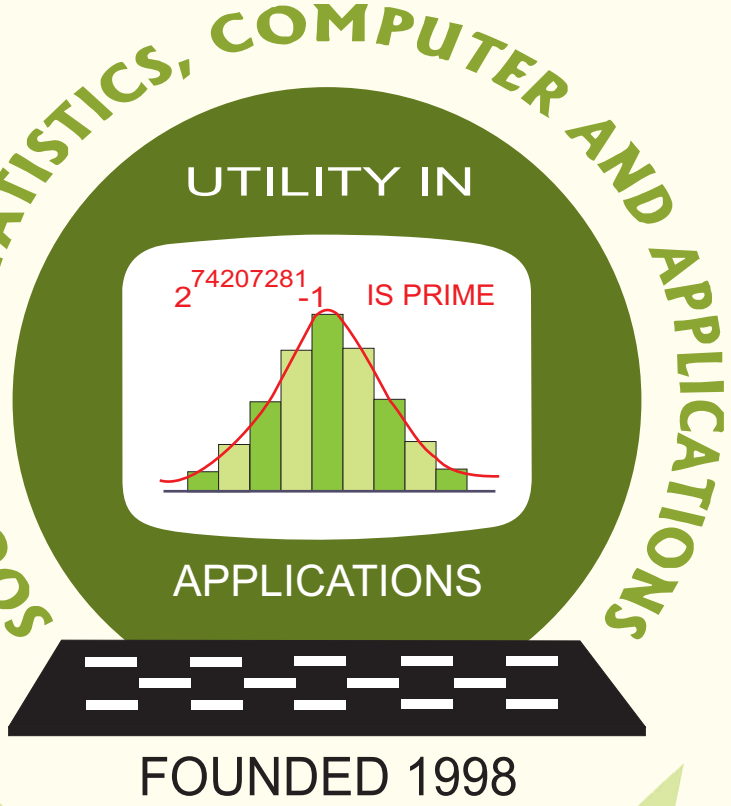
Special Proceedings
22nd Annual Conference of SSCA
held at Savitribai Phule Pune University, Pune Maharashtra

02-04 January 2020



Society of Statistics, Computer and Applications
https://ssca.org.in/
2020

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

A. Dhandapani, Lal Mohan Bhar, P. Venkatesan
S.D. Sharma, V.K. Bhatia

Secretary
D. Roy Choudhury

Foreign Secretary
Sudhir Gupta

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Hukum Chandra, Shibani Roy Choudhury

Council Members

Alka Arora, Anil Kumar Yadav, Anshu Bhardwaj, Manish Sharma, Manish Trivedi
Manisha Pal, P. Rath, Piyush Kant Rai, Praggya Das, R. Vishnu Vardhan
Ranjit Kumar Paul, Rao Saheb Latpate, S.A. Mir, Sanjeev Panwar, V. Srinivasa Rao

Ex-Officio Members (By Designation)

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

Citation: Special Proceedings of 22nd Annual Conference of SSCA (2020). Society of Statistics, Computer and Applications, New Delhi.

Special Proceedings
22nd Annual Conference of SSCA
held at Savitribai Phule Pune University, Pune Maharashtra

02-04 January 2020

Editors

V.K. Gupta

Baidya Nath Mandal

Rajender Parsad

Hukum Chandra

Ranjit Kumar Paul

K.J.S. Satyasai

D. Roy Choudhury

Society of Statistics, Computer and Applications

https://ssca.org.in/

2020

CONTENTS

Special Proceedings of 22nd Annual Conference of SSCA 2020
held at Savitribai Phule Pune University, PUNE during 02-04 January 2020

Preface

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

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

The twenty-second Annual Conference of the SSCA was organised during 02-04 January 2020 at the Department of Statistics, Savitribai Phule Pune University, Pune, Maharashtra. The theme of the conference was *Importance of Statistics in Emerging Global Scenario (ISEGS* 2020*)*. The conference was academically enriching with important and significant presentations made by scientists of international repute and eminence. Among the many technical sessions organised, were Professor C.R. Rao Birth Centenary Lecture Session, participated by renowned international and national statisticians and a session on Financial Statistics, in which renowned statisticians and leading practitioners from National Bank for Agriculture and Rural Development, Symbiosis Institute of Management Studies, Pune, Savitri Phule Pune University and Reserve Bank of India made presentations.

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

Distinguished speakers shortlisted for making contributions to the special proceedings were invited to submit their research papers for possible inclusion in the special proceedings. After the usual review process, **16** research papers were accepted for publication and are included in the special proceedings. We would like to express our sincere thanks to all the authors for responding to our request and submitting their research for publication in these special proceedings in time. The reviewers have also made a very big contribution by way of finishing the review process in a short span of time and it is a pleasure to thank each one of them individually. We would like to place on record our gratitude to all members and office

bearers of the Executive Council of SSCA for their support. We would also like to express our sincerest thanks to Prof. T.V. Ramanathan, Dr. V.K. Gedam, and Dr. R.V. Latpate for organising the Conference.

Guest Editors

V.K. Gupta
Baidya Nath Mandal
Rajender Parsad
Hukum Chandra
Ranjit Kumar Paul
K.J.S. Satyasai
Dipak Roy Choudhury

New Delhi
September 2020

Special Proceedings of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 1-10

Correlated Inverse Gaussian Frailty Model

David D. Hanagal
Symbiosis Statistical Institute, Symbiosis International University, Pune-411016, India
and
Department of Statistics, Savitribai Phule Pune University, Pune-411007, India.

Received: 31 January 2020; Revised: 07 February 2020; Accepted: 09 February 2020

Abstract

Frailty models are used in the survival analysis to account for the unobserved heterogeneity in individual risks to disease and death. To analyze the bivariate data on related survival times, the shared frailty models were suggested. Shared frailty models are used despite their limitations. To overcome their disadvantages correlated frailty models may be used. In this paper, we introduce the inverse Gaussian correlated frailty models.

Key words: Vivariate survival; Copula; Correlated inverse Gaussian frailty; Cross-ration function; Hazard rate.

1. Introduction

The frailty model is a random effect model for time to event data which is an extension of the Cox's proportional hazards model. Shared frailty models are the most commonly used frailty models in literature, where individuals in the same cluster share a common frailty. Frailty models (Vaupel et al. 1979) are used in the survival analysis to account for the unobserved heterogeneity in the individual risks to disease and death. The frailty model is usually modeled as an unobserved random variable acting multiplicatively on the baseline hazard function. Hanagal and Dabade (2013), Hanagal and Bhambure (2015, 2016) and Hanagal and Pandey (2014a, 2014b, 2015a, 2015b, 2016, 2017a) analyzed kidney infection data and Australian twin data using shared gamma and inverse Gaussian frailty models with different baseline distributions for the multiplicative model. Hanagal and Sharma (2013, 2015a, 2015b, 2015c) analyzed acute leukemia data, kidney infection data and diabetic retinopathy data using shared gamma and inverse Gaussian frailty models for the multiplicative model. Hanagal and Bhambure (2014) developed shared inverse Gaussian frailty model based on the reversed hazard rate for Australian twin data. Hanagal et al.(2017) discussed correlated gamma frailty models for bivariate survival data to analyze kidney infection data and Hanagal and Pandey (2017b) proposed correlated gamma frailty models for bivariate survival data based on reversed hazard rate for Australian twin data. Hanagal (2017) gave extensive literature review on different shared frailty models.

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

In a univariate frailty model, let a continuous random variable T be a lifetime of an individual and the random variable Z be frailty variable. The conditional hazard function for a given frailty variable, $Z = z$ at time $t > 0$ is,

$$h(t \mid z) = zh_0(t)e^{\mathbf{X}\boldsymbol{\beta}}, \tag{1}$$

where $h_0(t)$ is a baseline hazard function at time $t > 0$, $\mathbf{X}$ is a row vector of covariates, and $\boldsymbol{\beta}$ is a column vector of regression coefficients. The conditional survival function for given frailty at time $t > 0$ is,

$$S(t \mid z) = e^{-\int_0^t h(x|z)dx} = e^{-zH_0(t)e^{\mathbf{X}\boldsymbol{\beta}}}, \tag{2}$$

where $H_0(t)$ is the cumulative baseline hazard function at time $t > 0$. Integrating over the range of frailty variable Z having density $f_Z(z)$, we get the marginal survival function as,

$$\begin{aligned} S(t) &= \int_0^\infty S(t \mid z) f_Z(z) dz \\ &= \int_0^\infty e^{-zH_0(t)e^{\mathbf{X}\boldsymbol{\beta}}} f_Z(z) dz \\ &= L_Z(H_0(t)e^{\mathbf{X}\boldsymbol{\beta}}), \end{aligned} \tag{3}$$

where $L_Z(.)$ is the Laplace transformation of the distribution of Z. Once we get the survival function at time $t > 0$, of life time random variable for an individual, we can obtain probability structure and make their inferences based on it.

Shared frailty explains correlation's between subjects within clusters. However, it does have some limitations. Firstly, it forces the unobserved factors to be the same within the cluster, which may not always reflect reality. For example, at times it may be inappropriate to assume that all partners in a cluster share all their unobserved risk factors. Secondly, the dependence between survival times within the cluster is based on marginal distributions of survival times. However, when covariates are present in a proportional hazards model with gamma distributed frailty the dependence parameter and the population heterogeneity are confounded (Clayton and Cuzick, 1985). This implies that the joint distribution can be identified from the marginal distributions (Hougaard, 1986). Thirdly, in most cases, a one-dimensional frailty can only induce positive association within the cluster. However, there are some situations in which the survival times for subjects within the same cluster are negatively associated. For example, in the Stanford Heart Transplantation Study, generally the longer an individual must wait for an available heart, the shorter he or she is likely to survive after the transplantation. Therefore, the waiting time and the survival time afterwards may be negatively associated.

To avoid these limitations, correlated frailty models are being developed for the analysis of multivariate failure time data, in which associated random variables are used to characterize the frailty effect for each cluster. Correlated frailty models provide not only variance parameters of the frailties as in shared frailty models, but they also contain additional parameter for modeling the correlation between frailties in each group. Frequently one is interested in construction of a bivariate extension of some univariate family distributions (e.g., gamma). For example, for the purpose of genetic analysis of frailty one might be interested in estimation of correlation

of frailty. It turns out that it is possible to carry out such extension for the class of infinitely-divisible distributions (Iachine 1995a, 1995b). In this case an additional parameter representing the correlation coefficient of the bivariate frailty distribution is introduced.

2. Inverse Gaussian Frailty

The gamma distribution is most commonly used frailty distribution because of its mathematical convenience. Another choice is the inverse Gaussian distribution. The inverse Gaussian makes the population homogeneous with time, whereas for gamma the relative heterogeneity is constant (Hougaard, 1984). Duchateau and Janssen (2008) fit the inverse Gaussian (IG) frailty model with Weibull hazard to the udder quarter infection data. The IG distribution has a unimodal density and is a member of the exponential family. While its shape resembles that of other skewed density functions, such as lognormal and gamma, it provides much flexibility in modeling. Furthermore, there are many striking similarities between the statistics derived from this distribution and those of the normal; see Chhikara and Folks (1986). These properties make it potentially attractive for modeling purposes with survival data. The models derived above are bases on the assumption that a common random effect acts multiplicatively on the hazard rate function.

Alternative to the gamma distribution, Hougaard (1984) introduced the inverse Gaussian as a frailty distribution. It provides much flexibility in modeling, when early occurrences of failures are dominant in a life time distribution and its failure rate is expected to be non-monotonic. In such situations, the inverse Gaussian distribution might provide a suitable choice for the lifetime model. Also inverse Gaussian is almost an increasing failure rate distribution when it is slightly skewed and hence is also applicable to describe lifetime distribution which is not dominated by early failures. Secondly, for the inverse Gaussian distribution, the surviving population becomes more homogeneous with respect to time, where as for gamma distribution the relative heterogeneity is constant. The inverse Gaussian distribution has shape resembles the other skewed density functions, such as log-normal and gamma. These properties of inverse Gaussian distribution motivate us to use inverse Gaussian as frailty distribution. The inverse Gaussian distribution has a history dating back to 1915 when Schrodinger and Smoluchowski presented independent derivations of the density of the first passage time distribution of Brownian motion with positive drift. Villman et al., (1990) have studied the histomorphometrical analysis of the influence of soft diet on masticatory muscle development in the muscular dystrophic mouse. The muscle fibre size distributions were fitted by an inverse Gaussian law. Barndorff-Nielsen (1994) considers a finite tree whose edges are endowed with random resistances, and shows that, subject to suitable restrictions on the parameters, if the resistances are either inverse Gaussian or reciprocal inverse Gaussian random variables, then the overall resistance of the tree follows a reciprocal inverse Gaussian law. Gacula and Kubala (1975) have analyzed shelf life of several products using the IG law and found to be a good fit. For more real life applications (see Seshadri, 1999).

Consider a continuous random variable Z follows inverse Gaussian distribution with

parameters μ and σ^2 then density function of Z is,

$$f_Z(z) = \begin{cases} \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}} z^{-\frac{3}{2}} e^{\frac{(z-\mu)^2}{2z\sigma^2\mu^2}} & ;\ z>0, \mu>0, \sigma^2>0 \\ 0 & ;\ otherwise, \end{cases} \tag{4}$$

and the Laplace transform is,

$$L_Z(s) = exp\left[\frac{1}{\mu\sigma^2} - \left(\frac{1}{\sigma^4\mu^2} + \frac{2s}{\sigma^2}\right)^{\frac{1}{2}}\right]. \tag{5}$$

The mean and variance of frailty variable are $E(Z) = \mu$ and $V(Z) = \mu^3\sigma^2$. For identifiability, we assume Z has expected value equal to one i.e. $\mu = 1$. Under this restriction, the density function and the Laplace transformation of the inverse Gaussian distribution reduces to,

$$f_Z(z) = \begin{cases} \left[\frac{1}{2\pi\sigma^2}\right]^{\frac{1}{2}} z^{-\frac{3}{2}} e^{\frac{(z-1)^2}{2z\sigma^2}} & ;\ z>0, \sigma^2>0 \\ 0 & ;\ otherwise, \end{cases} \tag{6}$$

and the Laplace transform is,

$$L_Z(s) = exp\left[\frac{1-(1+2\sigma^2 s)^{\frac{1}{2}}}{\sigma^2}\right], \tag{7}$$

with variance of Z as σ^2. The frailty variable Z is degenerate at $Z = 1$ when σ^2 tends to zero. Let T_1 and T_2 be failure times of the pair of individuals like kidney, lungs, eyes or any paired organ of an individual or lifetimes of twins. The unconditional bivariate distribution function of lifetimes T_1 and T_2 with inverse Gaussian frailty is,

$$\begin{aligned} L_Z(H_1(t_1) + H_2(t_2)) &= exp\left[\frac{1-(1+2\theta(H_1(t_1)+H_2(t_2)))^{\frac{1}{2}}}{\theta}\right] \\ &= S(t_1, t_2) \end{aligned} \tag{8}$$

where $H_1(t_1)$ and $H_2(t_2)$ are the cumulative baseline hazard functions of the lifetime T_1 and T_2 respectively. Clayton (1978) define cross-ratio function as,

$$\theta^*(t_1, t_2) = \frac{\frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} S(t_1, t_2)}{\frac{\partial S(t_1,t_2)}{\partial t_1}\frac{\partial S(t_1,t_2)}{\partial t_2}}$$

The cross ratio function of inverse Gaussian frailty is,

$$\theta^*(t_1, t_2) = 1 + \frac{1}{\frac{1}{\theta} - ln(S(t_1, t_2))}$$

The highest value is obtained at the start and equals $1 + \theta$, and goes to one as the survival function goes to zero. It is decreasing function of t_1, t_2.

The joint bivariate survival functions in (8) can be expressed in terms of survival copula as (see Nelsen (2006) for details)

$$\overline{C}(u,v) = exp\left\{\frac{1-[(1-\theta\log u)^2+(1-\theta\log v)^2-1]^{\frac{1}{2}}}{\theta}\right\}$$

where $u = S_{T_1}(\cdot)$ and $v = S_{T_2}(\cdot)$. This is a new copula and not appeared in the earlier literature.

3. Correlated Frailty

The correlated frailty model is the second important concept in the area of multivariate frailty models. It is a natural extension of the shared frailty approach on the one hand, and of the univariate frailty model on the other. In the correlated frailty model, the frailties of individuals in a cluster are correlated but not necessarily shared. It enables the inclusion of additional correlation parameters, which then allows the addressing of questions about associations between event times. Furthermore, associations are no longer forced to be the same for all pairs of individuals in a cluster. This makes the model especially appropriate for situations where the association between event times is of special interest, for example, genetic studies of event times in families. The conditional survival function in the bivariate case (here without observed covariates) looks like

$$S(t_1,t_2|Z_1,Z_2) = S_1(t_1|Z_1)S_2(t_2|Z_2) = e^{-Z_1H_{01}(t_1)}e^{-Z_2H_{02}(t2)}, \tag{9}$$

where Z_1 and Z_2 are two correlated frailties. The distribution of the random vector (Z_1, Z_2) needs to be specified and determines the association structure of the event times in the model. Integrating the above bivariate survival function over Z_1 and Z_2, we get unconditional bivariate survival function as

$$S(t_1,t_2) \;\; = \;\; E_{Z_1,Z_2}[e^{-Z_1H_{01}(t_1)}e^{-Z_2H_{02}(t2)}] \tag{10}$$

where (Z_1, Z_2) has some known bivariate frailty distribution.

Consider some bivariate event times – for example, the lifetimes of twins, or age at onset of a disease in spouses, time to blindness in the left and right eye, or time to failure in the left and right kidney of patients. In the (bivariate) correlated frailty model, the frailty of each individual in a pair is defined by a measure of relative risk, that is, exactly as it was defined in the univariate case. For two individuals in a pair, frailties are not necessarily the same, as they are in the shared frailty model. We are assuming that the frailties are acting multiplicatively on the baseline hazard function (proportional hazards model) and that the observations in a pair are conditionally independent, given the frailties. Hence, the hazard of the individual $i(i = 1, 2)$ in pair $j(i = j, ..., n)$ has the form

$$h(t|X_{ij},Z_{ij}) = Z_{ij}h_{0i}(t)e^{\beta' X_{ij}}, \tag{11}$$

where t denotes age or time, X_{ij} is a vector of observed covariates, $\boldsymbol{\beta}$ is a vector of regression parameters describing the effect of the covariates X_{ij}, $h_{0i}(.)$ are baseline hazard functions, and

Z_{ij} are frailties. Bivariate correlated frailty models are characterized by the joint distribution of a two-dimensional vector of frailties (Z_{1j}, Z_{2j}). If the two frailties are independent, the resulting lifetimes are independent, and no clustering is present in the model. If the two frailties are equal, the shared frailty model is obtained as a special case of the correlated frailty model with correlation one between the frailties (Wienke(2011)).

In order to derive a marginal likelihood function, the assumption of conditional independence of lifespans, given the frailty, is used. Let δ_{ij} be a censoring indicator for individual $i(i = 1, 2)$ in pair $j(j = 1, ..., n)$. Indicator δ_{ij} is 1 if the individual has experienced the event of interest, and 0 otherwise. According to (2.2), the conditional survival function of the ith individual in the jth pair is

$$S(t|X_{ij}, Z_{ij}) = e^{-Z_{ij}H_{0i}(t)}e^{\beta' X_{ij}}, \tag{12}$$

with $H_{0i}(t)$ denoting the cumulative baseline hazard function. The contribution of individual $i(i = 1, 2)$ in pair $j(j = 1, ..., n)$ to the conditional likelihood is given by

$$\left[Z_{ij}h_{0i}(t)e^{\beta' X_{ij}}\right]^{\delta_{ij}} e^{Z_{ij}H_{0i}(t_{ij})}e^{\beta' X_{ij}}, \tag{13}$$

where t_{ij} stands for observation time of individual i from pair j. Assuming the conditional independence of lifespans, given the frailty, and integrating out the frailty, we obtain the marginal likelihood function

$$\prod_{j=1}^{n} \int_{R\times} \int_{R} \left[u_{1j}h_{01}(t_{1j})e^{\beta' X_{1j}}\right]^{\delta_{1j}} e^{u_{1j}H_{01}(t_{1j})}e^{\beta' X_{1j}}$$
$$\left[u_{2j}h_{02}(t_{2j})e^{\beta' X_{2j}}\right]^{\delta_{2j}} e^{u_{2j}H_{02}(t_{2j})}e^{\beta' X_{2j}} f(z_{1j}, z_{2j})dz_{1j}dz_{2j} \tag{14}$$

where $f(.,.)$ is the probability density function of the corresponding frailty distribution. All these formulas can be easily extended to the multivariate case, but need a specification of the correlation structure between individuals in a cluster in terms of the multivariate density function, which complicates analysis. For more details see (Hanagal(2011, 2019) and Wienke(2011)).

4. Correlated Inverse Gaussian Frailty Model

Let Z be an infinitely divisible frailty variable with Laplace transformation $L_Z(s)$ and $\rho \in [0, 1]$, then there exist random variables Z_1, Z_2 each with univariate Laplace transform $L_Z(s)$ such that the Laplace transform of Z_1, Z_2 is given by:

$$L(s_1, s_2) = L_Z^{\rho}(s_1 + s_2)L_Z^{1-\rho}(s_1)L_Z^{1-\rho}(s_2) \tag{15}$$

If Z has a variance the $Corr(Z_1, Z_2) = \rho$.
The respective bivariate survival model is identifiable under mild regularity conditions on Z provided that $\rho > 0$. The case $\rho = 1$ is known as the shared frailty model.

The above equation can be extended to multivariate case ($\rho > 0$) as below.

$$L(s_1, s_2, ..., s_k) \;=\; L_Z^{\rho}(s_1, s_2, ..., s_k)L_Z^{1-\rho}(s_1)....L_Z^{1-\rho}(s_k).$$

The case $\rho = 1$ leads to shared frailty. If $\rho = 0$, $Z_1, ...Z_k$ are mutually independent.

Let Z_i be the inverse Gaussian distributed with mean 1, variance σ^2, and Laplace transform

$$L(s_i, \sigma^2) = \exp[\frac{1-(1+2\sigma^2 s_i)^{\frac{1}{2}}}{\sigma^2}] \tag{16}$$

The bivariate Laplace transform for the correlated inverse Gaussian frailty model is given by

$$\begin{aligned} L(s_1, s_2, \sigma^2, \rho) &= \exp\left[\rho\frac{1-(1+2\sigma^2(s_1+s_2))^{\frac{1}{2}}}{\sigma^2}\right]\exp\left[(1-\rho)\frac{1-(1+2\sigma^2 s_1)^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2 s_2)^{\frac{1}{2}}}{\sigma^2}\right] \end{aligned} \tag{17}$$

where $Corr(Z_1, Z_2) = \rho$.
The correlated frailty model with inverse Gaussian frailty distribution is characterized by the bivariate survival function of the form:

$$\begin{aligned} S(t,t_{2j}) &= \exp\left[\rho\frac{1-(1+2\sigma^2\eta_j(H_1(t_{1j})+H_2(t_{2j})))^{\frac{1}{2}}}{\sigma^2}\right]\exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_j H_1(t_{1j}))^{\frac{1}{2}}}{\sigma^2}\right] \\ &\quad \exp\left[(1-\rho)\frac{1-(1+2\sigma^2\eta_j H_2(t_{2j}))^{\frac{1}{2}}}{\sigma^2}\right] \end{aligned} \tag{18}$$

where $H_{01}(t_{1j})$ and $H_{02}(t_{2j})$ are the cumulative baseline hazard functions of the life time random variables T_{1j} and T_{2j} respectively.

According to different assumptions on the baseline distributions we get different correlated inverse Gaussian frailty models.

5. Likelihood Specification and Bayesian Estimation of Parameters

Suppose there are n individuals under study, whose first and second observed failure times are represented by (t_{1j}, t_{2j}). Let c_{1j} and c_{2j} be the observed censoring times for the j^{th} individual ($j = 1, 2, 3, ..., n$) for first and second recurrence times respectively. We also assume that independence between the censoring time and the life-times of individuals.

The contribution of the bivariate life time random variable of the j^{th} individual in likelihood function is given by,

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

and the likelihood function is,

$$L(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta) = \prod_{j=1}^{n_1} f_1(t_{1j}, t_{2j}) \prod_{j=1}^{n_2} f_2(t_{1j}, c_{2j}) \prod_{j=1}^{n_3} f_3(c_{1j}, t_{2j}) \prod_{j=1}^{n_4} f_4(c_{1j}, c_{2j}) \tag{19}$$

where θ, $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$ are respectively the frailty parameter $(\sigma_1, \sigma_2, \rho)$, the vector of baseline parameters and the vector of regression coefficients.

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

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}} \\ f_2(t_{1j}, c_{2j}) &= \frac{\partial S(t_{1j}, c_{2j})}{\partial t_{1j}} \\ f_3(c_{1j}, t_{2j}) &= \frac{\partial S(c_{1j}, t_{2j})}{\partial t_{2j}} \\ and\ f_4(c_{1j}, c_{2j}) &= S(c_{1j}, c_{2j}) \end{aligned} \tag{20}$$

Usually maximum likelihood estimators can be used to estimate the parameters invloved in the model. Unfortunately computing the maximum likelihood estimators (MLEs) involves solving a fourteen dimensional optimization problem for Model I and Model III and eleven dimensional optimization problem for Model II and Model IV. As the method of maximum likelihood fails to estimate the parameters due to convergence problem in the iterative procedure, so we use the Bayesian approach. The traditional maximum likelihood approach to estimation is commonly used in survival analysis, but it can encounter difficulties with frailty models. Moreover, standard maximum likelihood based inference methods may not be suitable for small sample sizes or situations in which there is heavy censoring (see Kheiri et al. (2007)). Thus, in our problem a Bayesian approach, which does not suffer from these difficulties, is a natural one, even though it is relatively computationally intensive

To estimate parameters of the model, the Bayesian approach is now popularly used, because computation of the Bayesian analysis become feasible due to advances in computing technology [for more details on Bayesian estimation of the parameters and data analysis based on correlated inverse Gaussian frailty model, see Hanagal and Pandey, 2020].

References

Bain, L. J. (1974). Analysis for the linear failure-rate life-testing distribution. *Technometrics*, **16(4)**, 551 – 559.

Chhikara, R. S. and Folks, J. L. (1986). *The Inverse Gaussian Distribution*. Marcel Dekker, New York.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society*, **A148**, 82-117.

Gacula, M. C. Jr. and Kubala, J. J. (1975). Statistical models for shelf life failures. *Journal of Food Science*, **40**, 404-409.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer. New York.

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

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

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

Hanagal, D. D. and Bhambure, S. M. (2014). Shared inverse Gaussian frailty model based on reversed hazard rate for modeling Australian twin data. *Journal of the Indian Society for Probability and Statistics*, **15**, 9-37.

Hanagal, D. D. and Bhambure, S. M. (2015). Comparison of shared gamma frailty models using Bayesian approach. *Model Assisted Statistics and Applications*, **10**, 25-41.

Hanagal, D. D. and Bhambure, S. M. (2016). Modeling bivariate survival data using shared inverse Gaussian frailty model. *Communications in Statistics, Theory and Methods*, **45(17)**, 4969-4987.

Hanagal, D. D. and Dabade, A. D. (2013). Modeling of inverse Gaussian frailty model for bivariate survival data. *Communications in Statistics, Theory and Methods*, **42(20)**, 3744-3769.

Hanagal, D. D. and Pandey, A. (2014a). Inverse Gaussian shared frailty for modeling kidney infection data. *Advances in Reliability*, **1**, 1-14.

Hanagal, D. D. and Pandey, A. (2014b). Gamma shared frailty model based on reversed hazard rate for bivariate survival data. *Statistics and Probability Letters*, **88**, 190-196.

Hanagal, D. D. and Pandey, A. (2015a). Gamma frailty models for bivarivate survival data. *Journal of Statistical Computation and Simulation*, **85(15)**, 3172-3189.

Hanagal, D. D. and Pandey, A. (2015b). Inverse Gaussian shared frailty models with generalized exponential and generalized inverted exponential as baseline distributions. *Journal of Data Science*, **13(2)**, 569-602.

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

Hanagal, D. D. and Pandey, A. (2017a). Shared frailty model based on reversed hazard rate for left censoring data. *Communications in Statistics, Simulation and Computation*, **46(1)**, 230-243.

