$, where $\mu = \mathrm{E}(y)$. Here, again, E without subscript represents expectation with respect to the true distribution, which may be unknown but is not model-dependent. Following Jiang *et al.* (2011), our interest is prediction of a vector of mixed effects that can be expressed as $\theta = F'\mu + R'v$, where F, R are known matrices. Suppose that G, Σ are known. Then, the best predictor (BP) of θ, in the sense of minimum MSPE, under the assumed model is given by $\mathrm{E_a}(\theta|y) = F'\mu + R'\mathrm{E_a}(v|y) = F'X\beta + R'GZ'V^{-1}(y - X\beta)$, where $\mathrm{E_a}$ denotes expectation under the assumed model, $V = \mathrm{Var}(y) = \Sigma + ZGZ'$ and β is the true vector of fixed effects, under the assumed model. If we write $B = R'GZ'V^{-1}$ and $\Gamma = F' - B$, then the BP can be expressed as

$$\mathrm{E_a}(\theta|y) \;=\; F'y - \Gamma(y - X\beta). \tag{1}$$

Now let $\breve{\theta}$ denote the right side of (1) with a fixed, but arbitrary β. Then, it can be shown that $\mathrm{MSPE}(\breve{\theta}) = \mathrm{E}(I_1 - 2I_2 + (y - X\beta)'\Gamma'\Gamma(y - X\beta)\}$, where I_1, I_2 do not depend on β. Thus, the best predictive estimator (BPE) (Jiang *et al.* 2011) of β is obtained by minimizing the expression inside the expectation, that is, $\tilde{\beta} = (X'\Gamma'\Gamma X)^{-1}X'\Gamma'\Gamma y$, assuming that $\Gamma'\Gamma$ is nonsingular and X is full rank. Once the BPE is obtained, the OBP of θ (Jiang *et al.* 2011), is given by the right side of (1) with β replaced by $\tilde{\beta}$. On the other hand, the BLUP of θ is given by the right side of (1) with β replaced by $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$, which is the MLE of β. To compare the predictive performance of the OBP and BLUP, let us consider a class of empirical best predictors (EBPs) that can be expressed as

$$\breve{\theta} \;=\; F'y - \Gamma(y - X\breve{\beta}), \tag{2}$$

where $\breve{\beta}$ is a weighted least squares (WLS) estimator of β expressed as $\breve{\beta} = (X'WX)^{-1}X'Wy$ and W is a positive definite weighting matrix. Note that (2) is the BP (1) with β replaced by $\breve{\beta}$ (and hence explains the name EBP). Also note that the BPE and MLE are special cases of the WLS, hence the OBP and BLUP are special cases of the EBP.

2.1. Special Case 1: The Fay-Herriott Model

The Fay-Herriot model (Fay and Herriot 1979) is widely used in small area estimation (SAE). It was proposed to estimate the per-capita income of small places with population size less than 1,000. The model can be expressed as a mixed effects model:

$$y_i \;=\; x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m, \tag{3}$$

where x_i is a vector of known covariates, β is a vector of unknown regression coefficients, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that the v_i's

and e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. The variance A is unknown, but the sampling variances D_i's are assumed known. The problem of interest is estimation of the small area means, which, at least by a higher-order approximation, are equal to the mixed effects $\theta_i = x_i'\beta + v_i, i = 1, \ldots, m$. Thus, without loss of generality, we treat the θ_i's as the small area means, so the problem is prediction of the mixed effects.

One should note that the true small area means should not depend on the assumed model. In fact, it is easy to see that under the weak assumption

$$y_i = \mu_i + v_i + e_i, \quad \text{we have} \quad \theta_i = \mathrm{E}(y_i|v_i) = \mathrm{E}(y_i) + v_i, \quad i = 1, \ldots, m, \tag{4}$$

where $\mu_i = \mathrm{E}(y_i)$. The advantage of expressions (4) is that they are not model-dependent, which is a key to our approach. Here, E denotes the expectation with respect to the true distribution of y_i, which may be unknown, but not model-dependent. The most popular precision measure of a predictor is its mean squared prediction error (MSPE; *e.g.*, Prasad & Rao 1990, Das *et al.* 2004). Write $\theta = (\theta_i)_{1\le i\le m}$ and let $\tilde{\theta} = (\tilde{\theta}_i)_{1\le i\le m}$ be a predictor of θ. Then, the (overall) MSPE of $\tilde{\theta}$ is given by

$$\mathrm{MSPE}(\tilde{\theta}) = \mathrm{E}(|\tilde{\theta} - \theta|^2) = \sum_{i=1}^{m} \mathrm{E}(\tilde{\theta}_i - \theta_i)^2. \tag{5}$$

Once again, the expectation in (5) is with respect to the true underlying distribution (of whatever random quantities that are involved), which is unknown but not model-dependent. Under the assumed Fay-Herriot model, and given the parameters $\psi = (\beta', A)'$, the BP is given by

$$\tilde{\theta}(\psi) \;=\; \mathrm{E}_{\mathrm{m},\psi}(\theta|y) \;=\; \left[x_i'\beta + \frac{A}{A + D_i}(y_i - x_i'\beta)\right]_{1\le i\le m}, \tag{6}$$

or $\tilde{\theta}(\psi)_i = x_i'\beta + B_i(y_i - x_i'\beta), 1 \le i \le m$, where $B_i = A/(A + D_i)$, and $\mathrm{E}_{\mathrm{m},\psi}$ represents (conditional) expectation under the assumed model with ψ being the true parameter vector. Note that $\mathrm{E}_{\mathrm{m},\psi}$ is different from E unless the model is correct, and ψ is the true parameter vector. For simplicity, let us assume, for now, that A is known. Then, the precision of $\tilde{\theta}(\psi)$, which is now denoted by $\tilde{\theta}(\beta)$ because A is no longer a parameter, is measured by

$$\mathrm{MSPE}\{\tilde{\theta}(\beta)\} = \sum_{i=1}^{m} \mathrm{E}\{B_i y_i - \theta_i + x_i'\beta(1 - B_i)\}^2 = I_1 + 2I_2 + I_3, \tag{7}$$

where $I_1 = \sum_{i=1}^m \mathrm{E}(B_i y_i - \theta_i)^2$, $I_2 = \sum_{i=1}^m x_i'\beta(1 - B_i)\mathrm{E}(B_i y_i - \theta_i)$, and $I_3 = \sum_{i=1}^m (x_i'\beta)^2(1 - B_i)^2$. Note that I_1 does not depend on β. As for I_2, we have $\mathrm{E}(B_i y_i - \theta_i) = (B_i - 1)\mathrm{E}(y_i)$. Thus, we have $I_2 = -\sum_{i=1}^m (1 - B_i)^2 x_i'\beta \mathrm{E}(y_i)$. It follows that the left side of (7) can be expressed as

$$\mathrm{MSPE}\{\tilde{\theta}(\beta)\} = \mathrm{E}\left\{I_1 + \sum_{i=1}^{m}(1 - B_i)^2(x_i'\beta)^2 - 2\sum_{i=1}^{m}(1 - B_i)^2 x_i'\beta y_i\right\}. \tag{8}$$

The right side of (8) suggests a natural estimator of β, by minimizing the expression inside the expectation, which is equivalent to minimizing $Q(\beta) = \sum_{i=1}^m (1 - B_i)^2(x_i'\beta)^2 - 2\sum_{i=1}^m (1 -$

$B_i)^2 x_i' \beta y_i = \beta' X' \Gamma^2 X \beta - 2y' \Gamma^2 X \beta$, where $X = (x_i')_{1 \le i \le m}$, $y = (y_i)_{1 \le i \le m}$ and $\Gamma = \text{diag}(1 - B_i, 1 \le i \le m)$. A closed-form solution is given by

$$\tilde{\beta} = (X'\Gamma^2 X)^{-1} X'\Gamma^2 y = \left\{ \sum_{i=1}^{m} (1 - B_i)^2 x_i x_i' \right\}^{-1} \sum_{i=1}^{m} (1 - B_i)^2 x_i y_i. \tag{9}$$

Here we assume, without loss of generality, that X is of full column Note that $\tilde{\beta}$ is different from the MLE of β,

$$\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1} y = \left(\sum_{i=1}^{m} \frac{x_i x_i'}{A + D_i} \right)^{-1} \sum_{i=1}^{m} \frac{x_i y_i}{A + D_i}, \tag{10}$$

where $V = \text{diag}(A + D_i, 1 \le i \le m) = \text{Var}(y)$. While $\hat{\beta}$ maximizes the likelihood function, $\tilde{\beta}$ minimizes the "observed" MSPE which is the expression inside the expectation on the right side of (8). We call $\tilde{\beta}$ given by (9) the *best predictive estimator*, or BPE, of β. Note that the BPE has the property that its expected value,

$$\text{E}(\tilde{\beta}) \;=\; (X'\Gamma^2 X)^{-1} X'\Gamma^2 \text{E}(y), \tag{11}$$

is the β that minimizes $\text{MSPE}\{\tilde{\theta}(\beta)\}$. However, the expression (11) is not computable.

A predictor of the mixed effects θ is then obtained by replacing β in the BP (6) by its BPE. We call this predictor the *observed best predictor*, or OBP. The reason is that the BPE is the minimizer of the observed MSPE. If the observed MSPE were the true MSPE, the BPE would give us the BP. However, because, instead, we are dealing with the observed MSPE, the corresponding predictor (obtained by the same procedure with the MSPE replaced by the observed MSPE) should be called the observed BP.

2.2. Special Case 2: The nested error regression model

Consider sampling from finite subpopulations $P_i = \{Y_{ik}, k = 1, \ldots, N_i\}, i = 1, \ldots, m$. Suppose that auxiliary data $X_{ikl}, k = 1, \ldots, N_i, l = 1, \ldots, p$ are available for each P_i, and a super-population nested-error regression model (Battese *et al.* 1988) hold for all subpopulations:

$$Y_{ik} \;=\; X_{ik}'\beta + v_i + e_{ik}, \quad k = 1, \ldots, N_i, \tag{12}$$

where $X_{ik} = (X_{ikl})_{1 \le l \le p}$, the v_i's are small-area specific random effects, and e_{ik}'s are additional errors, such that the random effects and errors are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ik} \sim N(0, \sigma_e^2)$. The small area mean for P_i is then $\mu_i = N_i^{-1} \sum_{k=1}^{N_i} Y_{ik}$.

Now suppose that $y_{ij}, j = 1, \ldots, n_i$ are observed for the ith subpopulation, $i = 1, \ldots, m$. Let the corresponding auxiliary data be $x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m$. Write $y_i = (y_{ij})_{1 \le j \le n_i}$, $y = (y_i)_{1 \le i \le m}$, $\bar{y}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $\bar{x}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$. Let $\psi = (\beta', \sigma_v^2, \sigma_e^2)'$ denote the vector of parameters under the nested-error regression model (12). Under this model with ψ being the true parameter vector, the BP for μ_i is

$$\begin{aligned} \tilde{\mu}_i(\psi) \;&=\; \text{E}_{\text{m},\psi}(\mu_i | y) = \frac{1}{N_i} \left\{ \sum_{j=1}^{n_i} y_{ij} + \sum_{k \notin I_i} \text{E}_{\text{m},\psi}(Y_{i,k} | y_i) \right\} \\ &=\; \bar{X}_i'\beta + \left\{ \frac{n_i}{N_i} + \left(1 - \frac{n_i}{N_i}\right) \frac{n_i \sigma_v^2}{\sigma_e^2 + n_i \sigma_v^2} \right\} (\bar{y}_{i\cdot} - \bar{x}_{i\cdot}'\beta), \end{aligned} \tag{13}$$