Hanagal, D. D. and Pandey, A. (2017b). Correlated gamma frailty models for bivariate survival data based on reversed hazard rate. *International Journal of Data Science*, **2(4)**, 301-324.

Hanagal, D. D., Pandey, A. and Ganguly, A. (2017). Correlated gamma frailty models for bivariate survival data. *Communications in Statistics, Simulation and Computation*, **46(5)**, 3627-3644.

Hanagal, D. D. and Pandey, A. (2020). Correlated inverse Gaussian frailty models for bivariate survival data. *Communications in Statistics, Theory and Methods*, **49(4)**, 845-863.

Hanagal, D. D. and Sharma, R. (2013). Modeling heterogeneity for bivariate survival data by shared gamma frailty regression model. *Model Assisted Statistics and Applications*, **8**, 85-102.

Hanagal, D. D. and Sharma, R. (2015a). Bayesian inference in Marshall-Olkin bivariate exponential shared gamma frailty regression model under random censoring. *Communications in Statistics, Theory and Methods*, **44(1)**, 24-47.

Hanagal, D. D. and Sharma, R. (2015b). Comparison of frailty models for acute leukaemia data under Gompertz baseline distribution. *Communications in Statistics, Theory and Methods*, **44(7)**, 1338-1350.

Hanagal, D. D. and Sharma, R. (2015c). Analysis of bivariate survival data using shared inverse Gaussian frailty model. *Communications in Statistics, Theory and Methods*, **44(7)**, 1351-1380.

Hougaard, P. (1984). Life table methods for heterogeneous populations. *Biometrika*, **71(1)**, 75-83.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396.

Iachine, I. A. (1995a). *Correlated frailty concept in the analysis of bivariate survival data*. Bachelor project, Department of Mathematics and Computer Science, Odense University, Denmark.

Iachine, I. A. (1995b). *Parameter estimation in the bivariate correlated frailty model with observed covariates via the EM-algorithm*. Working Paper Series: Population Studies of Aging 16, CHS, Odense University, Denmark.

Kheiri, S., Kimber, A. and Meshkani M. R. (2007). Bayesian analysis of an inverse Gaussian correlated frailty model. *Computational Statistics and Data Analysis*, **51**, 5317-5326.

Seshadri, V. (1999). *The Inverse Gaussian Distribution: Statistical Theory and Applications*. Springer Science, New York.

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

Wienke, A.(2011). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC. New York.

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 11-24

Review of Temporal Point Process Models for Modeling Earthquake Aftershocks

Rupal Shah and K. Muralidharan
Department of Statistics, The M.S. University of Baroda, Gujarat, India

Received: 29 March 2020; Revised: 10 April 2020; Accepted: 14 April 2020

Abstract

It is well known that the occurrence of earthquake is likely to increase another earthquake or the number (or sequence) of aftershocks in the nearby space and time. Similarly, prior to the next major earthquake, pre-seismic foreshocks are expected to occur in the focal region. Thus, it may be of interest for seismologists to study the pattern of sequence of foreshocks and aftershocks for prediction of earthquake. To investigate this problem with the help of quantitative information like, magnitude of earthquake, latitude, longitude, time etc., it might be possible to develop statistical models for the definition and detection of the occurrence of earthquake and sequence of aftershocks. The point process models are most frequently used to model such time to event data. In the time to event data, because of their temporal context, the point process models are also referred to as temporal point process models. Some of the frequently used models are Self-Exciting point process models, Epidemic Type Aftershock-Sequences models, stress release model etc. In this paper, we review various other models specific to temporal events like earthquake and other situations.

Key words: Ground intensity function; Trigger function; Self-exciting point process; Marked point process; ETAS models; Stress release model (SRM).

1. Introduction

In variety of applications, like, occurrence of crime events, posts or likes or clicks on social media, occurrence of earthquakes and its aftershocks, trading in financial markets etc., we come across the sequence of events which are asynchronous in nature, unlike the time series in which the observations are taken at regular time intervals. The time points at which such events of interest occur are also known as temporal points and such data can carry some important relevant information other than the temporal points.

Some examples are:

1. In study of crime patterns, an act of violence by an individual or a group might provoke or stimulate counter attacks by another group of individuals; where the temporal events are the times when violence by one gang took place and relevant information may be the retaliatory attacks by rivalry gang, the locations, the race of victims etc.
2. One post on social media is followed by number of related posts or one video is repeatedly watched by number of viewers; here the temporal events are time of post or video and

Corresponding Author: K. Muralidharan
Email: muralikustat@gmail.com

relevant information may be number of likes or number of times it is watched or tagged, number of positive or negative comments, the geographic location of the respondents etc.

3. In market research, the sales of a product might get a jump following repeated advertisements; where the temporal events are the time of purchase of a pre specified consumer product and the relevant information may be the impact of television advertisements of the product, income, size of the household etc.
4. In medical context, a targeted treatment like chemotherapy can impact the diseased cells; the temporal events in this case may be the time points at which the patient is given a dose of chemotherapy and relevant information may be the dose, the effects (or side effects) etc.
5. In earthquake modeling, an earthquake of higher magnitude might induce more aftershocks as compared to that with a smaller magnitude; the temporal events in this case may be the time points at which earthquakes and major aftershocks (of magnitude beyond some predefined intensity) occur and the relevant information may be longitude and latitude of the location, magnitude of the aftershock etc.
6. In financial markets, when the stock prices are very much volatile or fluctuating (ultra-high frequency); the time points at which the prices rise or decline by a predefined number may be considered to be temporal points and the relevant information may be the factors like international crude oil price, political situation or stability of the country, some event which has international relevance or concerns etc.

In general, temporal events may be either statistically independent, following a Poisson process, or temporally correlated. The term "event dependence" captures the idea that an initial event can increase or decrease the likelihood of subsequent events in the future. Point processes are often found to be appropriate for modeling a series of asynchronous events occurring at points in time (Ascher and Feigngold, 1984) and are also useful for modeling some peculiar pattern of the events observed in time. Among the few examples cited above, we focused on the earthquake data from the perspective of statistical modeling and analysis.

The behavior of the temporal point process may be described with the help of the conditional intensity function, conditional on the history of the process over time. The conditional intensity function proposes the probability of occurrence of the subsequent event in the upcoming instance of time given the history of the past events until the present event. Thus it represents the rate for the occurrence of a new event, conditioned on the history of the process up to time *t*, say, $\mathcal{H}_t = \{t_i : t_i < t\}$, as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{E[N([t, t+\Delta t]) \mid \mathcal{H}_t]}{\Delta t} \tag{1}$$

where $N[\dots]$ is the number of events occurring in $(t, t + \Delta t)$.

The Nonhomogeneous Poisson process (NHPP) models are the most frequently used models to represent the temporal events. The NHPP models are models for reliability growth due to the nature of their intensity function which is a compromise between 'as good as new' and 'as bad as old' models [Duane (1964), Bassin (1969, 1973), Cox and Lewis (1966), Lewis (1970, 1972)]. At times we come across the temporal events with branching structure wherein the occurrence of an event of one type increases the chance of occurrence of similar kind of event in

future. For example, each earthquake gives rise to aftershock activity in the neighboring areas depending upon the magnitude of the earthquake. Hawkes (1971) proposed a process to model such a branching structure of temporal events which was named Self-Exciting Point Process (SEPP) which implemented the idea that an earthquake can trigger aftershocks. As it can be observed from historical data on earthquakes, it is obvious to observe that the earthquakes with high intensity induces more aftershocks as compared to those with low intensity, that is, the earthquake and its aftershock sequence shows an epidemic behavioral pattern. Ogata (1988) developed a point process model to model such an epidemic behavior of earthquakes and named it Epidemic Type Aftershock Sequences (ETAS) model which is an extension of SEPP model.

Another point process model that is specifically defined to model the earthquake occurrences is a 'stress release model' (SRM) proposed by Knopoff (1971) and Vere-Jones (1978). The SRM is considered to be a stochastic version of the so-called 'elastic rebound theory'. It considers the increased pressure in an area and the pressure released during an earthquake over a period of time. It is possible to measure the probability of earthquake occurrence using the conditional intensity function of the SRM.

Also the historical data of various earthquakes demonstrates that the high intensity earthquakes are normally followed by high intensity earthquakes in surrounding area and this resists the aftershocks. The interactions among such affected areas have influence on time and intensity of earthquake occurrences in the presence of stress movement. Bebbington and Harte (2001) proposed Linked Stress Release Model (Linked SRM) to model temporal events which exhibits such a characteristic.

Hawkes and Oakes (1974) defined cluster point process assuming that, at every point in temporal point process on $(0, \infty)$, a cluster of activities starts. This point can be interpreted as the first arrival time of a point process which triggers a random stream of events of the similar types. This model exhibits chain ladder type of characteristic which may be used to forecast the total number of events and the time of occurrence of events in the future course of time. Vere-Jones and Davies (1966) and Vere-Jones (1970) proposed trigger models which assumes that a series of primary events (say, main shocks) is distributed completely randomly in time and each of these primary events are capable of generating a series of secondary events (say, aftershocks).

Sometimes the temporal events might be just one of the components of the complex model which is carrying much more information apart from just the time of occurrence of events under study. This additional information may themselves have a stochastic structure and stochastic dependency. Daley and Vere-Jones (2003) classified such temporal point processes as marked point processes (MPP) or marked temporal point processes (MTPP). The behavior of the MTPP may also be described with the help of the conditional intensity function, conditional on the history of the process over time, having two components: the ground intensity function and the mark distribution. The ground intensity function describes the rate of occurrence of events with respect to time and the mark distribution describes the behavior of other variables (referred to as marks), that are associated with the event, and will also be usually dependent on the history of the occurrence of events.

The remaining part of this paper is organized this way: Section 2 offers various models and its technical descriptions. The likelihood estimation and related inferences are presented in

Section 3, followed by data analysis in Section 4. The Section 5 concludes with some useful discussion including the limitations.

2. Models and Descriptions

In case of earthquake occurrences, it can be commonly observed that an earthquake with very high magnitude is usually followed by a sequence of aftershocks and some of the aftershocks themselves may also be of such a high magnitude that they might have aftershocks caused by them. Obviously, both the frequency and magnitude diminish over a period of time, or with respect to space (distance from epicenter) etc. Also prior to a major earthquake or aftershock, pre-seismic activities and foreshocks might take place. But foreshocks are not very easy to identify as compared to aftershocks. Some researchers made attempts to study pre-seismic quiescence and the gaps between two major earthquakes of aftershocks. On the other hand, some researchers were of the opinion that the pre-seismic quiescence and the gaps are nothing but the result of decaying activity of aftershocks which were followed by the last major shock and so they are not of much importance to be studied. Thus, most of the studies focus on the main shocks and their aftershock sequences. In the context of earthquake studies, we consider the time points at which the main shock or major aftershocks occur as temporal points, or equivalently, the times of occurrence of main shocks or major aftershocks as temporal events.

2.1. NHPP models

In general, temporal events may be statistically independent or temporally correlated. Let $N(t_i, t_j)$ represents the number of events occurring between time t_i and t_j and let $(t_1, t_2), (t_3, t_4), \ldots, (t_{k-1}, t_k)$ be the disjoint sets where $t_1 < t_2 \leq \ldots \leq t_{k-1} < t_k$ are times of occurrence of temporal events, then N will be a Poisson process, if each of $N(t_1, t_2), N(t_3, t_4), \ldots, N(t_{k-1}, t_k)$ have a Poisson distribution and are independent, that is, $Cov[N(t_1, t_2), N(t_2, t_3)] = 0$ for any $t_1 < t_2 < t_3$. A Poisson process always has a deterministic conditional intensity $\lambda(t)$. If it is stationary, then $\lambda(t)$ is constant and the process is known as homogeneous Poisson process (HPP); otherwise the process is known as NHPP.

NHPP with a power-law intensity function is a frequently used model for temporal point process. According to the magnitude of the power law, one can ascertain whether the rate of occurrence of an event is decreasing, constant or increasing function of time. Crow (1974) proposed a model for which system failure times are assumed to occur according to a time dependent Poisson process with a Weibull intensity function of the form,

$$\lambda(t) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1}, \text{ where } \theta > 0 \; and \; \beta > 0. \tag{2}$$

Such a model is referred to as the *Weibull Process* or *Power Law Process* (PLP) and is used to model reliability growth. Another useful NHPP model was proposed was proposed by Cox and Lewis (1966) with intensity function of the form $\lambda(t) = e^{\alpha + \beta t}$, which is referred to as Log-Linear Process (LLP). This kind of models work well for any temporal event data exhibiting stationary characteristics in the long run.

2.2. Self-exciting point process models

When the temporal events are correlated, an initial event can increase or decrease the chances of occurrence of similar events in future. Hawkes (1971) introduced the *self-exciting process* as a point process for which $Cov[N(t_1, t_2), N(t_2, t_3)] > 0$ for any $t_1 < t_2 < t_3$. This means that if an event occurs, another event becomes more likely to occur in time and space (Hawkes and Oakes (1974), Daley and Vere-Jones (2003)). Hawkes (1971) defined the self-exciting process with an intensity function has the form:

$$\lambda(t) = \nu + \int_0^t g(t-s)\, dN(s) = \nu + k_0 \sum_{t_k<t} g(t-t_k) \tag{3}$$

where $\nu > 0$ is a baseline intensity or the background rate of events which is assumed to be constant in time and the second term describes the self-exciting part of the process having two components k_0 and g; k_0 reflects the magnitude of self-excitation and the function g measures the influence of an event on the intensity process or density at which self-excitation is triggered. For the earthquake study, the pre-seismic activities may be considered as the baseline intensity ν; intensity of the main shock which triggers the aftershocks may be considered as the magnitude of self-exciting part, that is k_0 and g is the density function at which self-excitation is triggered. It is assumed that $g(x) \geq 0$ for all $x \geq 0$, $g(x) = 0$ for $x < 0$, and $\int_0^t g(u)\, du < 1$, where t is the time up to which the events are observed. Many forms of $g(t)$ have been proposed and studied in literature, and in most of the cases, the choices and importance of the density depends on the situations and contexts. Egesdal et. al. (2010) considered g as an exponential distribution in their study, and they considered the intensity function having the form

$$\lambda(t) = \mu + k_0 \sum_{t_i<t} w\, e^{w(t_i - t)} \tag{4}$$

This intensity function describes the rate at which events occur over time, and is not only influenced by the current time, but also by the events that have occurred before the current time. The current events subsequently decrease over time exponentially. While studying earthquake and its aftershocks, similar pattern is likely to be observed and therefore it might be reasonable to consider the above form of intensity function while using SEPP to model earthquake occurrences.

2.3. ETAS model

The models with intensity function as proposed in SEPP were considered by Ogata (1988) for modeling the data regarding earthquake and its aftershocks, where it was assumed that the earthquake aftershock sequences can be modeled like an epidemic, that is, the earthquakes with larger magnitudes might have a sequence of more aftershocks in a given interval of time and also it may continue for a longer time after the main shock. He named such models as *Epidemic Type Aftershock-Sequences* (*ETAS*) models. Ogata (1988) studied various statistical models for the standard activity of sequence of earthquakes and compared them using likelihood methods. He proposed Epidemic-Type models with reference to the age-dependent birth and death process introduced by Kendall (1949), in which only the births or events are allowed to occur at a constant rate per unit time according to Poisson process and with each birth or event (with

reference to earthquake study, main shock) there is associated a cluster of subsidiary events (say, aftershocks) formed by the births of all of the descendants of all generations of the immigrant (Hawkes and Oakes, 1974). Thus one of the differences between the epidemic-type model and the trigger model is that the trigger model assumes only the first generation offspring whereas the epidemic-type model assumes that each event has the possibility of possessing offspring. Ogata (1988) in his study proved that the epidemic-type models that include the effect of magnitude of earthquake give a better fit to the data than any of the trigger models (considered in a restricted form).

These models can be constructed by assuming two activities associated with each occurrence of an earthquake, background events and aftershock events. It is assumed that the background events occur independently according to a stationary Poisson process $\mu(x, y)$, with magnitudes distributed independently of μ and each occurrence of an earthquake increases the risk of aftershocks. Also, it is reasonable to assume that the increased risk of aftershocks also spreads in the neighboring locations in space and time according to the kernel $g(t)$.

Ogata (1988) considered the model (4) with $(t) = \frac{K}{(t+c)^p}$, as defined by the modified Omori formula (Utsu, 1961), where K, c and p are parameters, t is the time since occurrence of shock. The number K depends on the lower bound of the magnitude of aftershocks counted in $N(t)$, whereas c and p are known to be independent of the choice of lower bound. The epidemic-type model is defined in terms of the conditional intensity rate, or seismic risk as a function of time, based on the following assumptions: (a) the background seismic activity is generated according to a stationary Poisson process with a constant hazard rate (b) each shock has a risk of stimulating aftershocks in proportion of the quantity $e^{\beta M}$, where M is the magnitude of the main shock, and (c) the hazard rate of aftershocks decreases with time according to the modified Omori law, $\frac{K}{(t+c)^p}$.

Ogata (1985, 1988, 1989) demonstrated that the ordinary seismic activity of a wide region can be described in terms of the conditional intensity by the superposition of a constant rate for background seismicity and the modified Omori functions of any shocks i which occurred at time t_i, in such a way that

$$\lambda(t \mid \mathcal{H}_t) = \mu + \sum_{t_i<t} \frac{K_i}{(t-t_i+c)^p} \tag{5}$$

where μ is the rate of occurrence of the background seismic activity. The sum $\sum_{t_i<t}$ is taken for all shocks i which occurred before time t, and the parameter K_i for each shock i contributes to the size of the corresponding aftershock. More importantly the parameter K_i is dependent on the magnitude M_i of the aftershock as well as the cut-off or threshold magnitude M_0 of the data set according to the exponential function form

$$K_i = A\, e^{\alpha(M_i - M_0)} \tag{6}$$

Above form is based on the empirical formula obtained by Utsu and Seki (1955) regarding the linear relation between the logarithms of aftershock areas and the magnitudes M of the main shock. It suggests that the number N of aftershocks with magnitudes over a threshold M_0 for a

fixed time span can be roughly estimated as proportional to $exp(\alpha\ (M - M_0))$ for a constant α. Model (5) with (6) for the ordinary seismicity in terms of the rate of occurrence of aftershocks is an ETAS model with $\mathcal{H}_t = \{(t_i, M_i);\ t_i < t\}$ as the history of occurrence times $\{t_i\}$ up to time t and their corresponding magnitudes $\{M_i\}$, with intensity function

$$\lambda(t \mid \mathcal{H}_t) = \mu + A \sum_{t_i<t} e^{\alpha(M_i - M_0)} \left(\frac{c}{t - t_i + c}\right)^p \tag{7}$$

The ETAS model assumes that the aftershocks are generated as events occurring according to Poisson process with rate μ. The parameters (μ, A, α, c, p) are all positive, t_i is the time of occurrence of the i^{th} event with magnitude M_i and M_0 is a threshold magnitude of aftershocks considered in the study. The term $e^{\alpha(M_i - M_0)}$ conveys the meaning that the main shock or aftershock with larger magnitude raise the intensity of occurrence of aftershocks more, and the term $\left(\frac{c}{t - t_i + c}\right)^p$ determines the length (time) till which the aftershock sequence will continue to occur. There will have to be put certain constraints on parameters, otherwise the aftershock sequence (epidemic) could explode and never die out. The parameters α and p characterizes the temporal pattern of seismicity. The parameter p represents the rate at which the aftershock sequence decays, and the parameter α represents the vulnerability of magnitude of an earthquake in generating the aftershocks. Ogata (1987) also proved that the swarm-type activity has a smaller value of α, that is, a small α value indicates the presence of main shock and aftershock activity whereas the larger value of α indicates that there are only few large aftershocks or magnitude of main shock is much larger than the maximum magnitude of aftershocks. Thus it can be said that α^{-1} represents the average time until a next aftershock occurs.

The appropriate selection of parameter values is very crucial part of the modeling process. The distance in space and time over which the risk of main shocks or aftershocks spreads, the number of aftershocks, the dependence of the increased risk of aftershocks on magnitude size of main shock, etc. can have great impact on the power of a point process model in predicting the space and time of next major earthquake.

2.4. Stress release models

The elastic rebound theory proposed by Reid (1910) is a classical model for earthquake mechanisms which postulates that elastic stress in a seismically active region accumulates due to movement of tectonic plates, and is released when the stress exceeds the strength of the medium. Thus this theory suggests that a large earthquake should be followed by a period of quiescence (passive period), whereas in reality a strong earthquake can be followed by a period of activation (another earthquake of comparable magnitude). This elastic stress within a region can be extracted by collecting various kinds of information but obtaining the information regarding the temporal variations of seismic activity might be the more suitable approach as it is expected to reflect directly the nature of earthquake generating stress. Vere-Jones (1978) proposed the stress release model as a stochastic version of the elastic rebound theory, incorporating the deterministic stress build-up within a region and its stochastic release through earthquakes, by developing the stochastic model for the occurrence of sequence of main shocks which was

proposed by Knopoff (1971). Many researchers applied stress release model to analyze several historical earthquakes, particularly to identify statistically distinct regions to which different stress models can be applied. One of the most interesting finding from these studies is that large earthquakes are often found to be followed by large earthquakes quite distant from the first. This seems consistent with a general consensus that the earthquakes taking place in the Earth's crust forms a tightly linked, near-critical process that exhibits the self-similarity, long-range correlation and power-law distributions. Thus it can be considered that there exists a class of models exhibiting self-organized criticality due to competition between local strengthening and weakening through interactions.

Suppose ρ is the loading rate which describes stress accumulation caused by large-scale tectonic plate movement, which is assumed to be constant and positive and $X(0)$ is the initial pressure, which is assumed to be positive. The stress is released when it exceeds the energy limit of plate strength in the form of earthquakes. Stress will increase linearly with the loading rate ρ at time t with initial stress $X(0)$ and thus the following equation can be obtained,

$$X(t) = X(0) + \rho t \tag{8}$$

Let t_i and S_i be the time and stress released, respectively, at the time of occurrence of event $i; i = 1, 2, \dots, n$ where n is the number of events occurred in the time interval $(0, t)$ under consideration. The value of stress S_i is correlated with the corresponding magnitude which is proportional to the seismic energy which is released during the occurrence of earthquake. Then the accumulated stress released by all events in the period $(0, t)$ can be expressed as $S(t) = \sum_{i;\, t_i<t} S_i$. An important variable in the stress release model is the level of stress in a certain area which controls the occurrence of an earthquake. The stress level at time t increases deterministically and decreases stochastically due to an earthquake. Therefore the stress released at time t, $X(t)$, is the difference between the accumulated stress that increases linearly with the loading rate ρ and the accumulated stress that is released in the period $(0, t)$, that is,

$$X(t) = X(0) + \rho\, t - S(t) \tag{9}$$

This is referred to as a Stress Release Model.

Further, the conditional intensity function of the above model can be obtained through its hazard function, say $\Psi(x)$, which denotes the probability of occurrence of an earthquake in the time interval $(t, t + \Delta t)$. Assuming that the hazard function is having an exponential form

$$\Psi(x) = \exp(\alpha + \beta x)\,;\ \alpha \in R, \beta \geq 0 \tag{10}$$

where the parameter α describes the initial stress value and parameter β describes the combination of strength and heterogeneity of the earth's crust in the area. Using equation (10), the conditional intensity function of the stress release model with the history $\mathcal{H}_t = \{(t_i, M_i);\ t_i < t\}$ is defined as

$$\lambda(t \mid \mathcal{H}_t) = \Psi\big(\mathrm{X(t)}\big) = exp\big(\alpha + \beta\,(X(0) + \rho\, t - S(t))\big) \tag{11}$$

Taking $\alpha + \beta X(0) = a$, $\beta \rho = b$ and $1/\rho = c$, equation (11) reduces to

$$\lambda(t \mid \mathcal{H}_t) = exp\big(a + b\,(t - c\,S(t))\big) \tag{12}$$

2.5. Linked stress release model

Apart from aftershocks, large events are often followed by large events quite distant from the first. On the other hand, very large events can resist occurrence of subsequent events or aftershocks on the same or other fault lines. Thus, interaction among areas can influence the time and magnitude of earthquake occurrence. Also, the regional stress itself evolves over centuries in the area which might be the result of the cumulative effects of all previous earthquakes and changes in tectonic loading. But the presence of stress movement and interaction among areas has not been incorporated in stress release model. The omission of this interaction between areas might underestimate the activity by the time and magnitude-predictable stress release model. These shortcomings of stress release model motivated its extension to include interactions among areas, by means of stress transfer and reduction and were named as Linked stress release model [Lu et al. (1999), Bebbington and Harte (2001)].

Zheng and Vere-Jones (1994) found that large geographical regions give better fits to the stress release model when broken down into subunits, and further noted some hints of clustering relating to some form of action at a distance, i.e. stress transfer and interaction. Thus, the multivariate extension to the stress release model was proposed as a linked stress release model by defining the evolution of stress $X_i(t)$ in the i^{th} region as

$$X_i(t) = X_i(0) + \rho_i\, t - \sum_j \theta_{ij}\; S^{(j)}(t) \tag{13}$$

where $S^{(j)}(t)$ is the accumulated stress release in region j over the period $(0, t)$, and the coefficient θ_{ij} measures the fixed proportion of stress drop, initiated in region j, which is transferred to region i. Here, θ_{ij} may be positive or negative, resulting in damping or excitation respectively. It is convenient to set $\theta_{ii} = 1$ for all i while ignoring aftershocks. If $\theta_{ij} = 0$ for all $i \neq j$, the model is reduced to an independent combination of simple forms as that of stress release model. Then the point process conditional intensity function of the linked stress release model will be of the form

$$\lambda_i(t) = \Psi\left(X_i(t)\right) = exp\left(\alpha_i + \beta_i\left(X_i(0) + \rho_i t - \sum_j \theta_{ij}\; S^{(j)}(t)\right)\right) \tag{14}$$

for each region i, where α_i, β_i, ρ_i and θ_{ij} are the parameters to be estimated. The above form of intensity function can be re-parameterized as

$$\lambda_i(t) = exp\big(a_i + b_i\big(t - \sum_j c_{ij}\; S^{(j)}(t)\big)\big) \tag{15}$$

with $a_i = \alpha_i + \beta_i\, X_i(0)$, $b_i = \beta_i\, \rho_i$ and $c_{ij} = \theta_{ij}/\rho_i$.

2.6. Marked point processes

The conditional intensity function of a point process, conditional on the past history of the temporal events, describes the instantaneous Poisson rate. Suppose we consider the events that are occurring in two-dimensional space, for example longitude and latitude in case of study regarding earthquakes, according to time. Then the history up to but not including time t of events may be denoted by $\mathcal{H}_t$ and may be defined as $\mathcal{H}_t = \{(t_i, x_i, y_i)$ for all i for which $t_i < t\}$. Here t_i is the time of occurrence of the i^{th} event and (x_i, y_i) is its spatial location, that is, longitude and latitude of the location of main shock or aftershock. Then the conditional intensity function as described in (1) may be written as

$$\lambda(t, x, y \mid \mathcal{H}_t) = \lim_{\delta,\xi,\eta \to 0} \frac{1}{\delta\xi\eta} \; P\left[\mathrm{N}_{\delta\xi\eta}(t, x, y) > 0 \mid \mathcal{H}_t\right] \tag{16}$$

where $\mathrm{N}_{\delta\xi\eta}(t, x, y)$ is the number of aftershocks occurring in the space $[t, t + \delta) \times [x, x + \xi) \times [y, y + \eta)$. As stated in section 1, the conditional intensity function of the marked point process should have two components, the ground intensity function which describes the rate at which the aftershocks occur over time and the history of main shock or aftershocks which occurred before the current time and mark distribution. Let $\mathrm{N}_\delta(t)$ be the number of aftershocks in time interval $[t, t + \delta)$, then the ground intensity function for the marked point process can be defined as

$$\lambda_g\left(t, \underline{\theta} \mid \mathcal{H}_t\right) = \lim_{\delta \to 0} \frac{1}{\delta} \, P[\mathrm{N}_\delta(t) > 0 \mid \mathcal{H}_t] \tag{17}$$

where $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta_{\mathrm{m}}$ are the parameters. $\lambda_g\left(t, \underline{\theta} \mid \mathcal{H}_t\right)$ can also be simply denoted as $\lambda_g(t \mid \mathcal{H}_t)$. Suppose the mark distribution of the variables x (univariate or multivariate) considered as marks may be denoted as $f(x \mid \mathcal{H}_t)$. Then the conditional intensity function of the marked point process may be considered to have form

$$\lambda(t, x \mid \mathcal{H}_t) = \lambda_g(t \mid \mathcal{H}_t) \, f(x \mid \mathcal{H}_t) \tag{18}$$

In the next section, we consider the estimation of model parameters in a couple of models described above.