where $\mathrm{E}_{\mathrm{m},\psi}$ denotes the model-based conditional expectation given that ψ is the true parameter vector, I_i is the set of sampled indexes such that Y_{ik} is in the sample iff $k \in I_i$, and $\bar{X}_i = N_i^{-1}\sum_{k=1}^{N_i} X_{ik}$ is the subpopulation mean of the X_{ik}'s for the ith subpopulation (which is known). Note that (13) is a model-based BP. The performance of the model-based BP is then evaluated by the design-based MSPE. This is because the latter is almost free of model assumptions, and therefore robust to model misspecifications [we could not do this under the Fay-Herriot model, however, because the sampling data were not available at the unit level; instead, we considered a model under very weak assumptions]. The design-based MSPE is given by $\mathrm{MSPE}\{\tilde{\mu}(\psi)\} = \mathrm{E}_{\mathrm{d}}\{|\tilde{\mu}(\psi) - \mu|^2\} = \sum_{i=1}^m \mathrm{E}_{\mathrm{d}}\{\tilde{\mu}_i(\psi) - \mu_i\}^2$, where $\tilde{\mu}(\psi) = [\tilde{\mu}_i(\psi)]_{1\le i\le m}$, $\mu = (\mu_i)_{1\le i\le m}$ and E_{d} denotes the design-based expectation. Assume, for simplicity, simple random sampling within each subpopulation P_i. Then, it can be shown that

$$\mathrm{MSPE} = \mathrm{E}_{\mathrm{d}}\left[\sum_{i=1}^m \{\tilde{\mu}_i^2(\psi) - 2a_i(\sigma_v^2,\sigma_e^2)\bar{X}_i'\beta\bar{y}_{i\cdot} + b_i(\sigma_v^2,\sigma_e^2)\hat{\mu}_i^2\}\right], \tag{14}$$

where $a_i(\sigma_v^2,\sigma_e^2) = (1-n_i/N_i)\sigma_e^2/(\sigma_e^2+n_i\sigma_v^2)$ and $b_i(\sigma_v^2,\sigma_e^2) = 1-2[n_i/N_i+(1-n_i/N_i)\{n_i\sigma_v^2/(\sigma_e^2+n_i\sigma_v^2)\}]$. Thus, the BPE of ψ is obtained by minimizing $Q(\psi) = \sum_{i=1}^m\{\tilde{\mu}_i^2(\psi)-2a_i(\sigma_v^2,\sigma_e^2)\bar{X}_i'\beta\bar{y}_{i\cdot}+b_i(\sigma_v^2,\sigma_e^2)\hat{\mu}_i^2\}$, which is the expression inside the expectation in (14). Here $\hat{\mu}_i^2$ is a design-based unbiased estimator of μ_i^2, given by $\hat{\mu}_i^2 = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}^2 - (N_i-1)\{N_i(n_i-1)\}^{-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$. A similar numerical procedure can be developed to compute the BPE. Given the BPE $\tilde{\psi} = (\tilde{\beta}', \tilde{\sigma}_v^2, \tilde{\sigma}_e^2)'$, the OBP of μ_i is given by $\tilde{\mu}_i = \tilde{\mu}_i(\tilde{\psi})$, $1 \le i \le m$, where $\tilde{\mu}_i(\psi)$ is given by (13).

2.3. Estimation of MSPE of OBP

Obtaining a measure of uncertainty for OBP is particularly challenging. This is because the OBP is derived by taking into account of the potential model misspecification. Therefore, to derive a measure of uncertainty, the potential model misspecification also needs to be taken into consideration when considering measures of uncertainty. More importantly, it is desirable to evaluate uncertainty due to the potential model misspecification.

A standard measure of uncertainty is the MSPE. Let us first consider this under a Fay-Herriot model. In proposing the OBP, Jiang *et al.* (2011) also proposed an MSPE estimator under potential model misspecification, which we call the JNR estimator in the sequel. The authors showed that the JNR estimator is second-order unbiased, that is, its bias is $o(m^{-1})$. However, the estimator is known to have large variation. To see why, note that the leading term of the JNR estimator has the expression

$$(\hat{\theta}_i - y_i)^2 + D_i(2\hat{B}_i - 1), \tag{15}$$

where $\hat{\theta}_i$ is the OBP of θ_i, $B_i = A/(A+D_i$, and $\hat{B}_i$ is B_i with A replaced by $\hat{A}$, the BPE of A. The direct estimator, y_i, is involved in (15), which has large variation compared to, for example, $\hat{A}$. The latter is an estimator based on all of the data, $y_i, x_i, 1 \le i \le m$, which has relatively small variation. In particular, the JNR estimator has a significantly nonzero chance of taking negative values.

In addition to the JNR estimator, Jiang *et al.* (2011) also proposed a bootstrap MSPE estimator. Although the bootstrap estimator is gauranteed non-negative, its bias was

shown to be significantly larger than the JNR estimator. The method also seemed lack of theoretical justification, in which the bootstrap samples were drown independently under the model $y_i^* \sim \hat{\theta}_i + e_i^*$, where $\hat{\theta}_i$ is the OBP and $e_i^* \sim N(0, D_i)$, $1 \leq i \leq m$.

Liu *et al.* (2022a) proposed a OBOR estimator for the MSPE of OBP. Here, OBOR is an abbreviation of one-bring-one-route. It is called OBOR because the estimator consists of averages of terms, where each term involves y_i, plus one other y_j for $j \neq i$. The average is over $m-1$ such y_j's for $j \neq i$. The idea can be generalized to one-bring-two, one-bring-three, etc., but the computational burden mounts as this moves on. In this regard, the JNR estimator may also be viewed as a special case of one-bring-none. Although the OBOR estimator reduces the variation over the JNR estimator, the result was not all satisfactory, compared to a much better estimator found later.

A well-known method for obtaining a second-order unbiased MSPE estimator is the Prasad-Rao (PR) linearization method (Prasad and Rao 1990). The method is developed under the assumption that the underlying model is correct. In fact, the assumed model is substantially used in the derivation of the P-R MSPE estimator. Given that, it would be surprising to learn that, in spite of the model misspecification, the PR MSPE estimator for OBP is, still, mostly correct. In fact, Liu *et al.* (2022b) found that the PR MSPE estimator remains first-order unbiased in the sense that the bias of the estimator is $O(m^{-1})$, even if the underlying model is misspecified in its mean function. Furthermore, the same authors showed the PR MSPE estimator can be modified to achieve the second-order unbiasedness, again under the potential model misspecification in the mean function.

3. Classified mixed model prediction (CMMP)

The world has been witnessing an information explosion in many areas of society from medicine to economics and business to social media for instance. The rapid increase in the unprecedented amount of data has resulted in many new important shifts of interest in the types of questions that can be potentially answered. These new shifts are focusing more and more attention on knowledge at individual or subject levels. One of the currently "hot" areas is *precision medicine*. The National Research Council of the United States in 2014 defined the latter as the "ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those disease they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not". Another area, in economic studies, is *family economics*, which applies basic economic concepts to families, which are viewed as (small) firms or companies. For example, China Household Finance Survey, the largest non-governmental household panel survey since 2009, has so far collected massive financial and economical data at household level. The latest wave, conducted in the summer of 2017, had more than 40,000 nationally and provincially representative households. More than 10,000 registered users worldwide are using the data for their studies about China. In particular, the data provide important information about household finance, which is a driving force of China's national economy (*e.g.*, Zhang *et al.* 2014, Gan, Yin and Tan 2016).

The target of classical statistical inference is a (large) population, from which data are collected, and to be used to make inference about the same population parameters, such

as the mean, proportion, and regression coefficients. However, in each of the above subject matter disciplines, the primary interest is inference at the subject levels. For example, in precision medicine, the subject may be a patient, or small group of patients sharing similar characteristics; in family economics, the subject is typically a family, whose definition may vary depending on factors such as culture or interest.

Nevertheless, it should be noted that making inference about a specific subject does not mean that the inference is based only on data collected from the subject, which is key. The idea can be best explained using a mixed effects model (MEM; *e.g.*, Jiang 2007). There are fixed effects and random effects in a MEM. The fixed effects are parameters that are common for all of the subjects; the random effects are typically subject-specific. The fixed effects are estimated using all of the data, that is, combining all subjects, but what about the random effects? This question has direct implications on how one predicts accurately from such models (*e.g.* Welham *et al.* 2004). In mixed model prediction (MMP), the best predictor for a characteristic of interest associated with a specific subject is derived, which depends on the subject-specific data as well as the fixed effects and variance components, which are population parameters. Thus, through MMP, inference about the subject-specific characteristic borrows strength from data from other subjects, as well as information from other sources.

In a significant recent development toward potentially much broader, modern-time applications, Jiang *et al.* (2018) proposed a method called classified mixed model prediction (CMMP) for two types of prediction problems - predicting the mixed effect associated with a set of new observations and predicting future values associated with new sets of covariates. The basic idea is to create a "match" between a group or cluster in the population, for which one wishes to make prediction, and a (massive) training data, with known groups or clusters. Once such a match is built, the traditional MMP method can be utilized to make accurate predictions. Even more interestingly, it can handle the situation where a real match may not exist.

To illustrate the CMMP method, let us focus on prediction of a mixed effect associated with new observations. Suppose that we have a set of training data, $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$ in the sense that their classifications are known, that is, one knows which group, i, that y_{ij} belongs to. The assumed model is a linear mixed model (LMM; *e.g.*, Jiang 2007):

$$y_i = \mathrm{E}(y_i|\alpha) + \epsilon_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \tag{16}$$

where $y_i = (y_{ij})_{1\le j\le n_i}$, $X_i = (x'_{ij})_{1\le j\le n_i}$ is a matrix of known covariates, β is a vector of unknown regression coefficients (the fixed effects), Z_i is a known $n_i \times q$ matrix, α_i is a $q \times 1$ vector of group-specific random effects, ϵ_i is an $n_i \times 1$ vector of errors, and $\alpha = (\alpha_i)_{1\le i\le m}$. It is assumed that the α_i's and ϵ_i's are independent, with $\alpha_i \sim N(0, G)$ and $\epsilon_i \sim N(0, R_i)$, where the covariance matrices G and R_i depend on a vector ψ of dispersion parameters, or variance components. Our goal is to make a classified prediction for a mixed effect associated with a new observation, y_{n}. Suppose that

$$y_{\mathrm{n}} = \mathrm{E}(y_{\mathrm{n}}|\alpha) + \epsilon_{\mathrm{n}} = x'_{\mathrm{n}}\beta + z'_{\mathrm{n}}\alpha_{\mathrm{I}} + \epsilon_{\mathrm{n}}, \tag{17}$$

where $x_{\mathrm{n}}, z_{\mathrm{n}}$ are known vectors, $I \in \{1, \ldots, m\}$ but one does not know which element i,

$1 \le i \le m$, is equal to I. Furthermore, $\epsilon_{\rm n}$ is a new error that is independent of $y_i, 1 \le i \le m$, and has mean zero. The mixed effect of interest is $\theta = {\rm E}(y_{\rm n}|\alpha) = y_{\rm n} - \epsilon_{\rm n}$. From the training data, one can estimate the parameters, β and ψ. Thus, we can assume that estimators $\hat{\beta}, \hat{\psi}$ are available for β, ψ, respectively. Suppose that $I = i$. Then, it can be shown that ${\rm E}(\theta|y_1, \ldots, y_m) = {\rm E}(\theta|y_i)$ and, according to the normal theory,

$$
{\rm E}(\theta|y_i) = x_{\rm n}'\beta + z_{\rm n}' G Z_i' (R_i + Z_i G Z_i')^{-1}(y_i - X_i\beta). \tag{18}
$$