3. Likelihood Functions and Parameter Estimation

The likelihood function for the temporal point process models where the conditional intensity function is a function of time only, can be derived in the following manner: Let τ be the time of the occurrence of the last event before time *t*, $F(t \mid \mathcal{H}_\tau) = P\{T \geq t \mid \mathcal{H}_\tau\}$ denote the conditional distribution of the time of occurrence of the next event after time *t* and $f(t \mid \mathcal{H}_\tau)$ be the corresponding conditional density function. Then

$$\lambda(t \mid \mathcal{H}_\tau) = \frac{f(t \mid \mathcal{H}_\tau)}{1 - F(t \mid \mathcal{H}_\tau)}$$

Solving the differential equation,

$$F(t\,|\mathcal{H}_\tau) = 1 - exp\left\{-\int_\tau^t \lambda(u \mid \mathcal{H}_\tau)\, du\right\},$$

hence

$$f(t\,|\mathcal{H}_\tau) = \lambda(t \mid \mathcal{H}_\tau)\, exp\left\{-\int_\tau^t \lambda(u \mid \mathcal{H}_\tau)\, du\right\}$$

Let $\ldots < t_{-2} < t_{-1} < t_0 < T_1 < t_1 < t_2 < \cdots < t_n < T_2 < t_{n+1} < t_{n+2} < \cdots$, where $t_i\,; i \in \mathbb{Z}$ are times of occurrence of main shock or aftershocks, and those main shock or aftershock events which occurs within the time interval $[T_1, T_2]$ are explicitly included in the likelihood. The events which occurred before time T_1, if any, are included in the history of the process. Then the log likelihood function of such models will be

$$\log L = \sum_{i:T_1 \le t_i \le T_2} log\, \lambda\big(t_i \mid \mathcal{H}_{t_i}\big) - \int_{T_1}^{T_2} \lambda(t \mid \mathcal{H}_t)\, dt$$

The likelihood function for the marked point process models for which the conditional intensity function is as defined in (18), can be derived similarly as above and will be of the form

$$\log L = \sum_{i:\, t_i \in \mathcal{T}} log\, \lambda\big(t_i, x_i, y_i \mid \mathcal{H}_{t_i}\big) - \int_{\mathcal{T}} \int_{\mathcal{Y}} \int_{\mathcal{X}} \lambda(t, x, y\,|\mathcal{H}_t)\; dx\, dy\, dt$$

where $\mathcal{T} \subseteq \mathbb{R}^+$ is a time interval and $\mathcal{X}$ and $\mathcal{Y}$ are the domains of x and y, respectively which represents the mark variables (for example, longitudes and latitudes in the earthquake study). This process can also be extended to have more than two mark variables and in that case the log likelihood function can be considered as

$$\log L = \sum_{i:\, t_i \in \mathcal{T}} log\, \lambda\big(t_i, \underline{x}_i \mid \mathcal{H}_{t_i}\big) - \int_{\mathcal{T}} \int_{\mathcal{X}} \lambda\big(t, \underline{x}\,|\mathcal{H}_t\big)\; d\underline{x}\, dt$$

where $\underline{x}$ is a multivariate variable referring the mark variables. The general form of the log likelihood function of a marked point process can be expressed as

$$\log L = \sum_{i:\, t_i \in \mathcal{T}} log\, \lambda_g\big(t_i \mid \mathcal{H}_{t_i}\big) - \int_{\mathcal{T}} \lambda_g(t\,|\mathcal{H}_t)\, dt + \sum_{i:\, t_i \in \mathcal{T}} \log f\big(\underline{x}_i \mid \mathcal{H}_{t_i}\big)$$

In different studies the mark density $f\big(\underline{x} \mid \mathcal{H}_t\big)$ is taken as exponential or gamma or Weibull etc. The model parameters can be estimated using the method of maximum likelihood (MLE), by obtaining the system of likelihood equations simultaneously by differentiating the log

likelihood function with respect to the parameters to be estimated and equating those equations to zero. The system of likelihood equations for the models discussed in this study does not possess the explicit solution and therefore a suitable iterative method for nonlinear optimization may have to be adopted for solving them.

4. Data Analysis

As an illustration, we consider the data of Kathmandu earthquake happened in the year 2015. The deadly earthquake of magnitude 7.9 measured on Richter scale, shook Nepal and sent tremors through Indian subcontinent, on April 25, 2015 at 11:56 (Nepalese time). The Complete description about the data, data set, and many inference-based studies on this data are available in Shah et al. (2019). The model check and validation were carried out using information functions like: Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Goodness-of-fit tests based on power law intensity and exponential intensity failed to fit the data. Therefore, we have fitted marked temporal process models with ETAS and SEP models as the ground intensity functions, considering marks as the magnitudes of earthquakes or aftershocks with intensity more than 5. In both the cases, the underlying distribution is assumed as Gamma. The summary of estimates is presented in Table 1.

Table 1. Parameter Estimates for Gamma model

Ground intensity function			
ETAS		**SEP**	
Parameters	**Estimates**	**Parameters**	**Estimates**
μ	0.001	μ	0.028
A	0.048	k_0	2.800
α	1.963	w	0.067
c	0.769	θ	0.400
p	1.305	β	2.700
θ	2.519	$\log L$	-339.770
β	0.169	AIC	689.540
$\log L$	-193.860	BIC	699.570
AIC	401.720		
BIC	415.769		

5. Conclusions and Future Directions

Temporal data are very sensitive to the assumptions underlying the situation. So one has to attempt to fit different models satisfying the history of events and the event occurrence mechanism, and then choose the best one. Among various competing models it might be possible that some other model still exists which provides the better fit to the data than the models which have already considered. Thus, it is logical to check whether the major features of the given data

are captured and reproduced by the assumed models. Graphical procedures are developed for intensifying the features of the data that deviate from the model, if any, as follows: If $\hat{\lambda}_g(t \mid \mathcal{H}_\tau)$ is the estimated ground intensity function which is fitted on the given data, after obtaining the MLEs of parameters of the temporal point process, then the transformed time points may be assumed to follow a stationary Poisson process with rate one. Thus if $t_i; i = 1, 2, \ldots, n$ are the event times, then the sequence of transformed times τ_i will be $\tau_i = \int_0^{t_i} \hat{\lambda}_g(t \mid \mathcal{H}_\tau)\,\mathrm{dt}$. A deviation from a property of $\{\tau_i\}$ from that which is expected from a stationary Poisson process implies the existence of a corresponding feature of the data $\{t_i\}$ that is not captured by the underlying model. The intensity $\hat{\lambda}_g(t \mid \mathcal{H}_\tau)$ represents a model for prediction, whereas the transformed data $\{\tau_i\}$ may be regarded as "noise" or "residuals" of the point process data $\{t_i\}$. This sequence of $\{\tau_i\}$ is referred to as the residual process by Ogata (1988). This is under investigation.

To check the goodness of fit of the model, the event number i can be plotted versus the transformed time τ_i. The plotted points should approximately follow a straight line. Significant departure from the straight line indicates a weakness in the model. Moreover, the slope of the line less than one implies that the transformed times τ_i are too small indicating that the fitted ground intensity function $\hat{\lambda}_g(t \mid \mathcal{H}_\tau)$ is too small, and the slope of the line greater than one implies that $\hat{\lambda}_g(t \mid \mathcal{H}_\tau)$ is too large. The changing pattern of ground intensity function is also an eye opener for understanding the seismological fluctuations of the occurrence. This is also under investigation. Although challenges are many, with the computing power of complex models, this issue can be resolved.

References

Ascher H. and Feingold H. (1984). *Repairable Systems Reliability – Modeling, Inference, Misconceptions and Their Causes.* Marcel Dekker, New York.

Bassin W. M. (1973). A Bayesian optimal overhaul interval model for the Weibull restoration process. *Journal of the American Statistical Association*, **68**, 575-578.

Bassin W. M. (1969). Increasing hazard functions and overhauls policy. *ARMS, IEEE-69 C 8-R*, 173-180.

Bebbington M. and Harte D. (2003). The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications. *Geophysical Journal International*, **154**, 925-946.

Bebbington M. and Harte D. (2001). On the statistics of the linked stress release process. *Journal of Applied Probability*, **38A**, 176-187.

Cox D. R. and Lewis P. A. (1966). *The Statistical Analysis of Series of Events*. Metheun, London.

Crow L. H. (1974). Reliability analysis for complex repairable systems. In *Reliability and Biometry*, Eds. F. Proschan and R. J. Serfling, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 379-410.

Daley D. J. and Vere-Jones D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. Second Edition. Springer-Verlag, New York. ISBN 0-387-95541-0.

Duane J. T. (1964). Learning curve approach to reliability monitoring. *IEEE Transactions*, **A-2**, 563-566.

Egesdal M., Fathauer C., Louie K. and Neuman J. (2010). *Statistical and Stochastic Modeling of Gang Rivalries in Los Angeles*. SIAM Undergraduate Research Online.

Hawkes A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society*, **B33**, 438-443.

Hawkes A. G. and Oakes D. A. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, **11**, 493-503.

Kendall D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society*, **B11**, 230-264.

Knopoff L. (1971). A stochastic model for the occurrence of main sequence events. *Reviews of Geophysics and Space Physics*, **9**, 175–188.

Lewis P. A. W. (1972). Recent results in the statistical analysis of univariate point processes. In *Stochastic Point Processes*, Ed. P.A.W. Lewis, 1-54. New York: Wiley.

Lewis P. A. W. (1970). Remarks on the theory, computation and application of the spectral analysis of series of events. *Journal of Sound and Vibration*, **12**, 353-375.

Lu C., Harte D. and Bebbington M. (1999). A linked stress release model for historical Japanese earthquakes: coupling among major seismic regions. *Earth Planets Space*, **51**, 907–916.

Ogata Y. (1989). Statistical model for standard seismicity and detection of anomalies by residual analysis. *Tectonophysics*, **169**, 159-174.

Ogata Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, **83**(**401**), 9-27.

Ogata Y. (1987). *Long Term Dependence of Earthquake Occurrences and Statistical Models for Standard Seismic Activity (in Japanese)*. Suri Zisin Caku (Mathematical Seismology) II (Ed. M. Saito), ISM Cooperative Research Report 3, 115-124, Institute of Statistical Mathematics, Tokyo.

Ogata Y. and Tanemura M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, **41**, 421-433.

Reid H. F. (1910). The mechanism of the earthquake. In *The California Earthquake* of April 18, 1906, Report of the State Earthquake Investigation Commission, 2, 16–28, Carnegie Institute of Washington, Washington, DC.

Shah R., Muralidharan K. and Parajuli A. (2019). Temporal point process models for Nepal earthquake aftershocks. *International Journal of Statistics and Reliability Engineering* (to appear).

Utsu T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine*. 30, 521-605.

Utsu T. and Seki A. (1955). Relation between the area of aftershock region and the energy of the main shock, *Zisin (Journal of the Seismological Society of Japan)*, 2nd Series, ii. 7, 233-240.

Vere-Jones D. (1978). Earthquake prediction - a statistician's view. *Journal of Physics of the Earth*, **26**, 129–146.

Vere-Jones D. (1970). Stochastic models for earthquake occurrence (with discussion). *Journal of the Royal Statistical Society*, **B32**, 1-62.

Vere-Jones D. and Davies R. B. (1966). A statistical survey of earthquakes in the main seismic region of New Zealand. Part 2. Time series analyses. *New Zealand Journal of Geology and Geophysics*, **9**, 251-284.

Zheng X. and Vere-Jones D. (1994). Further applications of the stochastic stress release model to historical earthquake data. *Tectonophysics*, **229**, 101-121.

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 25-38

District-level Estimates of Extent of Food Insecurity for the State of Uttar Pradesh in India by Combining Survey and Census Data*

Hukum Chandra
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Received: 18 April 2020; Revised: 26 April 2020; Accepted: 02 May 2020

Abstract

This paper describes small area estimation (SAE) method that incorporates the sampling information when estimating small area proportions. This method is applied to estimate the incidence of food insecurity in different districts of rural areas of the state of Uttar Pradesh in India by linking data from the 2011-12 Household Consumer Expenditure Survey collected by the National Sample Survey Office of India and the 2011 Population Census. A map showing district level inequalities in the distribution of food insecure households in Uttar Pradesh is also produced which provides an important information for analysis of spatial distribution of food insecurity in the state.

Key words: Food insecurity; SDG; Small area estimation; Precise, Representative.

1. Introduction

The food security is one of the highest priority of the Government of India to achieve the Sustainable Development Goal 2. In India, the Household Consumer Expenditure Survey (HCES) data collected by National Sample Survey Office (NSSO), Ministry of Statistics and Program Implementation, Government of India is used to generate the estimates of food insecurity indicators at state and national level for both rural and urban sectors separately. In spite of high importance, the estimates of food insecurity indicators are not available at local area or lower administrative unit (*e.g.* district) level in the country. Policy planners, researchers, government and public agencies are more and more interested in obtaining statistical summaries for smaller domains called small areas, created by cross classifying demographic and geographic variables such as small geographic areas (*e.g.* districts) or small demographic groups (*e.g.* age-sex groups, land category, social groups) or a cross classification of both. However, the sample sizes for such small areas in the existing large scale survey data (*e.g.* HCES in India) may be very small or even zero. The SAE methodology provides a viable and cost effective solution this problem of small sample sizes (Rao and Molina, 2015). The SAE methods produce reliable estimates

Corresponding Author: Hukum Chandra
Email: hchandra12@gmail.com

for small areas with small sample sizes by borrowing strength from data of other areas, other time periods or both.

The SAE methods are generally based on model-based methods. The idea is to use statistical models to link the variable of interest with auxiliary information, *e.g*. Census and Administrative data, for the small areas to define model-based estimators for these areas. Based on the level of auxiliary information available, the models used in SAE are categorized as area level or unit level. Area-level modelling is typically used when unit-level data are unavailable, or, as is often the case, where model covariates (*e.g.* census variables) are only available in aggregate form. The Fay–Herriot model (Fay and Herriot, 1979) is a widely used area level model in SAE that assumes area-specific survey estimates are available, and that these follow an area level linear mixed model with area random effects, Chandra (2013) and Chandra *et al* (2015). Standard SAE methods based on linear mixed models for continuous data can produce inefficient and sometime invalid estimates when the variable of interest is binary. If the variable of interest is binary and the target of inference is a small area proportion (*e.g.* for estimating food insecurity proportions), then the generalized linear mixed model with logit link function, also referred as the logistic linear mixed model (LLMM)) is generally used. An empirical plug-in predictor (EPP) under a LLMM is commonly used for the estimation of small area proportions, see for example, Chandra *et al*. (2012), Rao and Molina (2015) and references therein, although it is not the most efficient predictor under that model. An alternative to EPP is the empirical best predictor (EBP, Jiang, 2003). This predictor does not have a closed form and can only be computed via numerical approximation. This is generally not straightforward, and so national statistical agencies favour computation of an approximation like the EP.

In this context, when only area level data are available, an area level version of a LLMM is used for SAE, see for example, Johnson *et al*. (2010), Chandra *et al*. (2011), Chandra *et al*. (2017), Chandra *et al*. (2018), Anjoy *et al*. (2020). Unlike the Fay-Herriot model, this approach implicitly assumes simple random sampling with replacement within each area and ignores the survey weights. Unfortunately, this has the potential to seriously bias the estimates if the small area samples are seriously unbalanced with respect to key population characteristics, and consequently use of the survey weights appears to be inevitable for if one wishes to generate representative small area estimates. Chandra *et al*. (2019) deliberated the idea of Korn and Graubard (1998) and model the survey weighted estimates as binomial proportions, with an "effective sample size" chosen to match the binomial variance to the sampling variance of the estimates. Using the effective sample size rather than the actual sample size allows for the varying information in each area under complex sampling. This article considers Chandra *et al*. (2019) approach to model survey weighted small area proportions under a LLMM and attempts to produce the district level estimates of proportion of food insecurity (also refers as food insecurity prevalence or incidence of food insecurity) for rural areas of Uttar Pradesh. Throughout this article, proportion of food insecurity, food insecurity prevalence and incidence of food insecurity will be used interchangeably. The state of Uttar Pradesh is the most populous state in the country and accounts for about 16.16 percent of India's population. It covers 243,290 square km, equal to 6.88% of the total area of the country. The analysis is restricted to rural areas of Uttar Pradesh because about 78% of the population of the State live in rural areas according to 2011 Population Census.

Rest of the article is organized as follows. Next Section describes the data from the 2011-12 HCES of the NSSO and the 2011 Population Census that will be used to estimate the district-wise proportion of household food insecurity for rural areas of Uttar Pradesh. Section 3 presents the SAE methodology. The empirical results and a map showing district-level inequalities in the distribution of food insecurity in rural Uttar Pradesh along with various diagnostic measures are reported in Section 4. Finally, Section 5 provides concluding remarks.

2. Data and Model Specification

This section introduces the basic sources of the data, *i.e*. the 2011-12 HCES of the NSSO for rural areas of Uttar Pradesh and the 2011 Population Census, used in SAE application reported in this paper. Data obtained from these sources are then used to estimate the proportion of food insecurity (or incidence of food insecurity) at district level in Uttar Pradesh. The NSSO conducts nationwide HCE surveys at regular intervals as part of its "rounds", with the duration of each round normally being a year. The surveys are conducted through interviews of a representative sample of households selected randomly through a suitable sampling design and covering almost the entire geographical area of the country. The sampling design used in the 2011-12 HCES is stratified multi-stage random sampling with districts as strata, villages as first stage units and households as second stage units. Although, these surveys provide reliable and representative state and national level estimates, they cannot be used directly to produce reliable estimates at the district level due to small sample sizes. Although district is a very important domain of the planning process in India, there are no surveys aimed at producing estimates at this level. The lack of robust and reliable outcome measures at the district level puts constraints on the design of targeted interventions and policy development. In the 2011-12 HCES, a total of 5916 households from the 71 districts of rural areas of Uttar Pradesh were surveyed. The district sample sizes ranged from 32 to 128 with average of 83. It is evident that these district level sample sizes are relatively small, with an average sampling fraction of 0.0002 (see Table 1). Due to this sample size limitation, it is challenging to generate reliable district level direct estimates with associated standard errors from this survey (Rao and Molina, 2015 and Chandra *et al.*, 2011). This paper addresses this small sample size issue in the 2011-12 HCES data for producing district level estimates by adopting SAE approach and using auxiliary information from the 2011 Population Census to strengthen the limited sample data from the districts.

Table 1: Summary of sample size, number of food insecure households in sample (sample count) and sampling fraction in 2011 HCES data

Features	Minimum	Maximum	Average	Total
Sample size	32	128	83	5915
Sample count	10	111	53	3778
Sampling fraction	0.00015	0.00032	0.00023	0.01647

The target variable *Y* at the unit (household) level in the 2011-12 HCES survey data file is binary, corresponding to whether a household is food insecure (household consuming less than 2400 Kcal per day) or not. Average dietary energy intake per person per day in rural India is 2400 kilocalorie (Kcal), as defined by the Ministry of

Health and Family Welfare, Government of India. The target is to estimate the proportion of rural households that are not getting satisfactory proportion of calories consistently at small area level, also referred to as the incidence of food insecurity or proportion of household food insecurity.

As noted above, the auxiliary variables used in this analysis are taken from the 2011 Population Census of India. These auxiliary variables are only available as counts at district level, and so SAE methods based on area level small area models must be employed to derive the small area estimates. There are nearly 30 such auxiliary variables that are available for use in SAE analysis. We, therefore, carried out an exploratory data analysis to choose few auxiliary variables to determine appropriate covariates for SAE modelling. We also employed Principal Component Analysis (PCA) to derive composite scores for some selected groups of variables. In particular, we did PCA separately on two groups of variables, all measured at district level and identified as S1 and S2 below. The first group (S1) consisted of the proportions of main workers by gender, proportions of main cultivators by gender and proportions of main agricultural labourers by gender. The first principal component (S11) for this first group explained 44% of the variability in the S1 group, while adding the second component (S12) increased explained variability to 69%. The second group (S2) consisted of proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. The first principal component (S21) for this second group explained 52% of the variability in the S2 group, while adding the second component (S22) increased explained variability to 90%.

We fitted a generalised linear model using direct estimates of proportions of food insecure households as the response variable and the four principal component scores S11, S12, S21, S22 and few other selected auxiliary variables from the 2011 Population Census as potential covariates. The final selected model included five covariates namely proportional scheduled caste population (SC), literacy rate (Lit), proportion of working population (WP), index for main worker population (S11) and index for marginal worker population (S21), with Akaike Information Criterion (AIC) value of 636.34. For this model, null deviance is 430.88 on 70 degrees of freedom and including the five independent has decreased the deviance to 294.72 on 65 degrees of freedom, a significant reduction in deviance. The residual deviance has reduced by 136.16 with a loss of five degrees of freedom. We use Hosmer Lemeshow goodness of fit test to examine the fitted model (*i.e.* model fits depends on the difference between the model and the observed data). The *p*-value of Hosmer Lemeshow goodness is 0.9987. This indicates that model appears to fit well because we have no significant difference between the model and the observed data (*i.e*. the *p*-value is above 0.05). In this fitted model it can be noted that SC, Lit, WP, S11 influence proportion of food insecure households positively, while S21 has a slightly negative effect. Further, the coefficients of SC (–1.3741), Lit (–1.10334), WP (–5.0617), S11 (–0.385) and S21 (0.3123) are significant (p <0.001). This final model was then used to produce district wise estimates of food insecurity.

3. Small Area Estimation Methodology

Let us assume that a finite population U of size N consists of D non-overlapping and mutually exclusive small areas (or areas), and a sample s of size n is

drawn from this population using a probability sampling method. We use a subscript *d* to index quantities belonging to small area *d*. Let U_d and s_d be the population and sample of sizes N_d and n_d in area *d*, respectively such that $U = \bigcup_{d=1}^{D} U_d$, $N = \sum_{d=1}^{D} N_d$, $s = \bigcup_{d=1}^{D} s_d$ and $n = \sum_{d=1}^{D} n_d$. We use subscript *s* and *r* respectively to denote quantities related to sample and non-sample parts of the population. Let y_{di} denotes the value of the variable of interest for unit i $(i = 1, ..., N_d)$ in area *d*. The variable of interest, with values y_{di}, is binary (*e.g.*, $y_{di} = 1$ if household *i* in small area *d* is food insecure and 0 otherwise), and the aim is to estimate the small area population count, $y_d = \sum_{i \in U_d} y_{di}$, or equivalently the small area proportion, $P_d = N_d^{-1} y_d$, in area *d*. The standard direct estimator (denoted by Direct) for P_d is $\hat{p}_d^{Direct} = \sum_{i \in s_d} \tilde{w}_{di} y_{di}$, where $\tilde{w}_{di} = w_{di} \Big/ \sum_{i \in s_d} w_{di}$ with $\sum_{i \in s_d} \tilde{w}_{di} = 1$ and w_{di} is the survey weight for unit *i* in area *d*. The estimate of variance of direct estimator is $v(\hat{p}_d^{Direct}) \approx \sum_{i \in s_d} \tilde{w}_{di}(\tilde{w}_{di} - 1)(y_{di} - \hat{p}_d^{Direct})^2$. Under simple random sampling (SRS), $\hat{p}_d^{Direct} = n_d^{-1} y_{sd}$, and $v(\hat{p}_d^{Direct}) \approx n_d^{-1} p_d (1 - p_d)$, where $y_{sd} = \sum_{i \in s_d} y_{di}$ denotes the sample count in area *d*. Similarly, $y_{rd} = \sum_{i \in s_r} y_{di}$ denotes the non-sample count in area *d*. If the sampling design is informative, this SRS-based version of Direct may be biased. If we ignore the sampling design, the sample count y_{sd} in area *d* can be assumed to follow a Binomial distribution with parameters n_d and π_d, i.e. $y_{sd} | u_d \sim \text{Bin}(n_d, \pi_d)$. Similarly, for the non-sample count, $y_{rd} | u_d \sim \text{Bin}(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent binomial variables with π_d being a common success probability. This leads to $E(y_{sd} | u_d) = n_d \pi_d$ and $E(y_{rd} | u_d) = (N_d - n_d)\, \pi_d$.

Let $\mathbf{x}_d$ be the *k*-vector of covariates for area *d* from available from secondary data sources. Following Johnson *et al.* (2010), Chandra *et al.* (2011) and Anjoy *et al.* (2020), the model linking the probability π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model (LLMM) of form

$$logit(\pi_d) = \ln\left\{\pi_d (1 - \pi_d)^{-1}\right\} = \eta_d = \mathbf{x}_d^T \boldsymbol{\beta} + u_d, \tag{1}$$

with $\pi_d = \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)\left\{1 + \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)\right\}^{-1} = \text{expit}(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)$. Here $\boldsymbol{\beta}$ is the *k*-vector of regression coefficients and u_d is the area-specific random effect that capture the area dissimilarities. We assume that u_d is independent and normally distributed with mean zero and variance σ_u^2. The total population counts y_d can be written as $y_d = y_{sd} + y_{rd}$, where y_{sd}, the sample count is known whereas y_{rd}, the non-sample count, is unknown. Under (1), a plug-in empirical predictor (EPP) of y_d in area *d* is

$$\hat{y}_d^{EPP} = y_{sd} + (N_d - n_d)\left[\text{expit}(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{u}_d)\right]. \tag{2}$$

An estimate of the corresponding proportion in area d is $\hat{p}_d^{EPP} = N_d^{-1}\hat{y}_d^{EPP}$. It is obvious that in order to compute the small area estimates by equation (2), we require estimates of the unknown parameters $\boldsymbol{\beta}$ and $\mathbf{u} = (u_1, ..., u_D)^T$. We use an iterative procedure that combines the Penalized Quasi-Likelihood estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with REML estimation of σ_u^2 to estimate unknown parameters.

The model (1) is based on unweighted sample counts, and hence it assumes that sampling within areas is non-informative given the values of the contextual variables and the random area effects. The EPP predictor based on (2) therefore ignores the complex survey design used in HCES data. But, the sampling design used in HCES is informative. The precision of an estimate from a complex sample can be higher than for a simple random sample, because of the better use of population data through a representative sample drawn using a suitable sampling design. Following Chandra *et al.* (2019), we model the survey weighted probability estimate for an area as a binomial proportion, with an "effective sample size" that equates the resulting binomial variance to the actual sampling variance of the survey weighted direct estimate for the area. Hence, in our analysis we replaced the "actual sample size" and the "actual sample count" with the "effective sample size" and the "effective sample count" respectively. The mean squared error (MSE) estimation is followed from Chandra *et al.* (2019).

4. Results and Discussions

In this Section we first examine if sampling design in HCES sample data is informative. The sampling design is called informative design if the distribution in the sample is different from the distribution in the population. Such sampling design is also referred as non-ignorable design. The sampling design used in survey data collected must be incorporated in making the valid analytic inference about the population. For this purpose, we compute the effective sample sizes and the effective sample counts for the HCES data. Readers are suggested to refer Chandra *et al.* (2019) for details about calculation of the effective sample sizes and the effective sample counts.