The right side of (18) is the BP under the assumed LMM, if the true parameters, β and ψ, are known. Because the latter are unknown, we replace them by $\hat{\beta}$ and $\hat{\psi}$, respectively. The result is called an empirical best predictor (EBP), noted by $\tilde{\theta}_{(i)}$. In practice, however, I is unknown. In order to identify I, we consider the MSPE of predicting θ by the BP when I is classified as i, that is ${\rm MSPE}_i = {\rm E}\{\tilde{\theta}_{(i)} - \theta\}^2 = {\rm E}\{\tilde{\theta}_{(i)}^2\} - 2{\rm E}\{\tilde{\theta}_{(i)}\theta\} + {\rm E}(\theta^2)$. Using the expression $\theta = y_{\rm n} - \epsilon_{\rm n}$, we have ${\rm E}\{\tilde{\theta}_{(i)}\theta\} = {\rm E}\{\tilde{\theta}_{(i)}y_{\rm n}\} - {\rm E}\{\tilde{\theta}_{(i)}\epsilon_{\rm n}\} = {\rm E}\{\tilde{\theta}_{(i)}y_{\rm n}\}$. Thus, we have

$$
{\rm MSPE}_i = {\rm E}\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}y_{\rm n} + \theta^2\}. \tag{19}
$$

Note that the E in (19) denotes the true expectation, which may be unknown; nevertheless, the observed MSPE corresponding to (4) is the expression inside the expectation. Therefore, a natural idea is to identify I as the index i that minimizes the observed MSPE. Because θ^2 does not depend on i, this is equivalent to

$$
I = {\rm argmin}_i\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}y_{\rm n}\}. \tag{20}
$$

Denote the I identified by (20) by $\hat{I}$. Then, the classified mixed-effect predictor (CMMP) of θ is given by $\hat{\theta} = \tilde{\theta}_{(\hat{I})}$.

The basic idea of CMMP has been extended to multiple new observations from the same, unknown group, and to prediction of a future observation. See Jiang *et al.* (2018) for details. An important concept being exploited by CMMP is the idea that it *captures what is not captured by the fixed effect (the uncaptured)* through the classified random effect. It is important to note that the primary interest is not to identify the correct "match", I. In fact, in many applications such a match may not exist, that is, there is no group among the training that matches exactly the group corresponding to the new observations. Even if the exact match does exist, as the number of training data groups, m, increases, the probability of identifying the correct group, that is, ${\rm P}(\hat{I} = I)$, goes to zero (as opposed to going to one, as one might expect). But, regardless, the CMMP of the mixed effect, θ, is consistent (in fact, converges in L^2 to the true mixed effect), which is all we care about. For example, it was demonstrated that CMMP significantly outperform the traditional regression prediction whether or not the true match exists. The rationale behind the mismatch-led-correct-prediction is because, as m increases, the difference between different groups becomes smaller and smaller; thus, even though there is no exact match, there is a "close match" between one of the training data groups and the new group, of which CMMP is able to take advantage. This important, and interesting, feature makes the CMMP idea practically more attractive because, in practice, an exact match may not exist but a close resemblance may well be expected.

Following the initial work of CMMP, Sun *et al.* (2018) extended the idea to classified prediction of mixed effects associated with binary outcomes, such as conditional probabilities associated with a group of new observations, and demonstrated similar properties to CMMP for the resulting classified predictor. A number of recent results have been derived to extend the idea of CMMP to topics like functional data analysis (Xiu & Jiang (2024)) and a psuedo-Bayesian version of CMMP (Ma & Jiang (2023)).

3.1. Estimation of MSPE for CMMP

A standard uncertainty measure for a predictor is the MSPE. A "gold standard" for the MSPE estimation is to produce a second-order unbiased MSPE estimator, that is, the order of bias of the MSPE estimator is $o(m^{-1})$, where m is the total number of clusters in the training data. Typically, the $o(m^{-1})$ term is, in fact, $O(m^{-2})$, but this difference is usually ignored. For the most part, there have been two approaches for producing a second-order unbiased MSPE estimator. The first is the Prasad-Rao linearization method (Prasad & Rao 1990). The approach uses Taylor series expansion to obtain a second-order approximation to the MSPE, then corrects the bias, again to the second-order, to produce an MSPE estimator whose bias is $o(m^{-1})$. Various extensions of the Prasad-Rao method have been developed; see, for example, Datta & Lahiri (2000), Jiang & Lahiri (2001), Das, Jiang & Rao (2004), and Datta, Rao & Smith (2005). Although the method often leads to an analytic expression of the MSPE estimator, the derivation is tedious, and the final expression is likely to be complicated. More importantly, errors often occur in the process of analytic derivations as well as computer programming based on the lengthy expressions. Furthermore, the linearization method does not apply to situations where a non-differentiable operation is involved in obtaining the predictor, such as shrinkage estimation (*e.g.*, Tibshirani 1996), CMMP (Jiang *et al.* 2018), as well as the CMMP described in what follows in the next section.

The second approach to second-order unbiased MSPE estimation is resampling methods. Jiang, Lahiri & Wan (2002; hereafter JLW) proposed a jackknife method to estimate the MSPE of an empirical best predictor (EBP). The method avoids tedious derivations of the Prasad-Rao method, and is "one formula for all". On the other hand, there are restrictions on the class of predictors to which JLW applies. Namely, JLW only applies to empirical best predictor (EBP), that is, predictor obtained by replacing the parameters involved in the best predictor (BP), which is the conditional expectation, by their (consistent) estimators. The CMMP predictor, however, is not an EBP, because it involves a matching process. Jiang, Lahiri & Nguyen (2018) proposed a Monte-Carlo jackknife method, called McJack, which potentially applies to CMMP; however, the method is computationally very expensive. Another resampling-based approach is double bootstrapping (DB; Hall & Maiti 2006a,b). Although DB is capable of producing a second-order unbiased MSPE estimator, it is, perhaps, computationally even more intensive than the McJack. It is also unclear whether DB can be extended to CMMP.

In a way, the method to be proposed below may be viewed as a hybrid of the linearization method and resampling method, by combining the best part of each method. In short, we use a simple, analytic approach to obtain the leading term of our MSPE estimator, and a Monte-Carlo method to take care a remaining, lower-order term. The computational cost for the Monte-Carlo part is much lesser compared to McJack. For example, the computa-

tional burden of our method is about $1/m^3$ to $1/m^2$ of that for McJack. More importantly, the method provides a unified, conceptually easy solution to a difficult problem, that is, obtaining a second-order unbiased MSPE estimator for CMMP.

Let θ be the mixed effect corresponding to the new observations, and $\hat{\theta}$ the CMMP predictor of θ. The MSPE of $\hat{\theta}$ can be expressed as $\text{MSPE} = \text{E}(\hat{\theta} - \theta)^2 = \text{E}\left[\text{E}\{(\hat{\theta} - \theta)^2|y\}\right]$, where y represents the available data. Suppose that the underlying distribution of y depends on a vector of unknown parameters, ϕ. Then, the conditional expectation inside the expectation is a function of y and ϕ, which can be written as $a(y, \phi) = \text{E}\{(\hat{\theta} - \theta)^2|y\} = \hat{\theta}^2 - 2\hat{\theta}\text{E}(\theta|y) + \text{E}(\theta^2|y) = \hat{\theta}^2 - 2\hat{\theta}a_1(y, \phi) + a_2(y, \phi)$, where $a_j(y, \phi) = \text{E}(\theta^j|y), j = 1, 2$. If we replace the ϕ in $a(y, \psi)$ by $\hat{\phi}$, a consistent estimator of ϕ, the result is a first-order unbiased estimator, that is, we have $\text{E}\{a(y, \hat{\phi}) - a(y, \phi)\} = O(m^{-1})$. On the other hand, both $\text{MSPE} = \text{E}\{a(y, \phi)\}$ and $\text{E}\{a(y, \hat{\phi})\}$ are functions of ϕ, denoted by $b(\phi)$ and $c(\phi)$, respectively. It follows that $d(\phi) = b(\phi) - c(\phi) = O(m^{-1})$; thus, if we replace, again, to replace ϕ by $\hat{\phi}$ in $d(\phi)$, the difference is a lower-order term, that is, $d(\hat{\phi}) - d(\phi) = o_{\text{P}}(m^{-1})$ [see, *e.g.*, Jiang 2010, sec. 3.4 for notation like o_{P} and O_{P}]. Now consider the estimator

$$\widehat{\text{MSPE}} = a(y, \hat{\phi}) + d(\hat{\phi}) = a(y, \hat{\phi}) + b(\hat{\phi}) - c(\hat{\phi}). \tag{21}$$

We have $\text{E}(\widehat{\text{MSPE}}) = \text{E}\{a(y, \phi)\} + \text{E}\{a(y, \hat{\phi}) - a(y, \phi)\} + \text{E}\{d(\hat{\phi})\} = \text{MSPE} + \text{E}\{d(\hat{\phi}) - d(\phi)\} = \text{MSPE} + o(m^{-1})$. Essentially, this one-line, heuristic derivation shows the second-order unbiasedness of the proposed MSPE estimator, (21), provided that the terms involved can be evaluated.

Note that the leading term, $a(y, \hat{\phi})$, in (21) is guaranteed positive, a desirable property for an MSPE estimator. The lower-order term, $b(\hat{\phi}) - c(\hat{\phi})$, corresponds to a bias correction to the leading term. This term is typically much more difficult to evaluate than the leading term. We propose to approximate this term using a Monte-Carlo method. Let P_ϕ denote the distribution of y with ϕ being the true parameter vector. Given ϕ, one can generate y under P_ϕ. Let $y_{[k]}$ denote y generated under the kth Monte-Carlo sample, $k = 1, \ldots, K$. Then, by the law of large numbers, we have $b(\phi) - c(\phi) \approx K^{-1} \sum_{k=1}^{K} \left\{a(y_{[k]}, \phi) - a(y_{[k]}, \hat{\phi}_{[k]})\right\} \equiv d_K(\phi)$, where $\hat{\phi}_{[k]}$ denotes $\hat{\phi}$ based on $y_{[k]}$. If K is sufficiently large, which one has control over during the Monte-Carlo simulation, the difference between the two sides of the approximation is $o(m^{-1})$. Note that $y_{[k]}, k = 1, \ldots, K$ also depend on ϕ. Then, a Monte-Carlo assisted MSPE estimator (Nguyuen *et al.* 2022), is given by

$$\widehat{\text{MSPE}}_K = a(y, \hat{\phi}) + d_K(\hat{\phi}) = a(y, \hat{\phi}) + K^{-1} \sum_{k=1}^{K} \left\{a(y_{[k]}, \hat{\phi}) - a(y_{[k]}, \hat{\phi}_{[k]})\right\} \tag{22}$$

where $y_{[k]}, k = 1, \ldots, K$ are generated as above with $\phi = \hat{\phi}$, and $\hat{\phi}_{[k]}$ is, again, the estimator of ϕ based on $y_{[k]}$. (22) is called the Sumca estimator of the MSPE of $\hat{\theta}$ (Sumca is abbreviation of "simple, unified, Monte-Caro assisted").