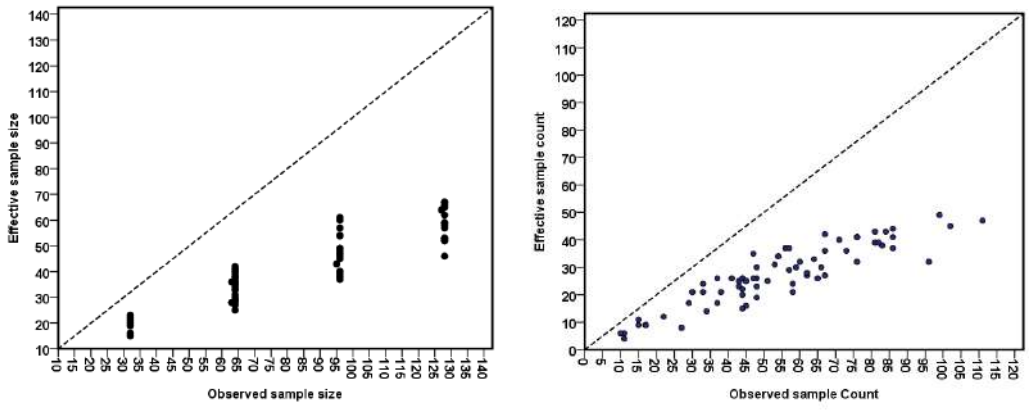


Figure 1: **Effective sample size versus observed sample size (left) and effective sample count versus observed sample count (right) in 2011 HCES data**

Figure 1 plots the effective sample sizes against the observed sample sizes (left side) and the effective sample counts against the observed sample counts (right side). It is evident from Figure 1 that the effective sample size is smaller than the observed sample sizes in almost all the districts. Similarly, the effective sample counts is lower than the observed sample counts. This indicates that the sampling design results in a loss in information, when compared with simple random sampling, in all the districts.

Figure 2 presents the district-wise survey weighted and unweighted direct estimates of proportion of household food insecurity. It can be seen from Figure 2 that the unweighted direct estimates underestimate the proportion of food insecurity, in majority of the districts. These examples are evident that the sampling design is informative and therefore must be accounted in SAE. Following the idea of Korn and Graubard (1998) and Chandra *et al.* (2019), we use the effective sample sizes in replace of observed sample sizes to incorporate the sampling design of HCES data.

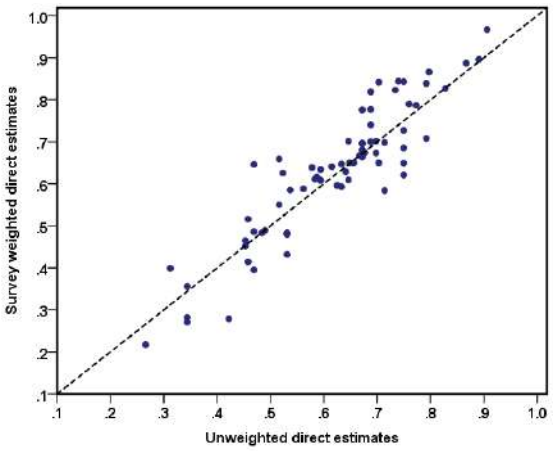


Figure 2: **District-wise survey weighted direct estimates versus unweighted direct estimates of proportion of food insecure households**

The estimates of proportion of food insecurity (or incidence of food insecurity) at district level for rural areas in the state of Uttar Pradesh is generated from the EPP method described in Section 3 using 5 significant covariates described in Section 2. Here we assume a binomial specification for the "effective" district level sample counts of food insecurity. Some important diagnostics measures are now discussed to examine the assumptions of the underlying models, and to validate the empirical performances of the EPP method. Generally, two types of diagnostics measures are advised in SAE applications. These are (i) the model diagnostics, and (ii) the diagnostics for the small area estimates. See Brown *et al.* (2001). The model diagnostics are applied to verify model assumptions. The other diagnostics are used to validate reliability of the model-based small area estimates of incidence of food insecurity generated by the EPP method. In LLMM (1) the random specific effects are assumed to have a normal distribution with mean zero and fixed variance. If the model assumptions are satisfied then the district level residuals are expected to be randomly distributed around zero. Histogram and normal probability (q-q) plot can be used to examine the normality assumption. Figure 3 shows the histogram (left plot), the normal probability (q-q) plot (centre plot) and the distribution of the district-level residuals (right plot). We also use the Shapiro-Wilk test (implemented using the *shapiro.test()* function in R) to examine the normality of the district random effects.

The Shapiro-Wilk test with p-value lower than 0.05 indicate that the data deviate from normality. Here, the value of Shapiro-Wilk test statistics is 0.988 with 71 degree of freedom and p-value 0.746. In Figure 3, the district level residuals appear to be randomly distributed around zero. Further, histogram and the q-q plot also provide evidence in support of the normality assumption. The Shapiro-Wilk p-value is larger than 0.05 and hence, the district random effects are likely to be normally distributed.

Following Chandra *et al.* (2011) and Brown *et al.* (2001), we use three commonly used measures for assessing the validity and the reliability of the model-based estimates generated by the EPP: the bias diagnostic, the percent coefficient of variation (CV) diagnostic and the 95 percent confidence interval diagnostic. The first diagnostics assesses the validity and last two assess the improved precision of the model based small area estimates. We also implemented a calibration diagnostic where the EPP estimates are aggregated to higher level and compared with direct estimates at this level. The bias diagnostic is based on following idea. The direct estimates are unbiased estimates of the population values of interest (i.e. true values), their regression on the true values should be linear and correspond to the identity line. If model-based small area estimates are close to these true values the regression of the direct estimates on these model-based estimates should be similar. We therefore plot direct estimates (y-axis) vs. model-based small area estimates (x-axis) and we looked for divergence of the fitted least squares regression line from the line of equality.

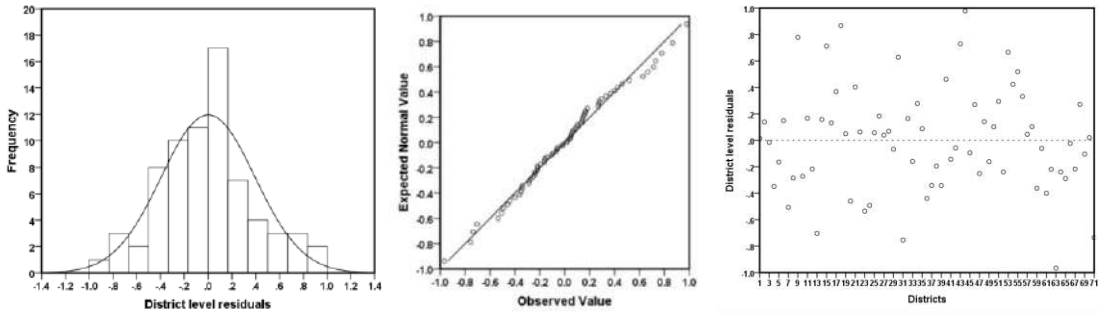


Figure 3: Histograms (left plot), normal q-q plots (centre plot) and distributions of the district-level residuals (right plot)

Figure 4 provides a bias diagnostic plot, defined by plotting direct estimates (*Y* axis) against corresponding small area estimates generated by the EPP (*X*-axis) and testing for divergence of the fitted least squares regression line (dashed line) from the line of equality, i.e. $Y = X$ line (solid line). The bias diagnostic plot in Figure 4 clearly indicate that the EPP estimates are less extreme when compared to the direct estimates, demonstrating the typical SAE outcome of shrinking more extreme values towards the average. The value of R^2 for the fitted regression line between the direct estimates and the EPP estimates is 95.6 per cent. The bias diagnostics indicates that the estimates generated by the EPP appear to be consistent with the direct estimates.

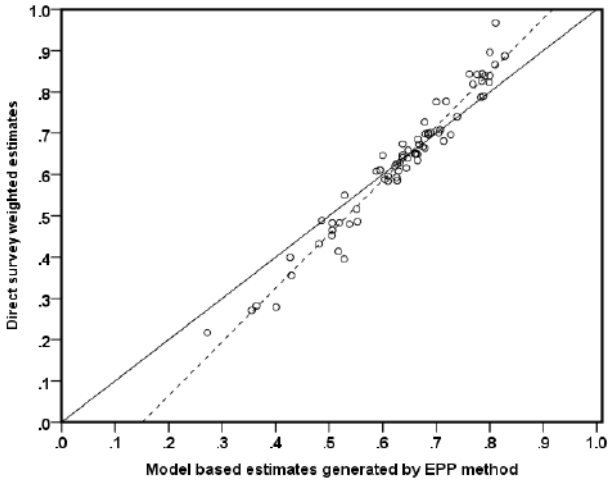


Figure 4: Bias diagnostic plot with $y = x$ line (solid) and regression line (dotted) for proportion of food insecurity for rural areas in Uttar Pradesh: EPP estimates versus direct survey estimates.

We now illustrate the second set of diagnostics to assess the extent to which the EPP estimates improve in precision compared to the direct estimates. The percent coefficient of variation (CV) is the estimated sampling standard error as a percentage of the estimate. Small area estimates with large CVs are considered unreliable. Table 2 provides a summary of CVs of the direct estimates and the EPP estimates. Figure 5 presents the District-wise values of CV for the direct and EPP methods. In one of the 71 districts, smaller CV (2.16%) of direct estimate is due to extreme value of proportion. Sample size and sample count for this district are 64 and 58 respectively while and direct estimate of proportion of food insecurity is 0.967. Note that the effective sample size and effective sample count for this districts are 25 and 24 respectively. In Table 2, we therefore presented the summary based on 70 districts (excluding one district extreme value of proportion). In further discussion we refer summary based on 70 districts only. The CVs of the direct estimates are larger than the EPP estimates.

Table 2: **Summary of area distributions of percentage coefficients of variation (CV, %) for the direct and EPP methods applied to HCES data**

Values	Summary of 71 Districts		Summary of 70 Districts	
	Direct	EPP	Direct	EPP
Minimum	2.16	5.12	5.53	5.12
Q1	8.97	7.90	9.06	7.99
Mean	14.41	10.60	14.59	10.65
Median	12.31	9.56	12.38	9.56
Q3	12.31	9.56	12.38	9.56
Maximum	45.52	24.29	45.52	24.29

Table 2 and Figure 5 show that direct estimates of incidence food insecurity are unstable with CVs that vary from 5.53 to 45.52 % with average of 14.59 %. In contrast, the CV values of EPP range from 5.12 to 24.29% with average of 10.65%. The relative performance of the EPP as compared to the direct survey estimates improve with decreasing district specific observed sample sizes. The estimates

computed from the EPP are more reliable and provide a better indication of food insecurity incidence. The district-wise plot of the 95 % confidence intervals (CIs) generated by direct and EPP methods are displayed in Figure 6, which shows that the 95% CIs for the direct estimates are wider than the 95% CIs for the EPP.

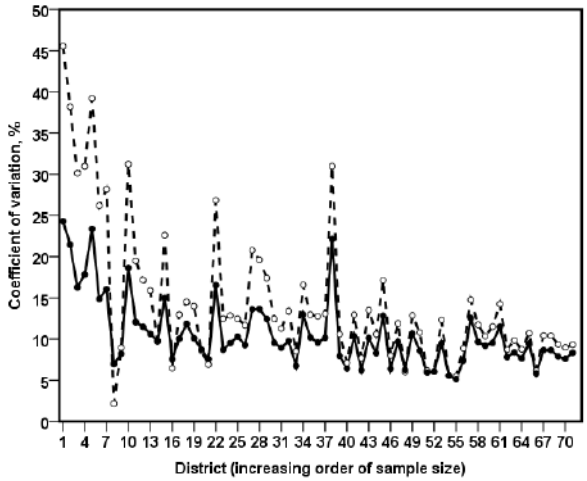


Figure 5: **District-wise percentage coefficient of variation (CV, %) for the direct (dotted line, o) and EPP (solid line, •) estimates for the food insecurity prevalence in Uttar Pradesh**

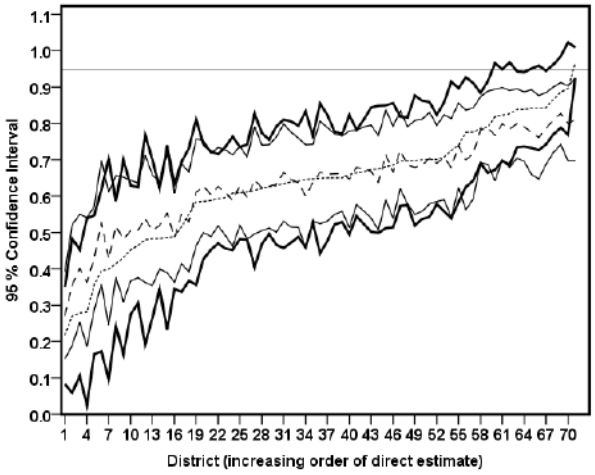


Figure 6: **District-wise 95 percentage nominal confidence interval (95% CI) for the direct (solid line) and EPP (thin line) methods. Direct (dotted point) and EPP estimates (dash point) for the food insecurity prevalence in Uttar Pradesh are shown in the 95% CI**

We inspect the aggregation property of the model-based district-level estimates generated by EPP at higher (*e.g.* State or Region) level. Let $\hat{P}_d$ and N_d denote the estimate of proportion of household food insecurity and population size for district *d.* The state-level estimate of the proportion of food insecure households is calculated as $\hat{P} = \sum_{d=1}^{D} N_d \hat{P}_d \Big/ \sum_{d=1}^{D} N_d$. The state of Uttar Pradesh is divided into Central, Eastern, Western and Southern regions, and calibration properties has been examined for these regions. State and regional level estimates of the proportion of food insecurity generated by the EPP is reported in Table 3. Comparing these with the corresponding direct estimates we see that the EPP estimates are very close to the direct estimates at

state level as well in each of the four regions. In Figure 7 we present a map showing the estimates of proportion of food insecurity in different districts in rural areas of Uttar Pradesh produced by the EPP method. This map provides the district-wise degree of inequality with respect to distribution of extent of food insecurity in rural areas of Uttar Pradesh. This map is supplemented by the results set out in Table 4, where we report the district-wise estimates along with CVs and 95 % confidence intervals generated by direct and EPP. The results indicate an east-west divide in the distribution of food insecurity. For example, in the western part of Uttar Pradesh there are many districts with low level of incidence of food insecurity. Similarly, in the eastern part and in the Bundelkhand region (north-east) we see districts with high incidence of food insecurity. This should prove useful for policy planners and administrators aiming to take effective financial and administrative decisions.

Table 3: Aggregated level estimates of incidence of food insecurity generated by direct and EPP method in different regions in Uttar Pradesh.

Estimator	State	Central	Eastern	Southern	Western
Direct	0.644	0.557	0.698	0.431	0.649
EPP	0.646	0.565	0.695	0.455	0.650

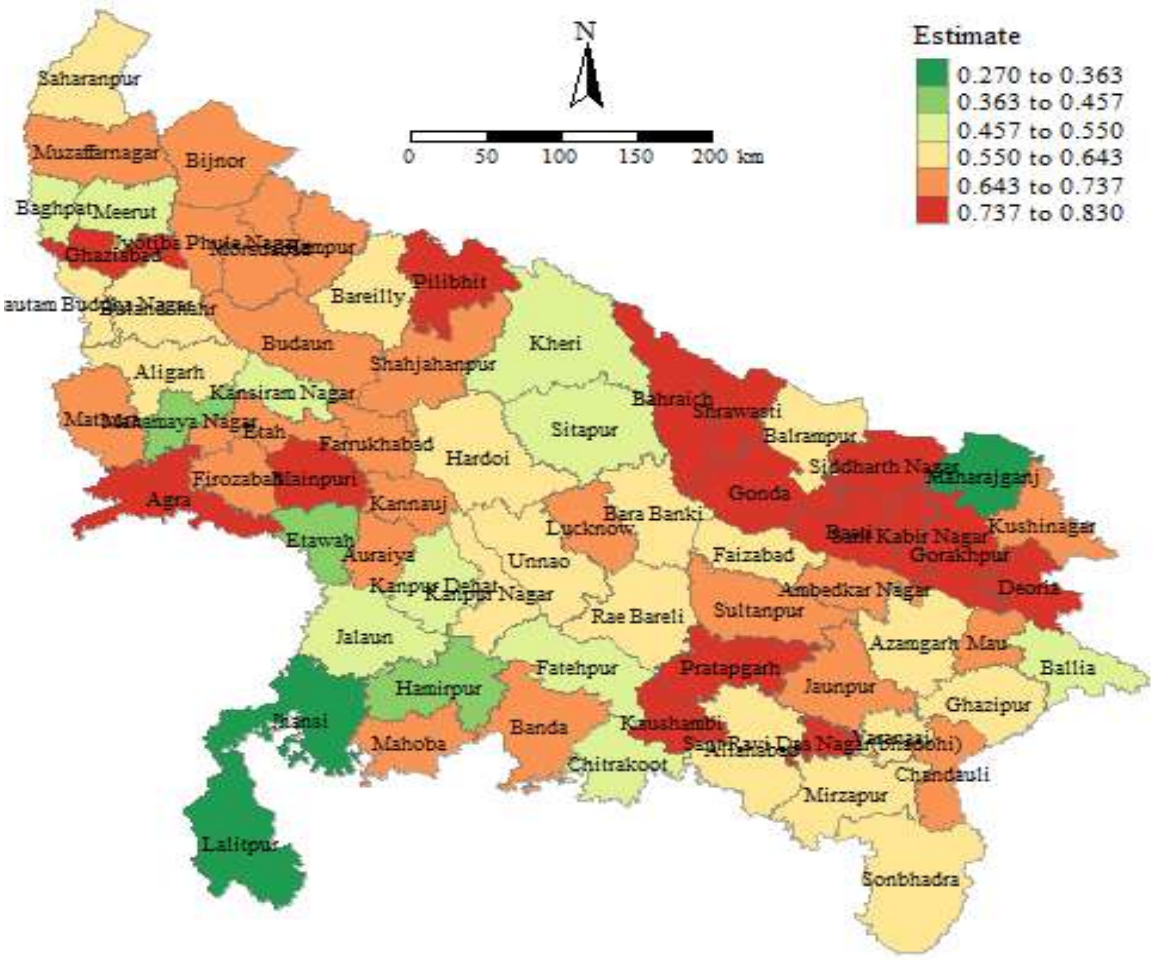


Figure 7: **EPP estimates showing the spatial distribution of incidence of food insecurity by District in Uttar Pradesh**

Table 4: Direct and EPP estimates along with 95 % confidence interval (95% CI) and percentage coefficient of variation (CV) of the incidence of food insecurity by District in rural areas of Uttar Pradesh

	Direct				EPP			
		95 % CI				95 % CI		
District	Estimate	Lower	Upper	CV	Estimate	Lower	Upper	CV
Saharanpur	0.641	0.488	0.794	11.90	0.637	0.514	0.760	9.65
Muzaffarnagar	0.696	0.575	0.817	8.72	0.685	0.578	0.792	7.79
Bijnor	0.667	0.523	0.811	10.81	0.675	0.559	0.791	8.59
Moradabad	0.664	0.545	0.783	8.98	0.679	0.576	0.782	7.58
Rampur	0.681	0.512	0.850	12.44	0.714	0.590	0.838	8.69
Jyotiba Phule Nr	0.700	0.537	0.863	11.67	0.685	0.558	0.812	9.26
Meerut	0.452	0.275	0.629	19.62	0.505	0.367	0.643	13.63
Baghpat	0.480	0.191	0.769	30.12	0.538	0.363	0.713	16.29
Ghaziabad	0.843	0.727	0.959	6.90	0.762	0.647	0.877	7.55
Gautam B. Nr	0.486	0.232	0.740	26.17	0.553	0.389	0.717	14.87
Bulandshahr	0.611	0.480	0.742	10.70	0.595	0.482	0.708	9.46
Aligarh	0.516	0.337	0.695	17.37	0.551	0.414	0.688	12.42
Hathras	0.356	0.165	0.547	26.82	0.429	0.287	0.571	16.53
Mathura	0.685	0.514	0.856	12.45	0.666	0.539	0.793	9.56
Agra	0.844	0.743	0.945	5.97	0.786	0.687	0.885	6.28
Firozabad	0.698	0.519	0.877	12.83	0.679	0.550	0.808	9.51
Etah	0.777	0.643	0.911	8.59	0.718	0.593	0.843	8.73
Mainpuri	0.967	0.925	1.009	2.16	0.811	0.698	0.924	6.96
Budaun	0.701	0.543	0.859	11.29	0.705	0.579	0.831	8.94
Bareilly	0.585	0.427	0.743	13.50	0.627	0.500	0.754	10.12
Pilibhit	0.842	0.733	0.951	6.46	0.776	0.659	0.893	7.54
Shahjahanpur	0.673	0.502	0.844	12.72	0.668	0.540	0.796	9.56
Kheri	0.465	0.306	0.624	17.12	0.506	0.376	0.636	12.82
Sitapur	0.483	0.345	0.621	14.25	0.519	0.400	0.638	11.50
Hardoi	0.626	0.496	0.756	10.35	0.627	0.512	0.742	9.18
Unnao	0.608	0.452	0.764	12.85	0.588	0.462	0.714	10.68
Lucknow	0.639	0.472	0.806	13.04	0.646	0.515	0.777	10.16
Rae Bareli	0.647	0.527	0.767	9.30	0.637	0.531	0.743	8.32
Farrukhabad	0.649	0.443	0.855	15.89	0.665	0.524	0.806	10.63
Kannauj	0.776	0.625	0.927	9.71	0.700	0.563	0.837	9.76
Etawah	0.279	0.105	0.453	31.22	0.401	0.252	0.550	18.63
Auraiya	0.659	0.495	0.823	12.45	0.647	0.514	0.780	10.31
Kanpur Dehat	0.483	0.265	0.701	22.60	0.506	0.354	0.658	14.99
Kanpur Nagar	0.646	0.459	0.833	14.51	0.600	0.458	0.742	11.83
Jalaun	0.550	0.368	0.732	16.57	0.529	0.392	0.666	12.96
Jhansi	0.217	0.083	0.351	30.96	0.272	0.152	0.392	22.12
Lalitpur	0.271	0.059	0.483	39.19	0.355	0.189	0.521	23.35
Hamirpur	0.399	0.095	0.703	38.16	0.427	0.244	0.610	21.44
Mahoba	0.282	0.025	0.539	45.52	0.364	0.187	0.541	24.29
Banda	0.727	0.539	0.915	12.96	0.678	0.542	0.814	10.04
Chitrakoot	0.432	0.165	0.699	30.95	0.481	0.310	0.652	17.82
Fatehpur	0.489	0.345	0.633	14.76	0.486	0.364	0.608	12.52
Pratapgarh	0.887	0.789	0.985	5.53	0.828	0.743	0.913	5.12
Kaushambi	0.896	0.769	1.023	7.10	0.800	0.697	0.903	6.43

Allahabad	0.621	0.468	0.774	12.31	0.623	0.505	0.741	9.48
BaraBanki	0.674	0.499	0.849	12.95	0.637	0.507	0.767	10.20
Faizabad	0.584	0.356	0.812	19.49	0.610	0.463	0.757	12.05
Ambedkar Nr	0.701	0.578	0.824	8.74	0.690	0.585	0.795	7.63
Sultanpur	0.650	0.529	0.771	9.35	0.662	0.558	0.766	7.88
Bahraich	0.740	0.583	0.897	10.60	0.739	0.622	0.856	7.93
Shrawasti	0.819	0.672	0.966	8.96	0.768	0.642	0.894	8.19
Balrampur	0.616	0.405	0.827	17.14	0.644	0.496	0.792	11.49
Gonda	0.866	0.770	0.962	5.56	0.810	0.720	0.900	5.56
Siddharthnagar	0.839	0.735	0.943	6.17	0.800	0.705	0.895	5.93
Basti	0.839	0.737	0.941	6.05	0.791	0.695	0.887	6.05
Sant Kabir Nr	0.823	0.697	0.949	7.68	0.799	0.699	0.899	6.23
Mahrajganj	0.708	0.558	0.858	10.60	0.707	0.590	0.824	8.27
Gorakhpur	0.787	0.690	0.884	6.19	0.783	0.692	0.874	5.81
Kushinagar	0.696	0.573	0.819	8.87	0.727	0.620	0.834	7.37
Deoria	0.790	0.664	0.916	7.98	0.788	0.687	0.889	6.40
Azamgarh	0.593	0.470	0.716	10.41	0.626	0.518	0.734	8.65
Mau	0.634	0.457	0.811	13.98	0.665	0.531	0.799	10.10
Ballia	0.414	0.242	0.586	20.77	0.517	0.377	0.657	13.58
Jaunpur	0.650	0.522	0.778	9.84	0.661	0.550	0.772	8.37
Ghazipur	0.609	0.482	0.736	10.43	0.629	0.520	0.738	8.63
Chandauli	0.650	0.476	0.824	13.40	0.660	0.531	0.789	9.75
Varanasi	0.596	0.456	0.736	11.71	0.610	0.492	0.728	9.66
Bhadohi	0.827	0.686	0.968	8.50	0.785	0.679	0.891	6.76
Mirzapur	0.588	0.453	0.723	11.50	0.604	0.489	0.719	9.54
Sonbhadra	0.629	0.466	0.792	12.94	0.632	0.501	0.763	10.36
Kanshiram Nr	0.395	0.173	0.617	28.14	0.528	0.358	0.698	16.06

Nr- Nagar

5. Concluding Remarks

In this paper we outlined a plug-in empirical predictor (EPP) for small area proportions and employed for estimating the district-wise incidence of food insecurity in rural areas of the state of Uttar Pradesh using the 2011-12 HCES data collected by the NSSO of India. The auxiliary variables used in this analysis were taken from the 2011 Population Census. The effective sample sizes in place of the observed sample sizes were used to account for sampling design information of the 2011-12 HCES. The use of survey information through effective sample size leads to better representative and realistic estimates of incidence of food insecurity. The empirical results were also evaluated through several diagnostic measures and showed that the model-based SAE method defined by EPP provide significant gains in efficiency for generating district level estimates of proportion of food insecurity. Spatial map produced from the estimates generated by the EPP provides an evidence of inequality in distribution of incidence food insecurity across different districts in Uttar Pradesh. Availability of reliable district level estimates can definitely be useful for various Departments and Ministries in Government of India as well as International organizations for their policy research and strategic planning. These estimates will also be useful for budget allocation and to target welfare interventions by identifying the districts/regions with high food insecurity incidence. This application clearly demonstrates the advantage of using SAE technique to cope up the small sample size

problem in producing the cost effective and reliable disaggregate level estimates and confidence intervals from existing survey data by combining auxiliary information from different published sources with direct survey estimates.

References

Anjoy, P., Chandra, H. and Parsad, R. (2020). Estimation and spatial mapping of incidence of indebtedness in the state of Karnataka in India by combining survey and census data. *Statistics and Applications*, **18(1)**, 21-33.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods - an application to the unemployment estimates from the UK LFS. In the *Proceedings of the Statistics Canada Symposium. Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. *Statistics Canada*.

Chandra, H. (2013). Exploring spatial dependence in area level random effect model for disaggregate level crop yield estimation. *Journal of Applied Statistics*, **40**, 823-842.

Chandra, H., Chambers, R. and Salvati, N. (2019). Small area estimation of survey weighted counts under aggregated level spatial model. *Survey Methodology*, **45**(**1**), 31-59.

Chandra, H., Chambers, R. and Salvati, N. (2012). Small area estimation of proportions in business surveys. *Journal of Statistical Computation and Simulation*. **82**(**6**), 783-795.

Chandra, H., Salvati, N. and Chambers, R. (2018). Small area estimation under a spatially non-linear model. *Computational Statistics and Data Analysis*, **126**, 19–38.

Chandra, H., Salvati, N. and Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics,* **20,** 30–56.

Chandra, H., Salvati, N. and Chambers, R. (2015). A spatially nonstationary Fay-Herriot model for small area estimation. *Journal of Survey Statistics and Methodology*, **3**, 109–135.

Chandra, H., Salvati N. and Sud U. C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India-an application of small area estimation technique. *Journal of Applied Statistics*, **38**(**11**), 2413-2432.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Johnson F. A., Chandra H., Brown J. and Padmadas S. (2010). Estimating district-level births attended by skilled attendants in Ghana using demographic health survey and census data: an application of small area estimation technique. *Journal of Official Statistics,* **26** (**2**), 341–359.

Korn, E. and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, **23**, 192-201.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning Inference*, **111**, 117–127.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. 2nd Edition. John Wiley and Sons.

Special Proceeding of the 22 nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02–04, 2020; pp 39-51

Linear Model Perspectives of fMRI Studies

Bikas Kumar Sinha
Former Professor, Indian Statistical Institute, Kolkata, India

Received: 29 April 2020; Revised: 09 May 2020; Accepted: 19 May 2020

Abstract

Functional Magnetic Resonance Imaging (fMRI) is a technology for studying how our brains respond to mental stimuli. It is interesting to note the potential developments of linear models in the study of 'design sequences' employed for fMRI studies. At the design stage, one is interested in developing a sequence of mental stimuli for collecting data in order to render information about some 'unknown' yet 'meaningful' parameters under an assumed statistical model. The simplest such model incorporates linear relation between 'mean response' and the 'parameters' describing the effects of the stimuli, applied at regularly spaced time points during the study period. In this paper, we introduce the linear model and discuss estimation issues. In the process, we take up a study of relative performances of comparable design sequences.

Key words: fMRI, Linear model; h-Parameters; Estimability; Information matrix; Generalized variance; Average variance; Design issues.

1. Introduction

It is interesting to note that **Statistics and Applications** published, in as early as 2008, an article dealing with **"event- related functional magnetic resonance imaging . . . "**. In fMRI studies, the brain functions of the experimental subjects are captured through response profiles at a number of instances. Each subject experiences onset of a stimulus at an instant if the stimulus is 'active' [denoted by code '1'] at that instant; otherwise, the subject is at 'resting state' [denoted by code '0'] at that instant. Each instant is defined as a compact duration of '4 seconds'. At any instant, the brain voxel captures the cumulative effects of a fixed (but unknown) parameter θ and other model parameters, known as h-parameters at the current instant as well as at each of the immediate past ordered $(K-1)$ instants - for some K - whenever there has been an onset of active stimulus at any of these instances. The reader familiar with the concept of 'carry-over effects' in the context of Repeated Measurement Designs [RMDs] or Cross-Over Designs will find a similarity in the model description. [Vide Shah and Sinha (1989)]. We will also mention about 'circular models' and for that we refer to Kunert (1984).

An anonymous referee has aptly pointed out another related piece of work by Maus *et al.* (2010).

This talk was delivered in C. R. Rao Birth Centenary Celebration session on January 02, 2020.
Correponding Author: Bikas K. Sinha
Email:bikassinha1946@gmail.com

A more general scenario exhibits itself in terms of different stages of activation of the brain stimuli, rather than just being 'active' - as coded by '1' in the above. We refer to Kao *et al.* (2008) for this and related considerations.

Below we introduce the linear (mean) model as has been described in the literature. There is a 'design sequence' in the form of a collection of 1s and 0s of, say length n. We denote it by D_n. As for example, for $n = 8$, the following describes an 8-point design: $D_8 = [0, 1, 1, 0, 1, 1, 0, 1]$. The implementation of the suggested design D_8 is described below. For any n, D_n is very much like D_8. The linear model to be described below is developed as a 'circular' model - a well-known consideration in the context of RMDs or Cross-Over Designs. Vide Kunert (1984) or Shah and Sinha (1989). To visualize a circular model, the same sequence (describing D_8) is used as a 'dummy' sequence and this is described as follows:

$< 0, 1, 1, 0, 1, 1, 0, 1 >$ $\rightarrow$ $[0, 1, 1, 0, 1, 1, 0, 1]$
Dummy Sequence followed by Data-generating Sequence

There are 8 data/time points and as such we observe y_1 to y_8 corresponding to the 8 time points in the data-generating sequence $[0, 1, 1, 0, 1, 1, 0, 1]$ - going from left to right. In the terminology of RMDs or Cross-Over Designs, for the first time point, the 'direct effect' [denoted by h_1] is to be captured along with the 'carry-over effects' $[h_2, h_3, ...]$ of the preceding time points as described in the Dummy Sequence - from right to left. Althrough, at each data point, only if the stimulus is active [denoted by 1], the corresponding h-parameter will be present in the mean model. Moreover, for n data/time points, we can incorporate at the most n 'parameters'- including the fixed parameter θ. This implies that we can incorporate in the model at the most $(n - 1)$ h-parameters. Otherwise/estimability issues creep in. In terms of K, it means that we assume - to start with - that $K \leq (n - 1)$.

We start with the following Table 1 describing the linear (mean) model underlying the design D_8. We assume $K = 7$. For clarity, we explain the derivation of the mean model for y_1. The co-efficients to be attached to the regression parameters i.e., h-parameters [h_1 to h_7] in the expression for the mean model corresponding to y_1 are: $(0, 1, 0, 1, 1, 0)$. This is seen as follows. In the data-generating sequence, extreme left-hand coefficient (0) is attached to h_1; then the coefficients in the dummy sequence are taken successively from right to left for attachment to h_2 to h_7. There are 6 h-parameters (in addition to h_1), and hence 6 of the coefficients are selected in the order from right to left in the dummy sequence. That gives the coefficients for h_2 to h_7 in the order $(1, 0, 1, 1, 0, 1)$. Hence the mean model for y_1 is given by $\theta + h_2 + h_4 + h_5 + h_7$. Likewise, for y_2, the coefficients start from the second member from the left of the data-generating sequence and proceeds along the left direction, cutting across the dummy-sequence and covers a total of 7 coefficients. The coefficients are thus $(1, 0, 1, 0, 1, 1, 0)$. All these are displayed in Table 1. Note that in Table 1, the h-parameters are listed in the reverse order.

Remark 1: It may be noted that the linear mean model developed above has similarity with one in the set-up of 'biased spring balance weighing designs'. Vide Raghavarao (1971) or Shah and Sinha (1989). It follows that θ-parameter represents the bias component in spring balance weighing design context. The co-efficient matrix $\mathbf{X} = ((x_{ij}))$ consists of 0s and 1s. However, the $\mathbf{X}$- matrix is shown in the reverse order. Multiplication by a permutation

Table 1: Linear Model with positional carry-over effects in terms of h-parameters

S1. No.	h_7	h_6	h_5	h_4	h_3	h_2	h_1	y	Mean Model
1	1	0	1	1	0	1	0	y_1	$\theta + h_2 + h_4 + h_5 + h_7$
2	0	1	1	0	1	0	1	y_2	$\theta + h_1 + h_3 + h_5 + h_6$
3	1	1	0	1	0	1	1	y_3	$\theta + h_1 + h_2 + h_4 + h_6 + h_7$
4	1	0	1	0	1	1	0	y_4	$\theta + h_2 + h_3 + h_5 + h_7$
5	0	1	0	1	1	0	1	y_5	$\theta + h_1 + h_3 + h_4 + h_6$
6	1	0	1	1	0	1	1	y_6	$\theta + h_1 + h_2 + h_4 + h_5 + h_7$
7	0	1	1	0	1	1	0	y_7	$\theta + h_2 + h_3 + h_5 + h_6$
8	1	1	0	1	1	0	1	y_8	$\theta + h_1 + h_3 + h_4 + h_6 + h_7$

matrix $\mathbf{P}$ will bring it to the right/standard order. Finally, the linear model $(\mathbf{Y}, \mathbf{X}^{(*)}\beta, \sigma^{\mathbf{2}}\mathbf{I})$ is obtained as usual where $\mathbf{X}^{(*)} = (\mathbf{1}, \mathbf{PX})$ and β=(θ, h_1, h_2, ...)'. It is assumed that the errors are, as usual, uncorrelated with zero means and equal variances.

Remark 2: We must note that a 'circular model' has been explicitly used in Table 1. The dummy - sequence is derived from the data - generating sequence on which the circular model is built. Another implication is that the columns $\mathbf{h}_1, \mathbf{h}_2, ...$ are circular in nature. That is, the columns of the matrix $\mathbf{X}$ are circular in nature. For a non- circular design/model, the carry - over effects are dependent on the nature of 1s and 0s - for each incoming unit/patient-at the two ends of the design sequence.

At times, the number of h-parameters may be specified and it may happen that there are $K^*[< K]$ h-parameters in the model. In that case, the understanding is that the initial set of K^* h-parameters viz. $h_1, h_2, \ldots, h_{K^*}$ are important and the rest can be ignored from the mean model. For $K^* = 4$, the model expectations of successive responses corresponding to the above design would be:

$$\theta + h_2 + h_4, \theta + h_1 + h_3, \theta + h_1 + h_2 + h_4, \theta + h_2 + h_3, \theta + h_1 + h_3 + h_4,$$

$$\theta + h_1 + h_2 + h_4, \theta + h_2 + h_3, \theta + h_1 + h_3 + h_4.$$

Note that the above design with $n = 8$ instances [for experimentation] generates more number of observations when only $K^* = 4$ h-parameters are assumed to be present. In such a situation, we might curtail the experiment from D_8 to D_5 since there are 5 parameters, including the common/fixed parameter θ. Use of $D_5 : [0, 1, 1, 0, 1]$ provides for the mean model the expressions:

$$\theta + h_2 + h_4, \theta + h_1 + h_3, \theta + h_1 + h_2 + h_4, \theta + h_2 + h_3, \theta + h_1 + h_3 + h_4.$$

On the other hand, use of $D_{alt.5} : [1, 1, 0, 1, 1]$ provides for the mean model the expressions:

$$\theta + h_1 + h_2 + h_3, \theta + h_1 + h_2 + h_3 + h_4, \theta + h_2 + h_3 + h_4, \theta + h_1 + h_3 + h_4, \theta + h_1 + h_2 + h_4.$$

Note that in both the cases, we have taken due consideration of circular nature of the sequence in working out the mean models. A natural question would be to search out

the difference, if any, between the two D_5 designs. Popular optimality criteria rest on the computation of the 'information matrix' for the h-parameters - based on the Gauss-Markov Model, assuming homoscedastic errors with mean 0 and variance σ^2. Minimization of the generalized variance [computed as reciprocal of the determinant of the information matrix] is an acceptable criterion for choice of the best design. This is the so-called D-optimality Criterion [Vide Shah and Sinha (1989)]. We will take up this comparative study in the next section.

2. Linear Estimation of Model Parameters

Since the linear model involves a fixed parameter (θ), for a given number of observations n, we can incorporate a maximal set of $(n-1)$ h-parameters. That is, we can develop the full model with θ and additional $(n-1)$ h-parameters. Naturally, the response vector $\mathbf{Y}$ of dimension $n \times 1$ will come under the standard Gauss-Markov Linear Model mentioned earlier. However, estimability of the h-parameters or of θ are not necessarily guaranteed for all choices of the design sequence.

We have already introduced the 'design matrix' $\mathbf{X}^{(*)} = (\mathbf{1}, \mathbf{PX})$ and the underlying parameters β=(θ, h_1, h_2, ...)'. For a given design D_n, when there are $K[\leq (n-1)]$ h-parameters viz., h_1, h_2, ..., h_K in the model, the h-parameters are all estimable iff Rank $(\mathbf{X}^{(*)}) = 1+K$ where $\mathbf{X}^{(*)}$ is based on K column vectors corresponding to the K h-parameters, in addition to the column vector $\mathbf{1}$. The 'if' part is easy to see. On the other hand, if all the h-parameters are estimable, θ is trivially so based on any single observation and hence the rank condition is satisfied.

In the above example, for $K = 7$, it can be seen that the design D_8 ensures estimability of all the model parameters. Explicit expressions for the estimates of h-parameters are shown below. For θ, expression for its estimator follows readily.

$$h_1 : y_6 - y_1;\ h_2 : -y_1 - y_2 + y_6 + y_7;\ h_3 : -y_1 - y_2 - y_3 + y_6 + y_7 + y_8;$$

$$h_4 : -y_2 - y_3 - y_4 + y_6 + y_7 + y_8;\ h_5 : -y_3 - y_4 - y_5 + y_6 + y_7 + y_8;\ h_6 : -y_4 - y_5 + y_7 + y_8; h_7 : -y_5 + y_8.$$

This suggests that $\mathbf{X}^{(*)}$ is a full rank square matrix of order 8. Hence, all its column vectors are linearly independent. Therefore, for all values of K^*, the number of non-negligible h-parameters, the above design sequence D_8 provides unbiased estimates for each one of them. This holds for all $1 \leq K^* \leq K = 7$.

At this stage, we may as well resolve two more cases. For $K^* = 4$, we may check the acceptability of the two D_5 design sequences listed above: $D_5 : [0, 1, 1, 0, 1]$ and $D_{alt.5} : [1, 1, 0, 1, 1]$. It turns out that both are acceptable from estimability point of view. It would be interesting to make a comparison of their performances with respect to, say, D-optimality criterion. Necessary computations are shown below.

$$I(\beta) = [(5, 3, 3, 3, 3), (3, 3, 1, 2, 2), (3, 1, 3, 1, 2), (3, 2, 1, 3, 1), (3, 2, 2, 1, 3)].$$

$$I(h) = [(6, -4, 1, 1), (-4, 6, -4, 1), (1, -4, 6, -4), (1, 1, -4, 6)], \quad Det(I) = 125.$$

$$I_{alt}(\beta) = [(5,4,4,4,4),(4,4,3,3,3),(4,3,4,3,3),(4,3,3,4,3),(4,3,3,3,4)].$$

$$I_{alt}(h) = [(4,-1,-1,-1),(-1,4,-1,-1),(-1,-1,4,-1),(-1,-1,-1,4)], \quad Det(I_{alt}) = 125.$$

It thus turns out that the two design sequences provide identical generalized variance of the estimates of h-parameters. We will return to this comparison later again in Remark 3.

3. Choice of D_n for given n and K^*

In the context of fMRI study, assume that it is a priori known that, for some K^*, h_1,h_2, ..., h_{K^*} are the only h-parameters present in the mean model. Therefore, we need $n \geq (1 + K^*)$ design points and the choice of D_n must be such that the formation of $\mathbf{X}$ enables one to ensure rank condition. For a chosen n, it is obvious that there are a large number of design sequences of length n-comprising of 1s and 0s. This count is 2^n. It is easy to note that the two extreme sequences $(1, 1, \ldots, 1)$ and $(0, 0, \ldots, 0)$ are inadmissible. In other words, no patient can be in resting phase or in active phase althrough the time duration of the experiment for collection of data. Generally, a mixture of the two phases is called for.

Below we examine the status of a special "Design Sequence [DS]" of length n. Consider the design sequence D_n: [1, 1, 0,..., 0, 0] which gives rise to [1, 1, 0, ..., 0, 0] dummy sequence followed by [1, 1, 0, ..., 0, 0] data-gathering sequence.

Therefore, model expectations of the resulting responses ys are given by: $[\theta + h_1, \theta + h_1 + h_2, \theta + h_2 + h_3, \ldots, \theta + h_{(n-2)} + h_{(n-1)}, \theta + h_{(n-1)}]$, assuming that there are $(n - 1)$ h-parameters in the model. It is interesting to note the following:

(i) For $n = 4,\ K = 3$, the joint information matrix is singular. (ii) For $n = 5,\ K = 4$, the joint information matrix is non - singular. (iii) For $n = 6,\ K = 5$, the joint information matrix is singular. (iv) For $n = 7,\ K = 6$, the joint information matrix is again non - singular

It turns out that for all n $(even) \geq 4,\ K = (n - 1)$, the joint information matrix is singular while for all n $(odd) \geq 5,\ K = (n - 1)$, the joint information matrix is non-singular. Let us fix $n = 8,\ K = 7$ so that DS_8 = [1,1, 0, 0, 0, 0, 0, 0] is not admissible. What if we replace the extreme right- end code 0 by 1? We are asking about the status of $DS_8^* = [1, 1, 0, 0, 0, 0, 0, 1]$. It follows that the 8×8 joint information matrix is given by

$$[(8,3,3,3,3,3,3,3);\ (3,3,2,1,0,0,0,1);\ (3,1,3,2,1,0,0,0);\ \ldots,\ (3,1,0,0,0,1,2,3)].$$

and it is of full rank.

Therefore, it pays off to change exactly one code in the above.

For n odd, each member of the above series of design sequences provides estimates of all the relevant h-parameters. For $n = 7, K = 6$, it follows that

$$V(\hat{h}_2) = 2\sigma^2,\ V(\hat{h}_4) = 4\sigma^2,\ V(\hat{h}_6) = 6\sigma^2,$$

while

$$V(\hat{h}_1) = 6\sigma^2,\ V(\hat{h}_3) = 4\sigma^2,\ V(\hat{h}_5) = 2\sigma^2.$$

Again, for $n = 9,\ K = 8$, we obtain

$$V(\hat{h}_2) = 2\sigma^2,\ V(\hat{h}_4) = 4\sigma^2,\ V(\hat{h}_6) = 6\sigma^2,\ V(\hat{h}_8) = 8\sigma^2,$$

while

$$V(\hat{h}_1) = 8\sigma^2,\ V(\hat{h}_3) = 6\sigma^2,\ V(\hat{h}_5) = 4\sigma^2,\ V(\hat{h}_7) = 2\sigma^2.$$

These expressions suggest general form of the variances of estimates of the h-parameters. For specified (n, K), we can also work out the variance-covariance matrix of the estimates of the h-parameters. For the choice $n = 7, K = 6$, we derive the form of the variance-covariance matrix as given below.

$$[(6, 1, 4, 3, 2, 5), (-, 2, 0, 2, 0, 2), (-, -, 4, 1, 2, 3),$$

$$(-, -, -, 4, 0, 4), (-, -, -, -, 2, 1), (-, -, -, -, -, 6)].$$

4. Comparison of Design Sequences

When we address this problem for design sequences of the same length n, there are effectively $2^n - 2$ such comparable sequences - barring the two extremes [all 0s and all 1's]. Actual number of admissible sequences may be much smaller - depending on the number K of non-negligible h-parameters. Anyway, such a comparison of two admissible sequences may rest on, say the criterion of smaller average variance or smaller generalized variance of the estimated h-parameters. Below we take up the case of a saturated model with $n = 7, K = 6$ and compare all available admissible design sequences. Note that we have already studied one such admissible design sequence in the above. In this case there are $2^7 - 2 = 126$ possible design sequences of length 7 each-barring the two inadmissible extreme allocations (viz., all 1's and all 0's). These design sequences can be classified into distinct types as follows.

$TypeI : (i)\ [1, 0, 0, 0, 0, 0, 0]$; $(ii)\ [1, 1, 0, 0, 0, 0, 0]$; $(iii)\ [1, 1, 1, 0, 0, 0, 0]$; $(iv)\ [1, 1, 1, 1, 0, 0, 0]$;

$$TypeI\ continued : (v)\ [1, 1, 1, 1, 1, 0, 0];\ (vi)\ 1, 1, 1, 1, 1, 1, 0]$$

and all their cyclic permutations-covering 42 design sequences;

$$TypeII : (i)\ [1, 0, 1, 0, 0, 0, 0];\ (ii)\ [1, 0, 0, 1, 0, 0, 0]$$

and all their cyclic permutations involving 2 non-consecutive 1's-covering 14 design sequences;

$TypeIII : (i)\ [1, 1, 0, 1, 0, 0, 0] -$ *replicated twice*; $(ii)\ [1, 1, 0, 0, 1, 0, 0]$; $(iii)\ [1, 0, 1, 0, 1, 0, 0]$

and all their cyclic permutations involving 3 non-consecutive 1's-covering 28 design sequences;

$TypeIV : (i)\ [1, 1, 1, 0, 1, 0, 0] -$ *replicated twice*; $(ii)\ [1, 1, 0, 1, 1, 0, 0]$; $(iii)\ [1, 1, 0, 1, 0, 1, 0]$

and all their cyclic permutations involving 4 non-consecutive 1's-covering 28 design sequences;

$$TypeV : (i)\ \ [1,1,1,1,0,1,0];\ (ii)\ \ [1,1,1,0,1,1,0]$$

and all their cyclic permutations involving 5 non-consecutive 1's-covering 14 design sequences.

Routine computations can be done to ascertain respective status of each of the design sequences listed above for any specified value of K-the number of non-negligible h-parameters. Below we show the detailed analysis of the design sequences of Type I.

Table 2: Type I(i): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	0	0	0	0	0	0	1	1
h_1	1	0	0	0	0	0	-1	2
h_2	0	1	0	0	0	0	-1	2
h_3	0	0	1	0	0	0	-1	2
h_4	0	0	0	1	0	0	-1	2
h_5	0	0	0	0	1	0	-1	2
h_6	0	0	0	0	0	1	-1	2

Table 3: Type I(ii) : Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	-1	1	-1	1	-1	1	7
h_1	0	1	-1	1	-1	1	-1	6
h_2	-1	1	0	0	0	0	0	2
h_3	0	0	0	1	-1	1	-1	4
h_4	-1	1	-1	1	0	0	0	4
h_5	0	0	0	0	0	1	-1	2
h_6	-1	1	-1	1	-1	1	0	6

Table 4: Type I(iii): Coefficient of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	-2	1	1	-2	1	13
h_1	0	0	1	-1	0	1	-1	4
h_2	-1	0	1	0	-1	1	0	4
h_3	0	-1	1	0	0	0	0	2
h_4	0	0	0	0	0	1	-1	2
h_5	-1	0	1	-1	0	1	0	4
h_6	0	-1	1	0	-1	1	0	4

Table 5: Type I(iv): Coefficient of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	-1	-1	3	-1	-1	-1	3	23
h_1	1	0	-1	1	0	0	-1	4
h_2	0	1	-1	0	1	0	-1	4
h_3	0	0	0	0	0	1	-1	2
h_4	0	0	-1	1	0	0	0	2
h_5	1	0	-1	0	1	0	-1	4
h_6	0	1	-1	0	0	1	-1	4

Table 6: Type I(v): Coefficient of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	-2	3	-2	3	-2	-2	3	43
h_1	0	0	1	-1	1	-1	0	4
h_2	0	0	0	0	0	1	-1	2
h_3	0	-1	1	-1	1	0	0	4
h_4	1	-1	0	0	0	1	-1	4
h_5	0	0	0	-1	1	0	0	2
h_6	1	-1	1	-1	0	1	-1	6

Table 7: Type I(vi): Coefficient of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	1	1	1	-5	1	31
h_1	0	0	0	0	0	1	-1	2
h_2	-1	0	0	0	0	1	0	2
h_3	0	-1	0	0	0	1	0	2
h_4	0	0	-1	0	0	1	0	2
h_5	0	0	0	-1	0	1	0	2
h_6	0	0	0	0	-1	1	0	2

To summarize the performances of the above design sequences of Type I, we find that in terms of average variance of the estimates of the h-parameters,

$$(i) = (vi) < (iii) = (iv) < (v) < (ii).$$

Table 8: Type II(i): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	-1	1	1	-1	-1	1	1	7
h_1	1	-1	0	1	0	-1	0	4
h_2	1	0	-1	1	1	-1	-1	6
h_3	0	0	0	0	1	0	-1	2
h_4	0	-1	0	1	0	0	0	2
h_5	1	-1	-1	1	1	-1	0	6
h_6	1	0	-1	0	1	0	-1	4

Table 9: Type II(ii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	1	-1	-1	-1	1	7
h_1	0	0	0	1	0	0	-1	2
h_2	-1	0	0	1	1	0	-1	4
h_3	-1	-1	0	1	1	1	-1	6
h_4	-1	-1	-1	1	1	1	0	6
h_5	0	-1	-1	0	1	1	0	4
h_6	0	0	-1	0	0	1	0	2

Table 10: Type III(i): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
2θ	2	-1	2	-1	-1	-1	2	4
$2h_1$	0	1	-1	1	0	0	-1	1
$2h_2$	-1	1	0	0	1	0	-1	1
$2h_3$	-1	0	0	1	0	1	-1	1
$2h_4$	-1	0	-1	1	1	0	0	1
$2h_5$	0	0	-1	0	1	1	-1	1
$2h_6$	-1	1	-1	0	0	1	0	1

We have completed computations of $\hat{h}$s along with their variances for all the effectively sixteen (16) competing design sequences. We may now display the totals of variances across all competitors.

$TypeI(i)$ 12; $TypeI(ii)$ 24; $TypeI(iii)$ 20; $TypeI(iv)$ 20; $TypeI(v)$ 22; $TypeI(vi)$ 12

$TypeII(i)$ 24; $TypeII(ii)$ 24

$TypeIII(i)$ 6; $TypeIII(ii)$ 20; $TypeIII(iii)$ 20

$TypeIV(i)$ 6; $TypeIV(ii)$ 20; $TypeIV(iii)$ 20

$TypeV(i)$ 24; $TypeV(ii)$ 24

Table 11: Type III(ii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	-2	1	-2	1	1	13
h_1	0	0	1	-1	1	-1	0	4
h_2	0	0	1	0	0	0	-1	2
h_3	-1	0	1	0	1	-1	0	4
h_4	0	-1	1	0	1	0	-1	4
h_5	-1	0	0	0	1	0	0	2
h_6	0	-1	1	-1	1	0	0	4

Table 12: Type III(iii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	1	1	-1	-1	1	13
h_1	0	0	0	0	1	0	-1	2
h_2	-1	0	0	0	1	1	-1	4
h_3	-1	0	1	0	1	-1	0	4
h_4	0	-1	-1	0	1	1	0	4
h_5	0	0	-1	-1	1	1	0	4
h_6	0	0	0	-1	0	1	0	2

Table 13: Type IV(i): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	1	1	-1	1	-1	-1	1	7
$2h_1$	0	0	1	-1	1	0	-1	1
$2h_2$	-1	0	1	0	0	1	-1	1
$2h_3$	-1	-1	1	0	1	0	0	1
$2h_4$	0	-1	0	0	1	1	-1	1
$2h_5$	-1	0	0	-1	1	1	0	1
$2h_6$	0	-1	1	-1	0	1	0	1

In conclusion, we find that the design sequences $TypeIII\ (i) : [1, 1, 0, 1, 0, 0, 0]$ and $TypeIV\ (i) : [1, 1, 1, 0, 1, 0, 0]$ are, together with their cyclic permutations, most efficient with respect to the average variance criterion. It is again readily observed that for both these designs, pair- wise covariance terms of the estimates of the h-parameters are all equal and it is the same for both. Therefore, as such, the two competing sequences are information-equivalent !

Our task will not be complete unless we discuss one more pertinent observation in this context. The above comparison may not be 'fair' since the design sequences are based on

Table 14: Type IV(ii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	3	-1	3	-1	-1	-1	-1	23
h_1	-1	1	-1	1	0	0	0	4
h_2	-1	0	0	0	1	0	0	2
h_3	-1	0	-1	1	0	1	0	4
h_4	-1	0	-1	0	1	1	0	4
h_5	0	0	-1	0	0	1	0	2
h_6	-1	1	-1	0	0	0	1	4

Table 15: Type IV(iii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	3	-1	-1	-1	-1	-1	3	23
h_1	0	1	0	0	0	0	-1	2
h_2	-1	1	1	0	0	0	-1	4
h_3	-1	0	1	1	0	0	-1	4
h_4	-1	0	0	1	1	0	-1	4
h_5	-1	0	0	0	1	1	-1	4
h_6	-1	0	0	0	0	1	0	2

Table 16: Type V(i): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	3	3	-2	-2	3	-2	-2	43
h_1	-1	0	1	0	-1	1	0	4
h_2	-1	-1	1	1	-1	0	1	6
h_3	0	-1	0	1	0	0	0	2
h_4	-1	0	0	0	0	1	0	2
h_5	-1	-1	1	0	-1	1	1	6
h_6	0	-1	0	1	-1	0	1	4

unequal number of 1s. Note that every sequence comprises of 1s and 0s and the understanding is that a 0-phase corresponds to 'idle' phase while a 1-phase is 'active'. So the number of active phases should also be considered while examining relative performances. We may apply the usual concept of "Efficiency" and work out "Efficiency per active phase". For a single parameter, efficiency is directly related to and measured by [Fisher] Information. For $K = 6$ h-parameters, we can compute the average variance of the estimates and multiply it by the number of 1s and minimize this quantity. If we are guided by this consideration, we find that the design sequence $TypeIII\ (i) : [1, 1, 0, 1, 0, 0, 0]$ is the best of all! We can argue that this is also the best with respect to generalized variance criterion as well.

Table 17: Type V(ii): Coefficients of estimates of h-parameters and their variances

parameter / coefficient	y_1	y_2	y_3	y_4	y_5	y_6	y_7	Variance coefficient
θ	3	3	-2	-2	-2	-2	3	43
h_1	0	0	1	0	0	0	-1	2
h_2	-1	0	1	1	0	0	-1	4
h_3	-1	-1	1	1	1	0	-1	6
h_4	-1	-1	0	1	1	1	-1	6
h_5	-1	-1	0	0	-1	-1	0	4
h_6	0	-1	0	0	0	1	0	2

Remark 3: At the end of Section 2, we had introduced two design sequences D_5 and $D_{alt.\ 5}$ for the case of $n = 5,\ K^* = 4$. We also observed that the two sequences possess the same generalized variance. However, it can be seen that the alternative sequence provides smaller average variance. Now we note that whereas in D_5 the number of active phases used was 3, in the alternative design this number was 4. As in the above, we borrow the concept of "Efficiency per observation" while this time we define the "Efficiency" as reciprocal of the generalized variance, raised to the power 1/4 since there are 4 h-parameters. Otherwise, we can also use the reciprocal of the average variance. Adjusting for the difference in the number of active phases, we conclude that (i) D_5 is better than $D_{alt.\ 5}$ under the generalized variance criterion, while (ii) alternative sequence is better under average variance criterion.