4. Classified mixed model projections

In many practical problems, there is interest in the estimation of mixed effect projections for new data that are outside the range of the training data. Examples include predicting extreme small area means for rare populations or making treatment decisions for patients who do not fit typical risk profiles. Standard methods have long been known to struggle with such problems since the training data may not provide enough information about potential model changes for these new data values (extrapolation bias). Rao *et al.* (2024) proposed a new framework called Prediction Using Random-effect Extrapolation (PURE) which involves constructing a generalized independent variable hull (gIVH) to isolate a minority training set which is "close" to the prediction space, followed by a regrouping of the minority data according to the response variable which results in a new (but misspecified) random effect distribution. This misspecification reflects "extrapolated random effects" which prove vital to capture information that is needed for accurate model projections. Projections were then made using classified mixed model prediction (CMMP) (Jiang et al. 2018) with the regrouped minority data. Let us assume that, for $i = 1, \ldots, m$, $\boldsymbol{y}^{(k)}$ follow a mixed model as follows:

$$\boldsymbol{y}_i^{(k)} = X_i^{(k)}\boldsymbol{\beta}_k + Z_i^{(k)}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{23}$$

where $\boldsymbol{y}_i^{(k)} = (y_{ij})_{1\le j\le n_i^{(k)}}$, $X_i^{(k)} = (x_{ij}^{(k)})^T_{1\le j\le n_i^{(k)}}$ is a matrix of known covariates, $Z_i^{(k)}$ is a matrix of known covariates, $\boldsymbol{\beta}_k$ is a p-vector of unknown regression coefficients (the fixed effects), $\boldsymbol{b}_i$ is q-vector of group-specific random effects, and $\boldsymbol{\varepsilon}_i$ is an vector of errors. *Notice the different notation for the random effects from the previous CMMP in order to distinguish the two methods.*

The subscript (k) denotes the population k, and $1 \le k \le K$. It is assumed that $\boldsymbol{b}_i \sim N(0, G)$, $\boldsymbol{\varepsilon}_i \sim N(0, R_i)$ and they are independent, and the covariance matrices G and R_i depend on a vector $\boldsymbol{\psi}$ of variance components. Note $\boldsymbol{\beta}_k$ is different for different population k, and the random effects $\boldsymbol{b}_i$ are the same across k populations. The total number of observations in each population is $n^{(k)} = \sum_{i=1}^m n_i^{(k)}$, and the overall total population $n = \sum_{i=1}^K n^{(k)}$. Note that $n = \sum_{i=1}^m n_i$ where n_i is the number of observations in the group i. If the data follows (23), people usually fit a one component mixed model that assumed only one set of fixed effects parameters when the true model information is unknown, which results in a convenient but "misspecified" model fit.

Assume new test observations, which follow:

$$y_{n,j} = x_n'\boldsymbol{\beta}_n + z_n'\boldsymbol{b}_I + \varepsilon_{n,j}, \quad 1 \le j \le n_{new}, \tag{24}$$

where x_n and z_n are known vectors, and I belongs to one of the m groups. The new errors $\varepsilon_{n,j}$ are independent with mean zero, and variance R_{new} and are assumed independent of the training data. Notice $\boldsymbol{\beta}_n \ne \boldsymbol{\beta}_k, 1 \le k \le K$. The mixed effect we wish to predict is $\theta_n = E(y_{n,j} \mid b_I) = x_n'\boldsymbol{\beta} + z_n'\boldsymbol{b}_I$ where $I \in \{1, \ldots, m\}$ but we do not know which group I belongs to.

4.1. Generalized independent variable hull

Conn et al. (2015) proposed one possible definition of "the range of observation data" which turns to early works on outlier detection in simple linear regression analysis. Cook

(1979) referred that the smallest convex set containing all design points of a full-rank linear regression model as the independent variable hull (IVH). The IVH definition is based on linear model which require full rank of design matrix and i.i.d Gaussian error. Therefore, it can not be applied to generalized models such as binary response or random effects. Cook (1979) notes that design points with maximum prediction variance will be located on the boundary of IVH, then Conn et al. (2015) defined a generalized independent variable hull (gIVH) as a set of all predicted locations $\boldsymbol{S}_0$ for which

$$\text{var}(\lambda_i) \leq \max(\text{var}(\boldsymbol{\lambda}_{\boldsymbol{S}})), \tag{25}$$

where $i \in \boldsymbol{S}_0$, λ_i corresponds to the mean prediction at i, $\boldsymbol{S}$ denoted the set of locations where data are observed, and $\boldsymbol{\lambda}_{\boldsymbol{S}}$ denotes predictions at $\boldsymbol{S}$. Conn et al. (2015) proposed that the gIVH can be applied to determine whether predictions are interpolations (predictive design points lying inside the gIVH) or extrapolations (predictive design points lying outside the gIVH). This uses the generalization,

$$\boldsymbol{\mu} = X_{aug}\boldsymbol{\beta}_{aug}, \tag{26}$$

where X_{aug} is an augmented design matrix to accommodate the random effects design matrix Z and $\boldsymbol{\beta}_{aug}$ is the corresponding regression parameter vector. We can then write the prediction variance as,

$$\text{var}(\hat{\boldsymbol{\lambda}}) = \text{var}(\hat{\boldsymbol{\mu}}) = X_{aug}\text{var}(\hat{\boldsymbol{\beta}}_{aug})X'_{aug}. \tag{27}$$

One possibility is to use a flexible generalized additive model (GAM) (Hastie and Tibshirani, 1990) and then estimate the appropriate form of $\text{var}(\hat{\boldsymbol{\beta}}_{aug})$. If y is not on the linear predictor scale (*e.g.* generalized linear models outside of the normal model), then the delta method can be used to estimate $\text{var}(\hat{\boldsymbol{\lambda}})$ (Conn et al. 2015). Outside of these situations, simulation based methods like bootstrapping can be used to estimate the variance.