Acknowledgments

The author is thankful to Professor Rajender Parsad of IASRI, New Delhi for bringing to his notice the publication by Kao *et al.* (2008) cited in the list of references. He also expresses his thanks to Professors Nripes K Mandal and Manisha Pal, Department of Statistics, Calcutta University, for taking keen interest in this topic of research and for fruitful discussions from time to time. The author also acknowledges citation of a related study in Maus *et al.* (2010) by an anonymous referee. One research collaborator, Dr Sobita Sapam, has kindly helped the author in formatting of the latex version following the template of instructions.

References

Cheng, C. S. and Kao, M. H. (2015). Optimal experimental designs for fMRI via circulant biased weighing designs. *Annals of Statistics*, **43(6)**, 2565–2587.

Kunert, J. (1984). Optimality of balanced uniform repeated measurements designs. *Annals of Statistics,* **12(4)**, 1006–1017.

Maus, B., Van Breukelen, G. J. P., Goebel, R. and Berger, M. P. F.(2010). Optimization of blocked designs in fMRI studies. *Psychometrika*, **75**, 373–390.

Kao, M.-H., Mandal, A. and Stufken, J. (2008: New series). Optimal design for event-related functional magnetic resonance imaging considering both individual stimulus effects and pairwise contrasts. *Statistics and Applications*, **6(1,2)**, 235–256.

Raghavarao, D. (1971). *Construction and Combinatorial Problems in Design of Experiments.* John Wiley and Sons Inc. New York.

Shah, K. R. and Sinha, Bikas K. (1989). *Theory of Optimal Designs.* Springer Lecture Notes in Statistics Series No. 54.

Special Proceeding of the 22 nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02–04, 2020, pp 53 –58

Statistical Inference for a One Unit System With Dependent Structure

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

Received: 30 May 2020; Revised: 16 June 2020; Accepted: 20 June 2020

Abstract

Under the assumption that the lifetime and repair time of a one unit system is bivariate exponential, measures of system performance such as system reliability, MTBF, point availability and steady state availability are obtained. Further, a $100(1-\alpha)\%$ asymptotic confidence interval for steady state availability of the system is derived.

Key words: Bivariate exponential distribution; CAN estimator; One unit system; Slutsky theorem; Steady state availability.

AMS Subject Classifications: 60K10, 60F05

1. Introduction

Analysis of one unit repairable system has received considerable attention and has been extensively studied by several researchers in the past. A system is said to be one unit system if it is made up of only one component or it has a single crucial component whose failure causes the system to fail (Barlow and Hunter, 1961). If the lifetime density and repair time density of the unit are assumed to be arbitrary, then one may obtain highly formal expressions for the probability distributions and other quantities of interest. These expressions are rarely suitable for numerical computations. In most of the cases, analytically explicit expressions are obtained only under negative exponential distributional assumptions. PH distribution introduced by Neuts (1975) is more general in the sense that Erlang distribution and negative exponential distribution are only particular cases of continuous PH distribution and this distribution can be used to describe the lifetimes or repair times of a unit when it is non-exponential. Chandrasekhar and Natarajan (2000) have obtained several measures of system performance of a complex one unit system assuming that the lifetime and repair time of the unit has PH distribution with different representations.

Generally speaking, the failure time and repair time are assumed to be independent random variables. However, this assumption of independence need not hold good always. For example, a component that fails frequently has to be thoroughly examined for defects before it is put back into the system so as to prevent future failures, thereby increasing its

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

repair time. Thus the repair time and lifetime (failure time) are dependent on each other. Hence, both from theoretical and application perspective, analysing reliability models with dependent structure will be of much use. In the past, a number of bivariate exponential distributions have been proposed and studied well in the literature to describe the lifetimes of two unit systems. But the bivariate exponential distribution proposed by Marshall and Olkin (1967) is widely used among researchers because of its properties. An attempt is made in this paper to derive performance measures of one unit system under the assumption that the lifetime and repair time of the unit are governed by Marshall-Olkin bivariate exponential distribution. Also, point and interval estimation of steady state availability of the system is carried out.

2. Model (One Unit System With Dependent Structure)

The system description and assumptions involved are given below:

The system under consideration consists of only one unit with dependent structure and when it fails, it is taken up for repair instantaneously. Let T and R denote respectively the lifetime and repair time of the failed unit. Since the lifetime and repair time is assumed to be dependent, Marshall-Olkin bivariate exponential distribution for T and R with the survival function given by

$$\bar{F}(t,r) = e^{-[\lambda_1 t+\lambda_2 r+\lambda_3 max(t,r)]}, t, r > 0; \lambda_1, \lambda_2 > 0, \lambda_3 \geq 0 \tag{1}$$

is considered. see Marshall and Olkin (1967).

The stochastic process underlying the behaviour of the system is an alternating renewal process. At time $t = 0$, the unit just begins to operate. It should be noted that

- The lifetime T and repair time R are exponential random variables each with parameters $(\lambda_1 + \lambda_3)$ and $(\lambda_2 + \lambda_3)$ respectively.
- $E(T) = \frac{1}{(\lambda_1+\lambda_3)}$ and $E(R) = \frac{1}{(\lambda_2+\lambda_3)}$.
- $V(T) = \frac{1}{(\lambda_1+\lambda_3)^2}$ and $V(R) = \frac{1}{(\lambda_2+\lambda_3)^2}$.
- The covariance between T and R is given by $Cov(T,R) = \frac{\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}$.
- T and R are independent if and only if $\lambda_3 = 0$.

3. Operating Characteristics of The System

In this section, various measures of performance that describe the operating characteristics of the system are discussed. To analyse the system, we note that at any given time t, the system will be found in any one of the following mutually exclusive and exhaustive states namely, state 0: the unit is operating (online) and state 1: the unit is under repair.

Here state 0 denotes the system upstate and state 1 denotes the system down state. Let $X(t)$ denote the state of the system (unit) at time t. Clearly, $\{X(t), t \geq 0\}$ is a continuous time Markov process with the state space given by $E = \{0, 1\}$. Since bivariate Marshall and Olkin distribution satisfy the bivariate lack of memory property and its marginal distributions are exponential, it follows that $\{X(t), t \geq 0\}$ is a Markov process with infinitesimal generator Q given by

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{pmatrix} -(\lambda_1+\lambda_3) & (\lambda_1+\lambda_3) \\ (\lambda_2+\lambda_3) & -(\lambda_2+\lambda_3) \end{pmatrix} \end{array}. \tag{2}$$

Let $p_i(t) = P[X(t) = i], i = 0, 1$ represent the probability that the system is in state i at time t with the initial condition $p_0(0) = 1$.

3.1. System reliability

Since system reliability $R(t)$ is the probability of failure free operation of the system in the time interval $(0, t]$ and the marginal distribution of T is exponential with mean $(\lambda_1 + \lambda_3)$, an expression for $R(t)$ can be readily obtained as

$$R(t) = e^{-(\lambda_1+\lambda_3)}. \tag{3}$$

Also, the system Mean Time Before Failure (MTBF) is given by

$$MTBF = \int_0^\infty R(t)dt = \frac{1}{(\lambda_1+\lambda_3)}. \tag{4}$$

3.2. Point and steady state availability of the system

The system availability $A(t)$ is the probability that the system operates within the tolerances at a given instant of time t and is obtained as follows:

From the infinitesimal generator given in (2), the following system of differential – difference equations are obtained.

$$dp_0(t)dt = -(\lambda_1+\lambda_3)p_0(t) + (\lambda_2+\lambda_3)p_1(t) \tag{5}$$

$$dp_1(t)dt = (\lambda_1+\lambda_3)p_0(t) - (\lambda_2+\lambda_3)p_1(t) \tag{6}$$

Let $L_i(s)$ be the Laplace transformation of $p_i(t), i = 0, 1$. Taking Laplace transforms on both the sides of the above differential-difference equations and solving for $L_i(s), i = 0, 1$ and inverting, we get $p_i(t), i = 0, 1$ as follows:

$$P_0(t) = \frac{(\lambda_2+\lambda_3)}{[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]} + \frac{(\lambda_1+\lambda_3)}{[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]} e^{-[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]t} \tag{7}$$

$$P_1(t) = \frac{(\lambda_1+\lambda_3)}{[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]} - \frac{(\lambda_1+\lambda_3)}{[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]} e^{-[(\lambda_2+\lambda_3)+(\lambda_1+\lambda_3)]t} \tag{8}$$

Hence, the point availability of the system is given by $A(t) = p_0(t)$. The system steady state availability is the expected fractional amount of time in a continuum of operating time that the system is in upstate and is given by

$$A_\infty = \lim_{t\to\infty} A(t) = \frac{(\lambda_2+\lambda_3)}{[(\lambda_1+\lambda_3)+(\lambda_2+\lambda_3)]}. \tag{9}$$

In the next section, point and interval estimation of the steady state availability of the system is carried out.

4. Point and Interval Estimation of A_∞

Let $(Y_{1i}, Y_{2i}), i = 1, 2, ..., n$ be a random sample of size n from a bivariate exponential population with the survival function given by (1). Here Y_1 denote the lifetime of the unit and Y_2 denote the repair time of the unit upon failure. Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the corresponding sample means.

4.1. Point estimator of A_∞

It can be established that $\bar{Y}_1$ and $\bar{Y}_2$ are the moment estimators of $\frac{1}{(\lambda_1+\lambda_3)}$ and $\frac{1}{(\lambda_2+\lambda_3)}$ respectively. Writing $\theta_1 = \frac{1}{(\lambda_1+\lambda_3)}$ and $\theta_2 = \frac{1}{(\lambda_2+\lambda_3)}$, A_∞ of the system given in (9) reduces to

$$A_\infty = \frac{\theta_1}{\theta_1+\theta_2}. \tag{10}$$

Hence, a point (moment) estimator of the steady state availability A_∞ of the system is given by

$$\hat{A}_\infty = \frac{\bar{Y}_1}{\bar{Y}_1+\bar{Y}_2}. \tag{11}$$

It may be noted that $\hat{A}_\infty$ given in (11) is a real valued differentiable function in $\bar{Y}_1$ and $\bar{Y}_2$. Applying multivariate central limit theorem, it is readily seen that
$\sqrt{n}[(\bar{Y}_1, \bar{Y}_2) - (\theta_1, \theta_2)] \xrightarrow{d} N_2(0, \Sigma)$ as $n \to \infty$, where the dispersion matrix Σ is given by

$$\Sigma = \begin{array}{c} \\ \bar{Y}_1 \\ \bar{Y}_2 \end{array} \begin{array}{c} \begin{array}{cc} \bar{Y}_1 & \bar{Y}_2 \end{array} \\ \begin{pmatrix} \theta_1^2 & \frac{\lambda_3\theta_1^2\theta_2^2}{\theta_1+\theta_2-\lambda_3\theta_1\theta_2} \\ \frac{\lambda_3\theta_1^2\theta_2^2}{\theta_1+\theta_2-\lambda_3\theta_1\theta_2} & \theta_2^2 \end{pmatrix} \end{array} \tag{12}$$

Applying the results in Chapter 6 of Rao (2009), we have

$\sqrt{n}(\hat{A}_\infty - A_\infty) \xrightarrow{d} N(0, \sigma^2(\theta))$, where $\theta = (\theta_1, \theta_2)$ and

$$\sigma^2(\theta) = \sum_{i=1}^{2} \theta_i^2 \left(\frac{\partial A_\infty}{\partial \theta_i}\right)^2 + \frac{2\lambda_3\theta_1^2\theta_2^2}{\theta_1+\theta_2-\lambda_3\theta_1\theta_2} \left(\frac{\partial A_\infty}{\partial \theta_1}\right)\left(\frac{\partial A_\infty}{\partial \theta_2}\right). \tag{13}$$

The expressions for the partial derivatives are obtained as

$$\frac{\partial A_\infty}{\partial \theta_1} = \frac{\theta_2}{(\theta_1+\theta_2)^2}$$

and

$$\frac{\partial A_\infty}{\partial \theta_2} = \frac{-\theta_1}{(\theta_1+\theta_2)^2}.$$

By substituting the partial derivatives in (13) and simplifying, we get

$$\sigma^2(\theta) = \frac{2\theta_1^2\theta_2^2}{(\theta_1+\theta_2)^4} - \frac{2\lambda_3\theta_1^2\theta_2^2}{(\theta_1+\theta_2)^4(\theta_1+\theta_2-\lambda_3\theta_1\theta_2)}. \tag{14}$$

Thus, $\hat{A}_\infty$ is a consistent and asymptotic normal (CAN) estimator of A_∞.

4.2. Interval estimation of A_∞

Let $\hat{\sigma^2}$ be an estimator of $\sigma^2(\theta)$ obtained by replacing θ by its consistent estimator namely, $\hat{\theta} = (\bar{Y}_1, \bar{Y}_2)$. Since $\sigma^2(\theta)$ is a continuous function of θ, $\hat{\sigma^2}$ is a consistent estimator of $\sigma^2(\theta)$ i.e., $\hat{\sigma^2} \xrightarrow{p} \sigma^2(\theta)$ as $n \to \infty$. Thus, applying Slutsky theorem, we get,

$$\frac{\sqrt{n}(\hat{A}_\infty - A_\infty)}{\sqrt{\hat{\sigma^2}}} \xrightarrow{d} N(0,1).$$

Hence, for $\alpha \in (0,1)$, we have

$$P\left(-Z_{\frac{\alpha}{2}} < \frac{\sqrt{n}(\hat{A}_\infty - A_\infty)}{\sqrt{\hat{\sigma^2}}} < Z_{\frac{\alpha}{2}}\right) = (1-\alpha)$$

where $Z_{\frac{\alpha}{2}}$ denote the upper $(\frac{\alpha}{2})^{th}$ percentile point of the standard normal distribution. Thus, a $100(1-\alpha)\%$ confidence interval for the steady state availability A_∞ of the system is given by $\hat{A}_\infty \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma^2}}{n}}$, where $\hat{\sigma^2}$ is obtained from (14).

5. Conclusion

Inferential aspects of performance measures within the framework of reliability models with dependent structure is less addressed in statistical literature. The present article has focused on point and interval estimation of system reliability and steady state availability of a one unit system assuming the lifetime and repair time to be modelled by bivariate Marshall-Olkin distribution. Such models can be applied in reliability studies involving electrical devices like, for example, water pumps used in commercial, agricultural and domestic activities, ceiling fans, wherein lifetime and repair time of the motor are dependant. The methodology adapted can be applied to models involving more than one component by using multivariate Marshall-Olkin distribution for modelling the dependency. Interested readers are encouraged to go through Yadavalli *et al.* (2017), Vaidyanathan and Chandrasekhar (2018) and the references cited therein.

Acknowledgements

I thank my collaborator Dr. P. Chandrasekhar, Associate Professor, Loyola College, Chennai, India for some discussions and technical assistance in writing this article.

References

Barlow, R. E. and Hunter, L. C. (1961). Reliability analysis of a one-unit system. *Operations Research*, **9(2)**, 145 - 291.

Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing.* Holt - Reinhart, New York.

Chandrasekhar, P. and Natarajan, R. (2000). On the applications of PH-distribution to a complex one unit system. In: *Proceedings of second International Conference on Mathematical Methods in Reliability (MMR2000)*, 269 - 272, held in Bordeaux, France, July 4 -7, 2000.

Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association*, **62**, 30 - 44.

Neuts, M. F. (1975). Probability distribution of phase type. In: *Liber Amicorum, Professor Emeritus H. Florin*, 173 - 206, Department of Mathematics, University of Louvain, Belgium.

Rao, C. R. (2009). *Linear Statistical Inference and its Applications.* Second edition. Wiley Series in Probability and Statistics, Wiley India Pvt. Limited, India.

Vaidyanathan, V. S. and Chandrasekhar, P. (2018). Inferential aspects of a three unit hot standby system with dependent structure. *IAPQR Transactions*, **43(1)**, 47 - 59.

Yadavalli, V. S. S., Vaidyanathan, V. S., Chandrasekhar, P. and Abbas, S. (2017). Applications of quadrivariate exponential distribution to a three unit warm standby system with dependent structure. *Communications in Statistics - Theory and Methods*, **46(14)**, 6782 - 6790.

Special Proceedings of the 22nd Annual Conference of SSCA held at S.P. Pune University, Pune, during January 02-04, 2020, pp 59–71

A Straight Forward Approach in Understanding the Improperness of ROC Curve

R. Vishnu Vardhan[1], S. Balaswamy[2] and G. Sameera[3]
[1]*Department of Statistics, Pondicherry University, Puducherry*
[2]*Department of Statistics, Indira Gandhi National Tribal University, Madhya Pradesh*
[3]*Biostatistics and Pharmacometrics, Global Drug Development, Novartis Healthcare Private Limited, Hyderabad*

Received: 12 June 2020; Revised: 20 June 2020; Accepted: 22 June 2020

Abstract

The present paper is based on the framework of a classification tool namely, Multivariate Receiver Operating Characteristic (MROC) curve, which is modelled to provide a better classification. In general, there are certain properties where the proposed ROC curve has to satisfy, violating any of such property leads to inappropriate conclusions about the classifier. In this paper, a straight forward approach is presented to explain the nature of 'Proper' and 'Improper' ROC curves. The methodology is supported with both simulated and real data sets.

Key words: MROC curve; AUC, Improper MROC curve; Inflection point; Crossing point.

AMS Subject Classifications: 92B15, 62P10

1. Introduction

Over the past seven to eight decades, the problem of detecting/identifying one's behavior and allocating them into one of the population gained lot of attention. Such work was majorly observed and initially related in the fields of Experimental Psychology and Signal Detection Theory (Tanner and Swets, 1954, Green and Swets, 1966). However, with the involvement of statistical essentials, this area branched to diversified fields of Science and Technology, namely Diagnostic Medicine, Banking, Finance and many more (Lusted, 1971, Krzanowski and Hand, 2009). All these come under the hub of classification tools/techniques (Statistical Decision Theory). The practice of allocation or separation is based on certain characteristics of univariate or multivariate in nature.

Initial application of this was in Medicine and used majorly to identify the individual's health status by defining an optimal threshold for a biomarker observed in the case of that particular disease. The first parametric ROC is the Binormal ROC Curve where the

Correponding Author: S. Balaswamy
Email:balaswamy.stats@gmail.com
(The third author, G. Sameera works for Novartis. Views expressed in this work are her own)

variable under study for two independent populations (healthy/diseased or signal/noise) follow Normal distributions (Green and Swets, 1966). The important properties of ROC curve are:

(i) $y = h(x)$ is the mathematical model of the ROC curve, where y denotes the true positive rate and x denotes the false positive rate. The curve is a monotonic increasing function in the positive quadrant, lying between $y = 0$ at $x = 0$ and $y = 1$ at $x = 1$.

(ii) The ROC curve is unaltered if the classification scores undergo a strictly increasing transformation.

(iii) The slope of the ROC curve (likelihood ratio of ROC curve) at threshold value 'c' is always positive and given by

$$\frac{dy}{dx} = \frac{P(U > c|1)}{P(U > c|0)}$$

When dealing with practical problems, we often come across the presence or involvement of several variables to have a classifier rule for a better classification. Su and Liu (1993), Reiser and Ferragi (1997), Schisterman *et al.* (2004), Liu *et al.* (2005), Yuan and Ghosh (2008), Chang and Park (2009) and Sameera *et al.* (2016) are a few to cite among those who proposed an extension of univariate ROC model to multivariate. However the present work is based upon the Multivariate ROC (MROC) model proposed by Sameera *et al.* (2016), as they showed that this model works better than the model proposed by Su and Liu (1993) and their model is applicable to data where the covariance structures of two populations can be proportional or non-proportional. As mentioned about the properties of the ROC curve, the most important one to verify is its concavity *i.e.*, slope of the ROC curve is always positive. Now the question that arises is, what happens if a curve is not satisfying the concavity property? If the curve violates this property, it might affect the accuracy of the test as well as the optimal cutoff point defined for that particular test. Mathematically, a meaningful decision variable should be an increasing function of the likelihood ratio (Pepe, 2003) and such MROC curve is said to be "Proper". A function whose first derivative is decreasing throughout an open interval is called concave in that interval, and a function whose first derivative is increasing throughout an open interval is called convex in that interval. Since the slope of an MROC curve for a continuous decision variable is equal to the likelihood ratio at the corresponding threshold, it follows that the slope of a MROC curve decreases as the false positive rate (FPR) increases, that is, a MROC curve will be concave everywhere ($0 \leq FPR \leq 1$). If the decision variable is not an increasing function of the likelihood function, then its model and corresponding MROC curve are said to be improper.

2. Illustration of Improper MROC (iMROC) Curve

Consider the following example which illustrates the Indian Liver Patients (ILP) dataset for which the MROC curve has been drawn and depicted in Figure 1. The fitted MROC curve seems to be proper but when observed keenly; the improperness of the curve can be

witnessed. In such situation, the usual MROC curve methodology might not project the true accuracy of a test and will not be used for future classification. Figure 1 visualizes the two crucial points namely, Crossing Reference line ((t_0) or Crossing Point) and Inflection Reference line ((t_1) or Inflection Point). Figure 1 shows the corresponding fitted MROC curve; note that there is a visible 'dip' in the curve crossing the chance line near the upper right hand corner of the unit square plot. In Figure 1, MROC curve crosses the chance line at the point (1-Specificity, Sensitivity) = (0.96, 0.96), shown by the intersection of the "crossing" reference line with the MROC curve. Furthermore, this MROC curve is concave for $FPR < 0.76$, but is convex for $FPR > 0.76$. Therefore, the MROC curve which separates the concave and convex portions of the curve is called the "Inflection Point (t_1)". Similarly, the MROC curve which crosses the chance line at the point where FPR=TPR is called the "Chance line crossing point or Crossing Point (t_0)". From Figure 1, though the

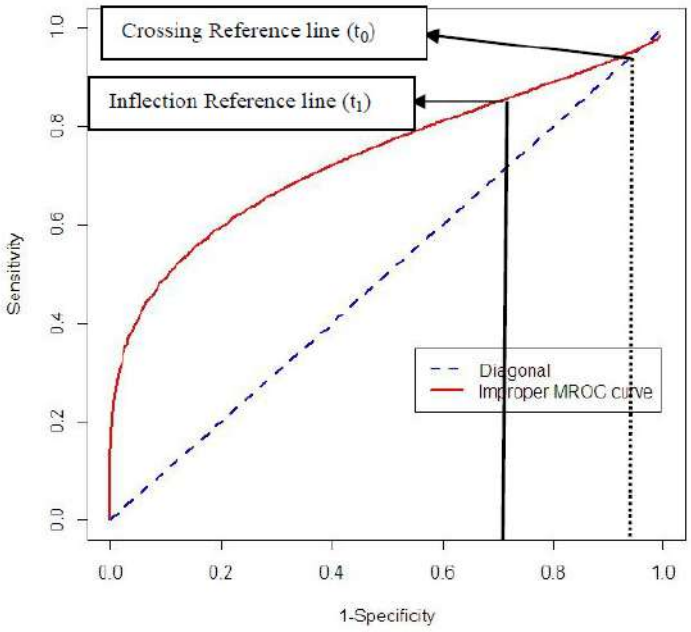


Figure 1: Improper MROC curve for ILP dataset

dip of the curve is visible *i.e.* the MROC curve is not concave everywhere, it is not possible to identify the inflection point visually. Even in the case of improper MROC curves, it is not that easy to identify the point where the curve changes from concave to convex. In order to deal with this situation, the ways to measure the improperness of an MROC curve is shown in subsequent sections with the help of real and simulated data sets.

2.1. MROC curve

Let U_0 and $U_0 \in$ U be the vectors of test scores of two independent multivariate normal populations with mean vectors μ_0, μ_1 and co-variance matrices Σ_0 and Σ_1 with m and n sample sizes respectively.

$$U_i \sim MVN(\mu_i, \Sigma_i); i = 0, 1$$

$$f(X|\mu_i, \Sigma_i) = \frac{1}{(2\Pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{N}{2}}} e^{\frac{-1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)}$$

Let $x(c)$ denote the false positive rate (FPR) and $y(c)$ denote the true positive rate (TPR) where 'c' is the threshold value. The expressions for FPR, TPR are

$$FPR = x(c) = P(U > c|0) = 1 - \Phi\left(\frac{c - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right) \tag{1}$$

$$TPR = y(c) = P(U > c|1) = \Phi\left(\frac{b'\mu_1 - c}{\sqrt{(b'\Sigma_1 b)}}\right) \tag{2}$$

where $b(\neq 0)$ be a $k \times 1$ vector. The threshold value thus obtained using (1) is given as

$$c = b'\mu_0 + \sqrt{(b'\Sigma_0 b)}\Phi^{-1}(1-x) \tag{3}$$

where $\Phi^{-1}(.)$ is the inverse function of $\Phi(.)$

substituting (3) in (2) implies that

$$TPR = y(c) = \Phi\left(\frac{b'(\mu_1 - \mu_0) - \sqrt{(b'\Sigma_0 b)}\Phi^{-1}(1-x)}{\sqrt{(b'\Sigma_1 b)}}\right) \tag{4}$$

which is the form of Multivariate ROC model (Sameera *et al.*, 2016)

The AUC of MROC curve is

$$AUC = \Phi\left\{\frac{b'(\mu_1 - \mu_0)}{\sqrt{(b'(\Sigma_0 + \Sigma_1)^{-1}b}}\right\} \tag{5}$$

2.2. Crossing point

In order to verify whether the generated ROC curve is 'proper' or 'improper', Balaswamy *et al.* (2020) came out with two measures namely crossing point and inflection point. The mathematical framework of these measures are adopted here to maintain the continuity of explanation about proper vs improper ROC curves. Let 'c' denote threshold to a chance line crossing FPR, then

$$P(U > c|0) = P(U > c|1)$$

$$\Rightarrow \Phi\left(\frac{c - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right) = \Phi\left(\frac{c - b'\mu_1}{\sqrt{(b'\Sigma_1 b)}}\right)$$

on further simplification, the expression for c_0 crossing threshold is

$$c_0 = \frac{(b'\mu_0)\sqrt{(b'\Sigma_1 b)} - (b'\mu_1)\sqrt{(b'\Sigma_0 b)}}{\sqrt{(b'\Sigma_1 b)} - \sqrt{(b'\Sigma_0 b)}} \tag{6}$$

Let t_0 denote the chance line crossing FPR corresponding to c_0. Then

$$t_0 = P(U > c_0|0) = 1 - \Phi\left(\frac{c_0 - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right)$$

on substituting (6) in the above expression, we obtain the expression for crossing point as,

$$t_0 = \Phi\left(\frac{b'\mu_1 - b'\mu_0}{\sqrt{(b'\Sigma_1 b)} - \sqrt{(b'\Sigma_0 b)}}\right) \tag{7}$$

Uniqueness of t_0 follows from the uniqueness of c_0.

2.3. Inflection point

The slope of ROC curve is twice differentiable. From basic calculus results concerning concave functions it follows that the MROC curve is concave (convex) over an open interval if its second derivative is negative (positive) throughout the interval (0, 1). The approach is to show that the second derivative of the MROC curve is negative throughout $(0, t_1)$ and positive throughout $(t_1, 1)$ if $v < 1$, and positive throughout $(0, t_1)$ and negative throughout $(t_1, 1)$ if $v > 1$.

Let t denote an FPR with corresponding threshold c. The derivative of the MROC curve evaluated at t is equal to the likelihood ratio evaluated at c, *i.e.*,

$$\frac{\partial ROC(t)}{\partial t} = LR(c)$$

i.e., at $t = t_0$

$$\frac{\partial ROC(t)}{\partial t}/t = t_0 = LR(c_0)$$

it follows, using the chain rule, that

$$\frac{\partial^2 ROC(t)}{\partial^2 t} = \frac{\partial LR(c)}{\partial c}\frac{\partial c}{\partial t} \tag{8}$$

since,

$$t = P(U > c|0) = 1 - P(U < c|0) = 1 - \Phi\left(\frac{c_0 - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right)$$

then t is a strictly decreasing function of c and

$$\frac{\partial c}{\partial t} = -\frac{1}{\sqrt{(b'\Sigma_0 b)}}\varphi\left(\frac{c_0 - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right)$$

therefore, the equation (8) can be rewritten as,

$$\frac{\partial^2 ROC(t)}{\partial^2 t} = \frac{\partial LR(c)}{\partial c}\left[-\frac{1}{\sqrt{(b'\Sigma_0 b)}}\varphi\left(\frac{c_0 - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right)\right]^{-1} \tag{9}$$

Since $\varphi\left(\frac{c_0-b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right) > 0$, it follows from Equation (9) that the second derivative of the MROC curve and the derivative of the likelihood ratio have opposites signs when evaluated at t and c, respectively (Balaswamy *et al.*, 2020).

The threshold value at the inflection point is given by

$$c_1 = \frac{(b'\Sigma_1 b)(b'\mu_0) - (b'\Sigma_0 b)(b'\mu_1)}{(b'\Sigma_1 b) - (b'\Sigma_0 b)} \tag{10}$$

then the corresponding FPR is

$$t_1 = 1 - \Phi\left(\frac{c_1 - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right)$$

on substituting c_1 in the above equation, the FPR at the corresponding c_1 is given by

$$t_1 = 1 - \Phi\left(\frac{\left\{\frac{(b'\Sigma_1 b)(b'\mu_0)-(b'\Sigma_0 b)(b'\mu_1)}{(b'\Sigma_1 b)-(b'\Sigma_0 b)}\right\} - b'\mu_0}{\sqrt{(b'\Sigma_0 b)}}\right) \tag{11}$$

on further simplification, the FPR value at the inflection point is as follows

$$t_1 = \Phi\left(\frac{(b'\mu_1 - b'\mu_0)\sqrt{(b'\Sigma_0 b)}}{(b'\Sigma_1 b) - (b'\Sigma_0 b)}\right) \tag{12}$$

Since the derivative of the log likelihood ratio will have opposite sign of the second derivative of the MROC curve evaluated at the corresponding FPR and thresholds less than c_1 correspond to FPRs greater than t_1 and vice versa, the FPR value

$$t_1 = \Phi\left(\frac{(b'\mu_1 - b'\mu_0)\sqrt{(b'\Sigma_0 b)}}{(b'\Sigma_1 b) - (b'\Sigma_0 b)}\right)$$

is the unique inflection point FPR and

$$c_1 = \frac{(b'\Sigma_1 b)(b'\mu_0) - (b'\Sigma_0 b)(b'\mu_1)}{(b'\Sigma_1 b) - (b'\Sigma_0 b)}$$

is its corresponding inflection point threshold.

3. Results and Discussion

3.1. Proper MROC curve

In order to explain the concept of Proper MROC curve, the Statlog (heart) data taken from UCI repository is used. The heart dataset consists of 270 samples of which 120 (44.4%) are diagnosed with presence of heart disease and 150 (55.6%) with absence of heart disease. The parameters age, sex (Male: 183, 67.78% & Female: 87, 32.22%), chest pain type

(4 nominal values), resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic (ECG) results (0, 1&2), maximum heart rate achieved, exercise induced angina, oldpeak, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy and thal (normal, fixed defect & reversible defect) are considered for diagnosis. MROC curve is fitted and it corresponding linear combination is

$$\begin{aligned} U = &-0.022 * Age + 1.323 * Sex + 0.829 * Chestpaintype + 0.019 * Restingbloodpressure \\ &+ 0.005 * SerumCholesterol - 0.724 * Fastingbloodsugar \\ &+ 0.358 * RestingECGresults - 0.025 * Maximumheartrate \\ &+ 1.091 * Exerciseinducedangina + 0.424 * Oldpeak + 0.534 * thal \\ &+ 0.398 * SlopeofthepeakexerciseSTsegment + 1.269 * Numberofmajorvessels \end{aligned}$$

This linear combination helps us to know the status of a new individual basing on the U value. From the results, the curve is found to be proper by satisfying the property of monotonic likelihood ratio of MROC curve, hence it is a Proper MROC curve and the figure is depicted in Figure 2.

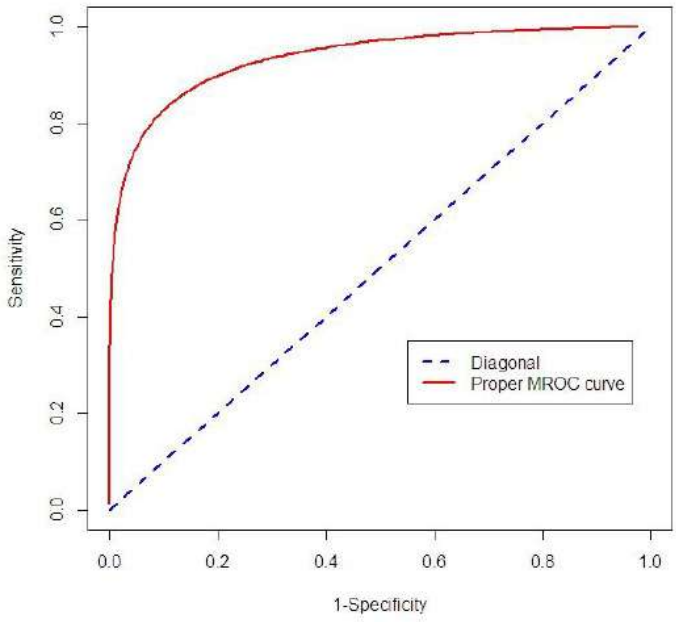


Figure 2: Proper MROC curve for Heart dataset

The optimal threshold value for identifying heart disease in an individual when the above mentioned characteristics studied is 7.27 with accuracy (AUC) of 93.7%. If score obtained for a new patient U is greater than 7.27, the individual will be allocated to heart disease group. The obtained threshold is observed to have 86.2% of sensitivity and 13.8% of 1-specificity (false positive rate). This means that the threshold is able to identify the true status of individual in a sensible manner with 86.2% by allowing 13.8% of false positive cases. This features out that the performance of the threshold has to be improved in such a way that the percentage of false positive rate can be minimized.

3.2. Improper MROC (iMROC) curve

The concept of iMROC curve is supported with the help of simulation studies as well as real datasets. The degree of improperness is also measured with the help of crossing point and inflection point and the results are reported along with the figures.

3.2.1. Simulation study

Two sets of multivariate normal random numbers are generated with mean vectors and covariance matrices (Table 1) for various samples sizes 25, 50, 100 and 300 respectively.

Table 1: Mean Vectors and Covariance Matrices of Simulation Studies

	μ_D	μ_H	Σ_D	Σ_H
1	$\begin{pmatrix}0.8606\\1.68\\5.1302\end{pmatrix}$	$\begin{pmatrix}0.8059\\1.5812\\4.7992\end{pmatrix}$	$\begin{pmatrix}0.0084 & 0.0057 & 0.1221\\0.0057 & 0.1183 & 0.0601\\0.1221 & 0.0601 & 2.3087\end{pmatrix}$	$\begin{pmatrix}0.0046 & 0.0001 & 0.0561\\0.0001 & 0.1274 & 0.0037\\0.0561 & 0.0037 & 0.7628\end{pmatrix}$
2	$\begin{pmatrix}0.7305\\1.39\\3.6302\end{pmatrix}$	$\begin{pmatrix}0.7057\\1.2811\\3.5992\end{pmatrix}$	$\begin{pmatrix}0.0084 & 0.0057 & 0.1221\\0.0057 & 0.1183 & 0.0601\\0.1221 & 0.0601 & 2.3087\end{pmatrix}$	$\begin{pmatrix}0.0046 & 0.0001 & 0.0561\\0.0001 & 0.1274 & 0.0037\\0.0561 & 0.0037 & 0.7628\end{pmatrix}$

The accuracy and intrinsic measures along with the linear combinations obtained for the simulated data sets are reported in Table 2.

Table 2: Measures of MROC curve for two sets of simulations at four different sample sizes

Simulation	Samples	c	AUC	TPR	FPR	Linear combination
I	25	2.1928	0.5817	0.5591	0.4408	$2.95 * X_1 + 0.50 * X_2$ - $0.12 * X_3$
	50	4.4107	0.6323	0.5979	0.4020	$6.01 * X_1$ - $0.10 * X_2$ - $0.07 * X_3$
	100	9.8698	0.6754	0.6444	0.3555	$1627 * X_1 + 0.62 * X_2$ - $0.90 * X_3$
	300	18.5639	0.7003	0.6640	0.3359	$29.69 * X_1 + 0.99 * X_2$ - $1.53 * X_3$
II	25	-1.6330	0.5713	0.5543	0.4456	$-2.93 * X_1 + 0.51 * X_2$ - $0.07 * X_3$
	50	2.0133	0.6109	0.5811	0.4188	$3.62 * X_1 + 1.05 * X_2$ - $0.53 * X_3$
	100	4.3673	0.6493	0.6111	0.3888	$5.37 * X_1 + 0.84 * X_2$ - $0.16 * X_3$
	300	2.5782	0.6519	0.5943	0.4056	$3.35 * X_1 + 1.16 * X_2$ - $0.23 * X_3$

Here, an observation made is that as the sample size increases the accuracy (AUC) is also slightly improving even though the expressions for the intrinsic measures FPR, TPR and the accuracy measure AUC are free from the sample size. This means that, there is slight deviation in the accuracy as the curve deviates from cancavity to convexity (Figure 3 and 4). Therefore, iMROC curve is not able to provide the maximum extent of correct classification with less misclassification rate due to the shift in the magnitude of the curve.

3.2.2. Real datasets

In order to demonstrate the iMROC curve, MCA and ILP datasets are used. Further, ILP dataset has been split according to gender of the patients. Of which, ILP male dataset

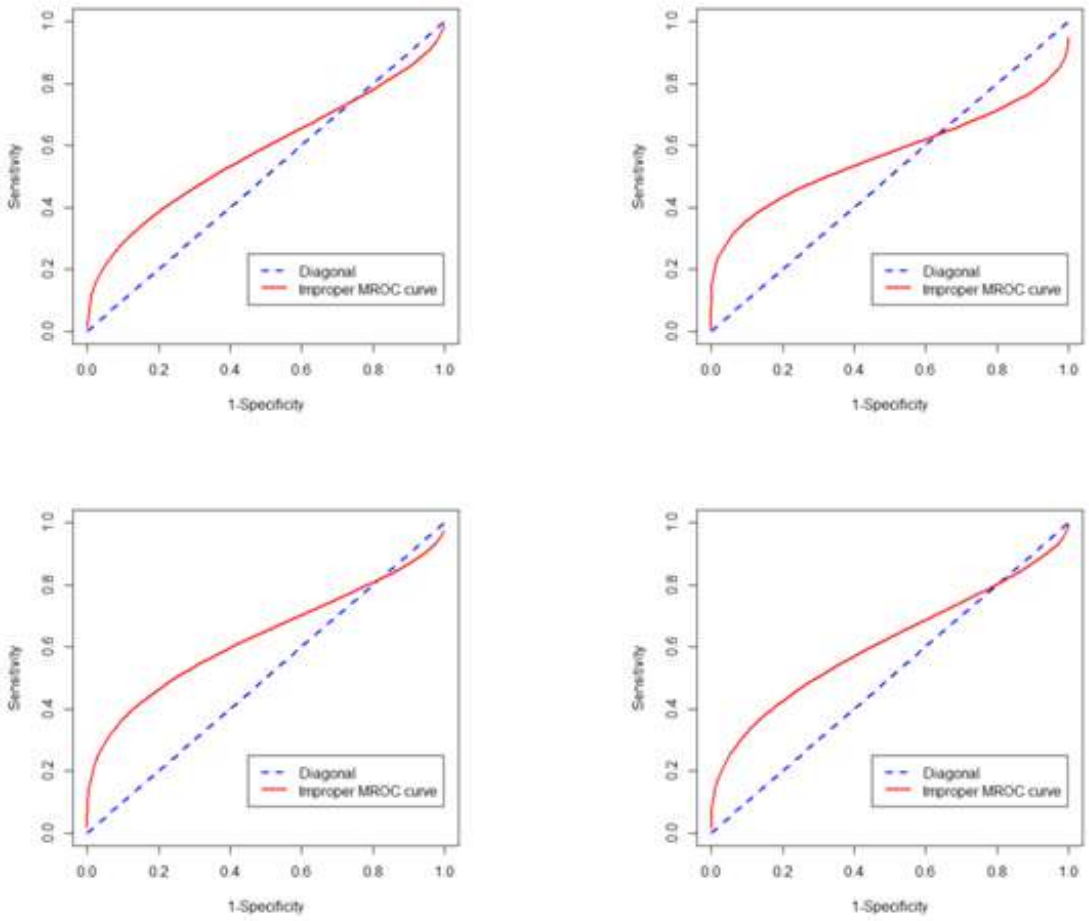


Figure 3: iMROC curves for Simulation datasets at different sample sizes with 25 and 50

has a form of Improper ROC curve and the same dataset has been chosen for demonstration purpose.

ILP Male dataset (Ramana *et al.*, 2012)

The intrinsic measures TPR and FPR, summary measure AUC and optimal cut point are computed using equations (1) to (5). The AUC observed is 0.7495 which provides moderate classification, TPR and FPR are 0.6992 and 0.3008 respectively at the optimal cutpoint c = 1.5372. The best linear combination is given by

$$U_{ILP} = 0.0172 * Age\text{–}0.0556 * TB + 0.3133 * DB + 0.0005 * Alkphos\text{–}0.0104 * sgpt + 0.0074 * sgot - 0.4164 * TP + 0.6726 * ALB\text{–}1.1341 * A.G$$

If the test score is greater than optimal cutoff *i.e.*, 1.5372 the individual is classified as diseased, otherwise healthy. The iMROC curve is drawn and depicted in the Figure (5). From Figure (5), it is clear that the fitted MROC curve crosses the chance line and is

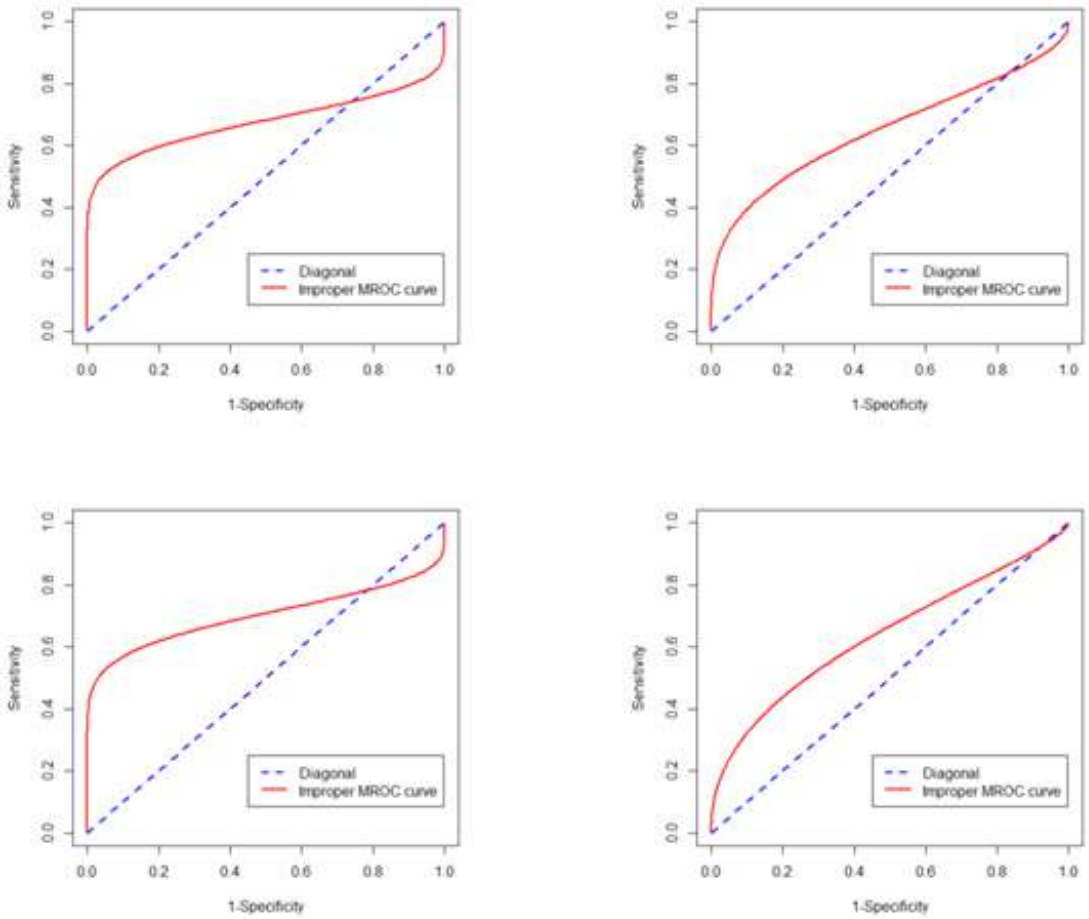


Figure 4: iMROC curves for Simulation datasets at different sample sizes with 100 and 300

moving towards the top right corner of the unit square plot, which generates an improper MROC curve. Using the proposed methodology, the inflection point (t_1) and chance line cross reference points (t_0) are obtained and are highlighted in the Figure (5). MROC curve is concave for $FPR < 0.5221$, but is convex for $FPR > 0.5221$. Due to this improperness, the true accuracy of the classifier cannot be obtained. Further, such contaminated AUC will mislead the interpretation and decision making too.

MCA dataset (Vishnu Vardhan *et al.*. 2015)

The neonatal dataset consists of two procedures: MCA and CPR used to check the blood flow from the womb of the mother to the baby for identifying the growth of the baby. Three indices were measured namely pulsatility index (PI), resistivity index (RI) and Systolic/Diastolic (S/D) ratio in all the procedures. The intrinsic measures TPR and FPR, summary measure AUC and optimal cut point are computed using equation (1) to (5). The AUC observed is 0.6253, which provides moderate classification, TPR and FPR are 0.5968

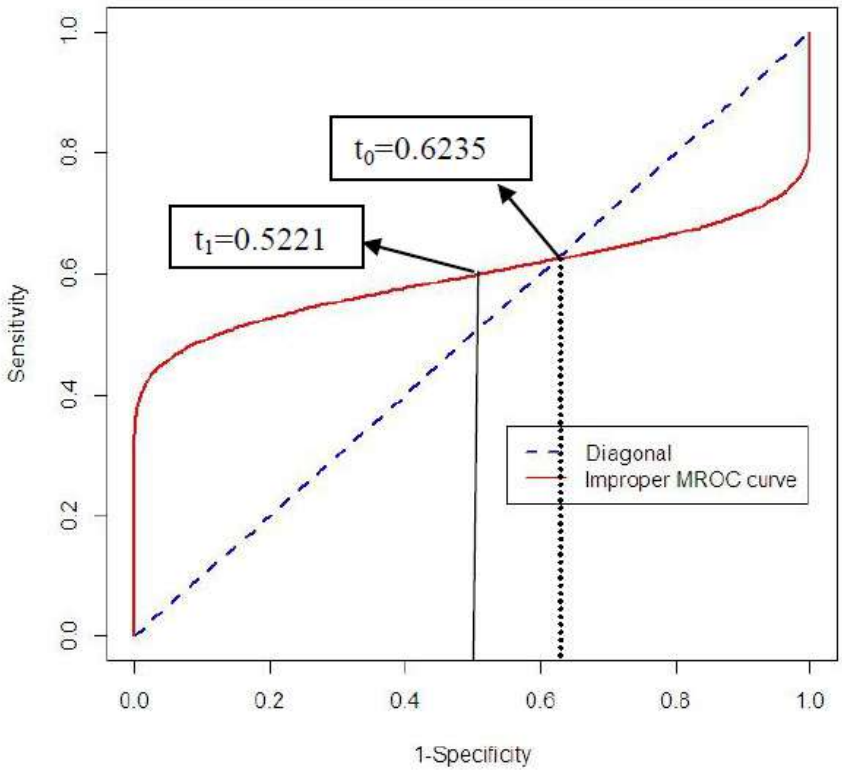


Figure 5: iMROC curve for ILP Male dataset

and 0.4032 at the optimal cutpoint $c = -3.1749$. The best linear combination is given by

$$U_{MCA} = -10.6711 * MCA.RI + 0.0226 * MCA.PI + 1.1733 * MCA.SD$$

The above linear combination can be used for identifying the status of new individual. If the test score is greater than optimal cutoff *i.e.*, -3.1749 the individual is classified as diseased, otherwise healthy.

Further, the MROC curve is drawn and depicted in Figure (6). From Figure (6), it is clear that the fitted MROC curve crosses the chance line and moves towards the top right corner of the unit square plot, which leads to an improper MROC curve. In this illustration also, it is shown that not all ROC curves that gets generated for the classification data is a "Proper" one and before fitting and computing the measures of ROC curve, one has to verify whether the data is satisfying the three properties or not. Doing so, we can overcome the misuse of the technique and misleading conclusions out of it.

4. Conclusion

In this paper, main focus was on establishing the fact that not all ROC curves that are generated through data will be "Proper", *i.e.* that they possesses the monotonic property. So, there is a need to have some mechanism to verify whether an ROC curve so obtained is proper or improper. To address this, crossing point and inflection point are defined, which work on concavity and convexity nature of the ROC curve. To have a better understanding

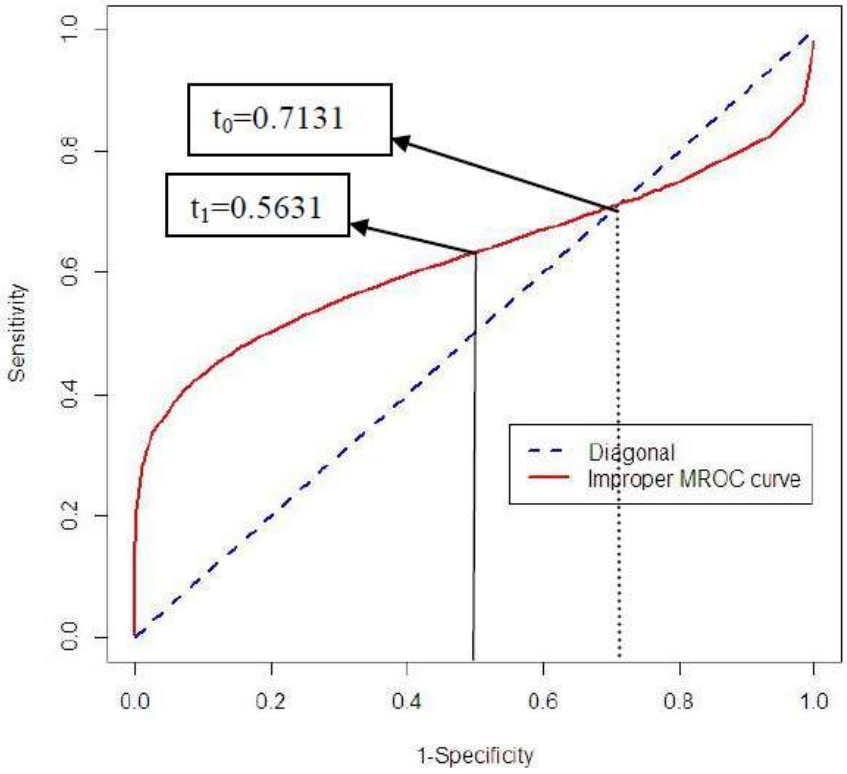


Figure 6: iMROC curve for MCA dataset

of these crossing and inflection points, simulations were carried out for different sample sizes and parameter combinations. Also support of real data sets is also taken. On the whole, the message emerging from this study is that before interpreting the outcomes of ROC curves, it is essential to check whether the curve is proper or improper. If the curve is satisfying the desirable properties then one can proceed for using the classifier for future classification, and if the curve is improper then it is not a better way to use the classifier anymore. So, here it is quite essential to work on a procedure that helps in correcting the ROC curve and making it to have the monotonicity.

References

Balaswamy, S., Vardhan, R. V. and Sameera, G. (2020). Improper multivariate receiver operating characteristic (iMROC) curve. *Statistics Optimization and Information Computing*, **8**, 1-9.

Ramana, B. V., Babu, M. S. P. and Venkateswarlu, N. B (2012). ILPD (Indian Liver Patient Dataset).

Charles E. Metz and Xiaochuan Pan. (1999). Proper Binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology*, **43**, 1 – 33.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York, NY: Academic Press.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley.

Krzanowski, W. J. and Hand, D. J. (2009). ROC curves for continuous data, *Monographs on Statistics and Applied Probability*. New York, NY: CRC Press, Taylor and Francis Group.

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science*, **171**, 1217–1219.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Limited.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford; New York.

Sameera, G., Vishnu Vardhan, R. and Sarma, K. V. S. (2016). Binary classification using multivariate receiver operating characteristic curve for continuous data. *Journal of Biopharmaceutical Statistics*.

Tanner, J. W. P., and Swets, J. A. (1954). A decision-making theory of visual detection, *Psychological Review*, **61**, 401-409.

Vishnu Vardhan, R., Sameera, G., , Chandrasekharan, P. A. and Thulasi B. (2015). Inferential procedures for comparing the accuracy and intrinsic measures of multivariate receiver operating characteristic (MROC) curve, *International Journal of Statistics in Medical Research*, **4**, 87-93.

Special Proceeding of the 22nd Annual Conference of SSCA
Savitribai Phule Pune University, Pune, January 02-04, 2020; pp 73–84

Index Tracking for NIFTY50

Divyashish Choudhary[1] and Rituparna Sen[2]
[1]*Indian Institute of Technology Madras, Chennai 600036, India*
[2]*Indian Statistical Institute, Bengaluru 560059, India*

Received: 28 May 2020; Revised: 25 June 2020; Accepted: 27 June 2020

Abstract

Passive investing has been on the rise in global markets for the last four decades now. It constitutes 45% of U.S. stock based funds up from 25% in just a decade. The prominent reason is the exorbitant costs of active management which are more often than not unjustified by their performance in relation to benchmark index. Passive management mitigates this issue by closely tracking the benchmark index with minimum transaction costs and management fees.

Although there has been expansive research on passive investing in developed markets like the U.S. capital market, the issue has been covered in marginal detail in emerging markets like that of India. Index tracking is at the heart of passive investing and this paper aims to discern the efficacy of our method of construction of an index tracker in the Indian market while focusing on NIFTY50 index. We employed lowess smoothing method and subsequently the partial correlation to create a tracker with subset of the 50 stocks that constitute the benchmark. Further, we quantified the effects that changing rebalancing frequency and number of constituent stocks in the tracker had on the tracking error and transaction costs and suggested optimal trackers.

Key words: Passive Investing; Lowess Smoothing; Transaction Costs; Tracking Error.

AMS Subject Classifications: 60G50, 05C81

1. Introduction

Active investing has been around since the inception of modern money and capital markets. It is build on the philosophy that one can get better returns than market on an average if the organisation or person allocating the capital is skilled enough to exploit the inefficiencies which are assumed to be present in the market.

Passive investing on the other hand assumes that markets are efficient and the best returns that one can get are by investing capital in the portfolio which mimics market composition. This philosophy has gained prominence mainly because majority of active portfolio managers have a track record that trails market returns when taken over a sufficiently large period like 10-20 yrs. This empirical evidence coupled with the fact that passive investing

Correponding Authors: Divyashish Choudhary, Rituparna Sen
Email: divyashishchoudhary@gmail.com and rsen@isibang.ac.in

is simple and has lower transaction charges has led to capital's exodus from active portfolio management into passive portfolio management. Although passive investing can have many forms, index investing is the most common. It entails building a portfolio which mimics a major benchmark index in the market and hence gives an investor returns in line with the broader market which is the essence of passive investment. This strategy has minimum buy/sell operations and hence reduces market friction which reduces the returns of an investor over the long term.