4.2. Prediction Using Random-effects Extrapolation (PURE)

Suppose we have a set of training data and test data as in (23) and (24). Let π_k denote the percentage of the population that comes from the population k, and $\sum_{k=1}^{K}\pi_k = 1$. If $K = 2$, we have π_1 percent of the population comes from the minority and the rest $1 - \pi_1$ population comes from the majority. We define the following relevant features:

1. Extreme data: This is the test dataset which may or may not be outside of range of the training data. Both cases can be handled here.

2. Majority data: Notationally, we can concatenate all observations in the full training data as $\mathscr{L} = \{(x_l, y_l); l = 1, \ldots, (n_1+n_2+\ldots+n_m)\}$. Then define the majority dataset as those further away from the test data. Let ‡ denotes the majority, we have a distance measure $d_‡ = |\text{median}(\text{var}(\lambda_‡)) - \max(\text{var}(\boldsymbol{\lambda}_{\boldsymbol{S}}))|$ where $\text{var}(\lambda_‡) > \max(\text{var}(\boldsymbol{\lambda}_{\boldsymbol{S}}))$ and $\lambda_‡$ denotes the λ that calculated from the majority data. Similarly, $d_†$ denotes the distance measure for the minority data and $d_‡ > d_†$. Therefore:

$$\mathscr{L}^‡ = \{(x_l, y_l)|d_‡ > d_†\}.$$

The original groupings are maintained so the majority data can be re-expressed according to the groupings.

3. Minority data: This portion of the data that is the complement of the majority data. This is found by a minority data decision rule to be described.

$$\mathscr{L}^{\dagger} = \{(x_l, y_l) | d_{\dagger} \leq d_{\ddagger}\}.$$

Again, the original groupings are maintained so the minority data can be re-expressed according to these groupings.

4. Re-grouped minority data $\mathscr{L}_R^{\dagger}$: For this, we take the minority data and re-group it according to a hierarchical clustering algorithm with respect to the responses $\boldsymbol{y}$ resulting in $m_r = m$ groupings with potentially revised memberships.

Rao *et al.* (2024) presented comprehensive simulation studies and analysis of data from the National Longitudinal Mortality Study (NLMS) which demonstrated superior predictive performance in these very challenging paradigms. An asymptotic analysis revealed why PURE resulted in more accurate projections.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of American Statistical Association*, **80**, 28-36.

Conn, P. B., Johnson, D. S., and Boveng, P. L. (2015). On extrapolating past the range of the observed data when making statistical predictions in ecology, *PLoS One*, **10**, e0141416.

Cook, R. D. (1979). Influential observations in linear regression, *Journal of American Statistical Association*, **74**, 169-174.

Das, K., Jiang, J., and Rao, J. N. K. (2004). Mean squared error of emprical predictor, *Annals of Statistics*, **32**, 818-840.

Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, **10**, 613–627.

Datta, G. S. and Rao, J. N. K. and Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**, 183-196.

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

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Annals of Statistics*, **34**, 1733-1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction, *Journal of Royal Statistical Society Series B*, **68**, 221-238.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, Boca Raton, Florida.

Gan, L., Yin, Z., and Tan, J. (2016). *Report on The Development of Household Finance in Rural China (2014)*, Springer, Singapore.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

Jiang, J. (2010). *Large Sample Techniques for Statistics*, Springer, New York.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data, *Annals of Institute of Statistics and Mathematics*, **53**, 217-243.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion), *TEST*, **15**, 1-96.

Jiang, J., Lahiri, P., and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation, *Annals of Statistics.*, **30**, 1782-1810.

Jiang, J., Lahiri, P., and Nguyen, T. (2018). A unified Monte-Carlo jackknife for small area estimation after model selection, *Annals of Mathematical Sciences and Applications*, **3**, 405-438.

Jiang, J., Nguyen, T., and Rao, J.S. (2011). Best predictive small area estimation, *Journal of American Statistical Association*, **106**.

Jiang, J., Rao, J. S., Fan, J., and Nguyen, T. (2018). Classified mixed model prediction, *Journal of American Statistical Association*, **113**, 269-279.

Liu, X., Ma, H., and Jiang, J. (2022b). That Prasad-Rao is robust: Estimation of mean squared prediction error of observed best preditor under potential model misspecification, *Statistica Sinica*, **32**, 2217-2240.

Liu, X. and Jiang, J. (2024). Classified functional mixed effects model prediction, *Statistics in Medicine*, **43**, 1329-1340.

Ma, H. and Jiang, J. (2023). Psuedo-Bayesian classified mixed model prediction, *Journal of American Statistical Association*, **118**, 1747-1759.

Nguyen, T., Jiang, J., and Rao, J. S. (2022). Assessing uncertainty for classified mixed model prediction, *Journal of Statistical Computation and Simulation*, **92**, 249-261.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error for small-area esitmators, *Journal of American Statistical Association*, **85**, 163-171.

Rao, J. N. K. (2003). *Small Area Estimation*, Wiley, New York.

Rao, J. S., Li, M., and Jiang, J. (2024). Classified mixed model projections, *Journal of American Statistical Association*, to appear.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion), *Statistical Science*, **6**, 15-51.

Sun, H., Nguyen, T., Luan, Y., and Jiang, J. (2018). Classified mixed logistic model prediction, *Journal of Multivariate Analysis*, **168**, 63-74.

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

Welham, S., Cullis, B., Gogel, B., Gilmour, A., and Thompson, R. (2004). Prediction in linear mixed models, *Australian & New Zealand Journal of Statistics*, **46**, 325-347.

Zhang, J., Gan, L., Xu, L. C., and Yao, Y. (2014). Health shocks, village elections, and household income: Evidence from rural China, *China Economic Review*, **30**, 155-168.

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2024

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203
Email: galaxystudio08@gmail.com