There are various methods to track an index, simplest being full replication of an index. This method gives minimum tracking error but incurs explicit and implicit trading costs which makes it the costliest strategy. Second method can be optimization based on mean/variance analysis (Roll (1992)). This is a solution which holds fewer stocks than full replication but still doesn't compromise much on the tracking error. However, studies (Focardi and Fabozzi (2004)) note that the noise can dominate the correlations of stocks which renders variance/covariance unreliable. This stems from the fact that high dimensional covariance matrices cannot be estimated consistently. Nakayama and Yokouchi (2018) propose a method that picks constituent stocks for a tracker based on the similarities they have to benchmark index. In turn, these similarities are arrived at by calculating distances of time series trends which are derived from decomposing original series using lowess smoothing.This does not yield appreciable results due to two factors. The first being the residual time series left after lowess decomposition and second being the high correlation between the stocks that constituted the tracker. To improve tracking performance, they propose a similarity-balanced approach in which different groups are formed based on ranks generated from the similarity approach and representative stocks are taken from each group.

We have built upon this method by utilizing iterative confounding variable approach to overcome the correlation issue instead of similarity-balanced approach. This method works in this case because the number of data points far exceed the number of constituent stocks and hence the variance-covariance matrix has eigenvalues different from those stipulated by random matrix theory distribution (Focardi and Fabozzi (2004)). Unfortunately, we were not able to find any study of similar kind done for the Indian market which left us with no reference or comparison source for the results we got.

This paper proceeds as follows. Section 2 describes the data used for the study. Section 3 expounds the methodology starting with lowess smoothening continuing on to the use of partial correlation and how the transaction costs were calculated. Section 4 presents the results we obtained for our index-tracking portfolio , the transaction costs incurred and the best optimal combination of rebalancing window and number of constituent stocks for our tracker. Section 6 concludes and presents scope for further research.

2. Data

Daily closing price adjusted for dividend, bonus and splits has been taken with starting date at 01/06/2008 and ending date at 01/06/2018 for the 50 constituent stocks and the NIFTY50 index. The data was retrieved from yahoo finance website. Below is the snippet of the raw data and processed data in which we combined the adjusted close data of all the stocks and discarded other data columns.

Date	Open	High	Low	Close	Adj Close	Volume
02/06/2008	159.100006	165.399994	152.009995	153.160004	135.345947	4245850
03/06/2008	148.199997	153.979996	146.669998	147.880005	130.680008	2664250
04/06/2008	148.399994	151.139999	138.449997	140.440002	124.105370	1915440
05/06/2008	140.199997	141.779999	127.019997	131.509995	116.214043	3579815
06/06/2008	133.979996	136.250000	127.000000	127.919998	113.041603	2841970
09/06/2008	124.000000	125.199997	118.260002	122.519997	108.269669	2306670
10/06/2008	119.839996	123.769997	116.000000	117.849998	104.142845	2586135
11/06/2008	119.599998	126.279999	119.040001	121.419998	107.297615	3841800
12/06/2008	119.580002	120.879997	114.000000	118.699997	104.893974	2530485

Date	Price_1	Price_2	Price_3	Price_4	Price_5	Price_7	Price_8
2008-06-02	153.160004	91.800003	126.129997	29.375000	149.869995	25.003599	58.099998
2008-06-03	147.880005	92.400002	124.375000	28.900000	151.479996	23.760000	58.633301
2008-06-04	140.440002	87.599998	125.000000	27.910000	147.360001	22.754400	53.716702
2008-06-05	131.509995	85.800003	125.665001	28.665001	154.179993	22.322100	50.308300
2008-06-06	127.919998	85.400002	126.080002	28.230000	152.110001	21.510799	50.158298
2008-06-09	122.519997	81.900002	125.605003	27.165001	140.619995	18.610600	46.474998
2008-06-10	117.849998	82.349998	125.665001	28.090000	136.300003	18.596100	47.433300
2008-06-11	121.419998	88.250000	128.785004	30.094999	141.160004	17.337900	45.541698
2008-06-12	118.699997	85.900002	125.904999	31.410000	142.869995	16.385700	44.891701

Figure 1: Left: raw data for a stock ; Right: pre-processed data for all stocks

3. Methodology

3.1. Lowess smoothing of the raw data

We have taken the data of all the stocks and the index in the time frame marked out in section 2. We have applied *Lowess* smoothing as proposed by Cleveland (1979); used by Shibata and Miura (1997) as well as Nakayama and Yokouchi (2018). *Lowess* smoothes values by applying locally weighted linear regression with weights determined by the distance of data points from the point selected for smoothing. It is a non-parametric method of smoothing the data which assumes no prior assumptions about economic cycles or specific models.

We have decomposed the price time series data into components of long term trends and short term trends. The long term trend was obtained by smoothing the normalised price time series of all the constituents and the index. Short term trends were then obtained by smoothing residuals from the long term smoothing. The residual after second smoothing was weakly-stationary. We have kept one year as the time frame for long term trends and 1 month as the time frame for short term trends. Figure 1 illustrates such break-up for the NIFTY50 index.

3.2. Tracker construction using partial correlation

Nakayama and Yokouchi (2018) used similarity between constituent stocks and the index to rank stocks; out of which the top few were selected to construct a tracker. This was unlike the previous studies which used similarity between the constituent stocks to form clusters (Focardi and Fabozzi (2004), Dose and Cincotti (2005)) or which used integration (Thomaidis (2013), Papantonis (2016)) for index tracking. However, they were not able to beat the results obtained by clustering techniques through the similarity approach. Hence, they developed similarity-balanced approach to tackle the issue of correlation between the constituent stocks and now were able to provide better returns than the similarity approach.

In our paper, we have incorporated the operational features of similarity-balanced approach by computing inter-day percentage change data for both long term and short term trends and then using partial correlation between the percentage change data of index and the constituent stocks to rank the stocks and construct an index based on this ranking. We implemented this procedure separately between long term trends of stocks & the index and

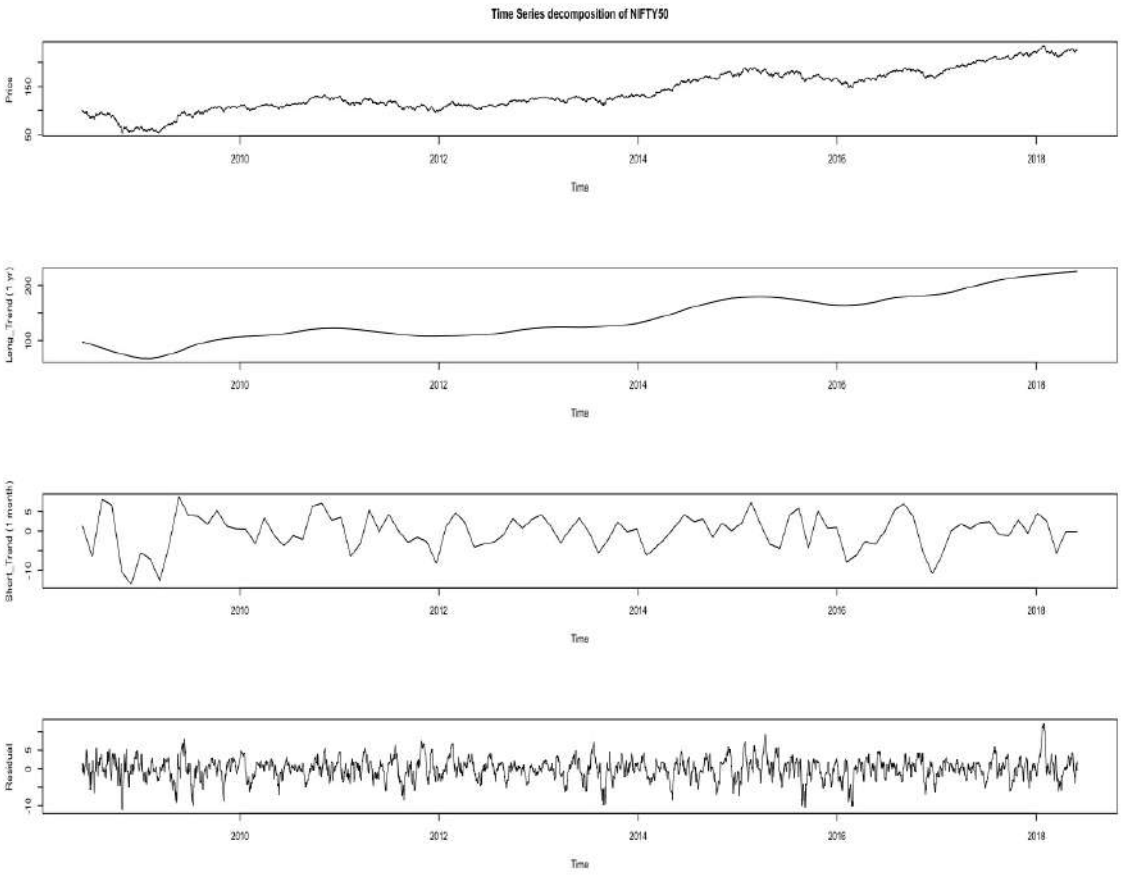


Figure 2: Lowess smoothing and subsequent decomposition of NIFTY50

between short term trends of stocks & the index. Two different rankings and subsequently two different trackers were obtained using this methodology. This technique allowed us to eliminate confounding bias but in a statistical fashion rather than heuristic fashion.

We have assigned equal weight to all the constituents of a tracker. On rebalancing, the partial correlation is again calculated in a similar manner as stated before and some constituents are replaced but the total capital at that point is again redistributed equally among all the new constituents and old constituents which still remain. Finally, we have compared and suggested winning tracker on the basis of reduction of tracking error and transaction costs. Tracking error in our context is defined as the difference in percentage change between benchmark index and the tracker after every rebalancing window. Transaction cost's definition and calculation will be detailed in the next sub-section. We have used three parameters to control the values of costs and tracking error - number of stocks in the tracker, rebalancing frequency and composition of hybrid long-short tracker.

3.3. Transaction costs

Transaction costs play the biggest role in determining the real returns that an index can provide. We aim to minimize the transaction costs which we will assume to have both

a proportional and fixed component for every trade (Jha and Srivastava (2013), Kellerer et.al (2000), Nakayama and Yokouchi (2018)). Let p denote the per rupee transaction cost, f denote the fixed cost per transaction, n denote the number of stocks in the tracker, R_i denote the return on stock i (value at end of rebalancing period/value at the beginning of it), MAD be mean absolute deviation, turnover be the ratio of number of stocks replaced at rebalancing point to the total number of stocks in the tracker, j be the set of stocks that will be replaced at rebalancing point and l be the total number of these replaced stocks. Then the proportional and fixed costs will be given as follows with detailed derivation attached in the appendix.

$$Cost_{proportional} = pMAD(R_i) + 2p(turnover)[\tfrac{1}{l}\textstyle\sum_j min(R_j, \bar{R})]$$

$$Cost_{fixed} = 2fl + f(n - l)$$

4. Results

4.1. Number of stocks in tracker

We have taken 2 years as the training period for construction of all the trackers in this and the ensuing sub-sections. Tables 1 and 2 show the results obtained when we varied the number of stocks in the 2 different categories of trackers - obtained by using short trends and long trends respectively with Quarterly rebalancing window.

Table 1: Effect of no. of stocks on short-trend tracker

No. of Stocks	Average Turnover	Average Rebalancing Cost	Iter. $\leq 1\%$ Tracking error	Iter. $\geq 4\%$ Tracking error	OU Ratio
5	0.29	0.25%	31%	38%	1.67
10	0.28	0.37%	28%	28%	1.46
15	0.26	0.49%	19%	34%	3
20	0.22	0.58%	22%	28%	3
25	0.18	0.67%	16%	28%	3
30	0.14	0.75%	16%	31%	2.56
35	0.09	0.82%	31%	09%	0.6
40	0.05	0.87%	16%	25%	3
45	0.01	0.93%	16%	19%	2.56

Table 2: Effect of no. of stocks on long-trend tracker

No. of Stocks	Average Turnover	Average Rebalancing Cost	Iter. $\leq 1\%$ Tracking error	Iter. $\geq 4\%$ Tracking error	OU Ratio
5	0.61	0.39%	16%	28%	1.29
10	0.58	0.54%	34%	22%	1.46
15	0.48	0.63%	19%	16%	1.29
20	0.38	0.70%	34%	16%	1.91
25	0.29	0.77%	22%	09%	2.56
30	0.25	0.85%	31%	06%	3
35	0.18	0.90%	31%	09%	2.2
40	0.11	0.94%	31%	09%	1.91
45	0.02	0.94%	19%	19%	2.56

Definitions for the columns in tables 1 and 2 : −

No. of stocks - Number of stocks in the tracker.
Average Turnover - Average of the proportion of stocks that were replaced by another set of stocks in the tracker.
Average Rebalancing Cost - Average cost in percentage terms to rebalance the tracker. The calculation includes both the proportional and fixed costs which are calculated based on the formulae in section 3. Hence, low value value of this parameter is favourable.
Iter. $\leq 1\%$ Tracking error - Proportion of tracking iteration in percentage terms which had a tracking error of 1% or less including the cases of both excess and under returns as compared to the index. Hence, high value of this parameter is favorable.
Iter. $\geq 4\%$ Tracking error - Proportion of tracking iteration in percentage terms which had a tracking error of 4% or more including the cases of both excess and under returns as compared to the index. Hence, low value of this parameter is favourable.
OU Ratio - This parameter gives the ratio of iterations in which tracker outperformed the benchmark to the iterations in which tracker underperformed the benchmark. Hence, higher OU value the better.

Results show that the cost of rebalancing goes up with increase in the number of stocks in the tracker owing to increasing fixed charges with increasing number of stocks in the tracker which have to be replaced and others which have to be re-balanced. Further, we note that tracking error does not follow a strictly increasing or decreasing trend with the number of stock. Finally, the tracking error favours the number of stocks in the 20 -30 range. We can assert this as the tracking error efficiency saturates beyond 30 stocks (also evident in figure 3) barring the case of 35 stocks in short tracker (but this tracker suffers from a poor OU ratio). Further, for less than 20 stocks, the tracking error is high which is not duly compensated with low transaction costs. Therefore, we will be considering trackers with **20, 25 or 30** stocks in them for constructing the index.

4.2. Rebalancing frequency

Table 3 and 4 show the results obtained when we varied the rebalancing frequency for short and long trend trackers taking 20, 25 and 30 stocks in them.

Table 3: Effect of rebalancing frequency on short-trend tracker

No. of Stocks	Rebal. Frequency.	Annual Rebal. Cost	Iter. $\leq 1\%$ Tracking error	Iter. $\geq 4\%$ Tracking error	OU Ratio
20	Monthly	6%	47%	0%	1.72
	Bi-Monthly	3.24%	22%	16%	3.08
	Quarterly	2.32%	22%	28%	3
	Half-Yearly	1.36%	06%	38%	7
	Yearly	0.79%	12%	75%	7
25	Monthly	6.96%	51%	0%	2.50
	Bi-Monthly	3.78%	18%	12%	2.77
	Quarterly	2.68%	16%	28%	3
	Half-Yearly	1.52%	06%	44%	4.33
	Yearly	0.85%	12%	88%	7
30	Monthly	8.04%	56%	01%	2.06
	Bi-Monthly	4.32%	31%	12%	2.27
	Quarterly	3.00%	16%	31%	2.56
	Half-Yearly	1.66%	12%	05%	4.33
	Yearly	0.89%	12%	75%	7

Table 4: Effect of rebalancing frequency on long-trend tracker

No. of Stocks	Rebal. Frequency.	Annual Rebal. Cost	Iter. $\leq 1\%$ Tracking error	Iter. $\geq 4\%$ Tracking error	OU Ratio
20	Monthly	7.68%	59%	01%	1.13
	Bi-Monthly	4.08%	41%	06%	2.5
	Quarterly	2.80%	34%	16%	1.91
	Half-Yearly	1.48%	25%	38%	3
	Yearly	0.79%	12%	38%	7
25	Monthly	8.64%	58%	01%	1.09
	Bi-Monthly	4.56%	37%	04%	2.77
	Quarterly	3.08%	22%	09%	2.56
	Half-Yearly	1.6%	31%	44%	1.67
	Yearly	0.86%	25%	50%	7
30	Monthly	9.72%	66%	01%	1.39
	Bi-Monthly	5.1%	39%	04%	2.50
	Quarterly	3.40%	31%	06%	3
	Half-Yearly	1.94%	31%	38%	2.2
	Yearly	0.94%	12%	50%	7

We observe that although monthly rebalancing gives us extremely favorable results in terms of tracking error, it does incur exorbitant annual transaction costs which makes it undesirable and hence this frequency cannot be used. Moreover, half-yearly and yearly rebalancing are on the other end of the spectrum where they provide lower transaction costs but their tracking error is large and hence these frequencies too cannot be used.

Bi-monthly and Quarterly rebalancing acceptably balance between costs and tracking error. Quarterly rebalancing offers an edge in costs while bi-monthly rebalancing gives us better results in tracking error.

4.3. Proportion of long/short tracker in composite tracker

Table 5 shows the results obtained for different combinations of number of stocks taken to construct long-short composite tracker for 20, 25 and 30 stocks in total with bi-monthly and quarterly rebalancing window. We observe that as the contribution of short tracker increases in the composite tracker, we get lower transaction charges but tracking error increases at the same time. The vice-versa is true for the contribution of long tracker in the composite tracker. Figure 3 depicts the relationship between transaction costs and the tracking error based on table 5. We infer that there is again no strict mono-directional relationship between the two which leads to multiple trackers being optimal for our purpose. Figure 4 depicts the performance of four trackers which are optimal trackers marked in boldface in table 5.s

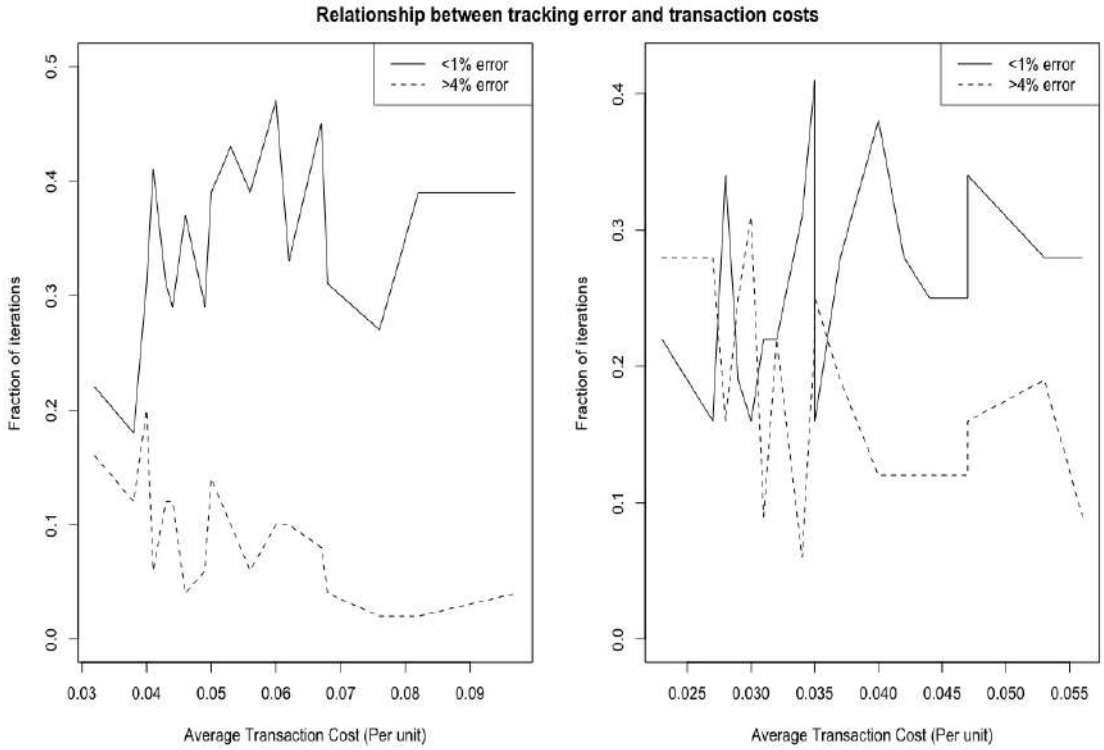


Figure 3: Left: bi-monthly rebalancing; Right: quarterly rebalancing

Table 5: Effect of tracker composition on long-short tracker with bi-monthly rebalancing

Rebal. Freq.	No. of Stocks	Composition	Annual Rebal. Cost	Iter. $\leq 1\%$ Tracking error	Iter. $\geq 4\%$ Tracking error	OU Ratio
Bi-Monthly	20	s = 0, l = 20	4.08%	41%	06%	2.5
		s = 5, l = 15	6.78%	31%	04%	1.58
		s = 10, l = 10	4.98%	39%	14%	1.45
		s = 15, l = 5	4.02%	31%	20%	1.58
		s = 20, l = 0	3.24%	22%	16%	3.08
	25	s = 0, l = 25	4.56%	37%	04%	2.77
		s = 5, l = 20	7.56%	27%	02%	2.27
		s = 10, l = 15	**6.00%**	**47%**	**10%**	**1.88**
		s = 15, l = 10	**5.28%**	**43%**	**10%**	**1.72**
		s = 20, l = 5	4.44%	29%	12%	2.06
		s = 25, l = 0	3.78%	18%	12%	2.77
	30	s = 0, l = 30	5.10%	39%	04%	2.5
		s = 5, l = 25	8.22%	39%	02%	2.5
		s = 10, l = 20	6.72%	45%	08%	3.45
		s = 15, l = 15	6.24%	33%	10%	1.88
		s = 20, l = 10	5.64%	39%	06%	2.27
		s = 25, l = 5	4.92%	29%	06%	3.08
		s = 30, l = 0	4.32%	31%	12%	2.27
Quarterly	20	s = 0, l = 20	2.80%	34%	16%	1.91
		s = 5, l = 15	4.72%	25%	12%	1.29
		s = 10, l = 10	**3.52%**	**41%**	**22%**	**1.91**
		s = 15, l = 5	2.88%	19%	25%	3
		s = 20, l = 0	2.32%	22%	28%	3
	25	s = 0, l = 25	3.08%	22%	09%	2.56
		s = 5, l = 20	5.28%	28%	19%	2.56
		s = 10, l = 15	**4.20%**	**28%**	**12%**	**2.2**
		s = 15, l = 10	3.72%	28%	19%	3.57
		s = 20, l = 5	3.20%	22%	22%	3
		s = 25, l = 0	2.68%	16%	28%	3
	30	s = 0, l = 30	3.40%	31%	06%	3
		s = 5, l = 25	5.56%	28%	09%	3.57
		s = 10, l = 20	4.72%	34%	16%	4.33
		s = 15, l = 15	4.36%	25%	12%	3
		s = 20, l = 10	4.00%	38%	12%	4.33
		s = 25, l = 5	3.48%	16%	25%	3.57
		s = 30, l = 0	3.00%	16%	31%	2.56

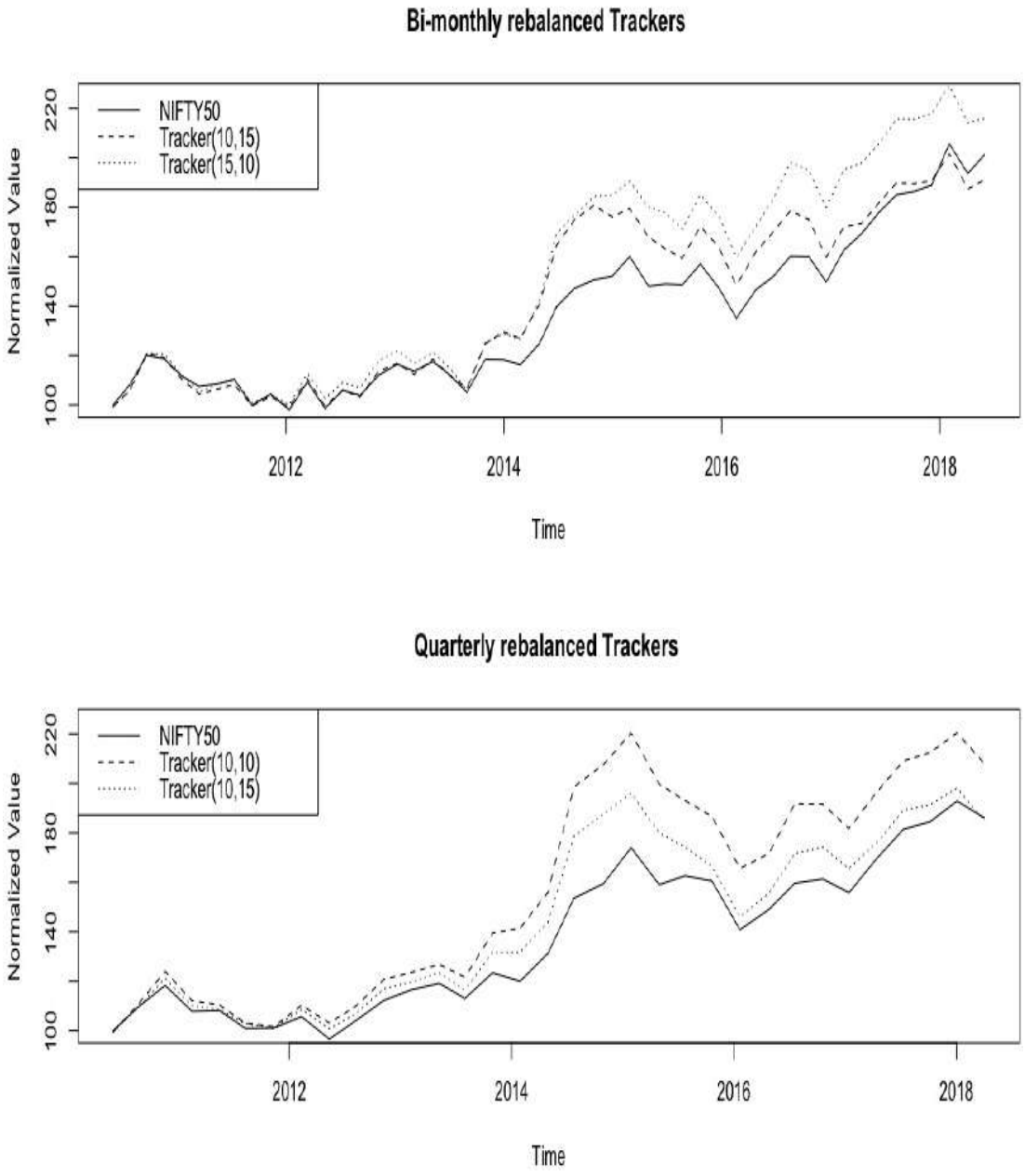


Figure 4: Performance of optimal trackers with bi-monthly and quarterly rebalancing

5. Conclusion

In this study, we proposed a method to track NIFTY50 which is one of the most representative benchmark of Indian capital markets covering over 65% of market capitalisation. We observed that we have multiple parameters at our disposal to enhance the performance of tracker. We also characterised the trade-offs involved in the process.

Further work can involve taking other benchmark indices of Indian capital markets *viz.* BSE Sensex *etc.* and implementing the aforementioned methodology on them to find the bandwidth of this procedure. Further work can also be done on dynamic weight adjustment of stocks in the tracker based on their performance in the previous iterations.

References

Nakayama, J. and Yokouchi, D. (2018). Applying time series decomposition to construct index-tracking portfolio. *Asia-Pacific Financial Markets (2018),* **25**, 341–352.

Roll, R. (1992). A mean/variance analysis of tracking error. *Journal of Portfolio Management, Summer,* **18**, 13–22.

Focardi, S. M. and Fabozzi, F. J. (2004). A methodology for index tracking based on time-series clustering. *Quantitative Finance,* **4**(**4**), 417–425.

Dose, C. and Cincotti, S. (2005). Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A,* **355**, 145–151.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* **74**(**368**), 829–836.

Shibata, R. and Miura, R. (1997). Decomposition of Japanese yen interest rate data through local regression. *Financial Engineering and the Japanese Markets,* **4**, 125–146.

Thomaidis, N. S. (2013). On the application of cointegration analysis in enhanced indexing. *Applied Economic Letters,* **20**, 391–396.

Papantonis, I. (2016). Cointegration-based trading: Evidence on index tracking and market-neutral strategies. *Managerial Finance,* **42**, 449–471.

Jha, M. and Srivastava, N. (2013). Portfolio rebalancing model using fuzzy optimization. *International Journal of Scientific Engineering and Research (IJSER),* **1**(**4**), 2347-3878.

Kellerer, H., Mansini, R. and Speranza, M. G. (2000). Selecting Portfolios with fixed costs and minimum transaction lots. *Annals of Operations Research,* **99**, 287–304.

NIFTY 50 Index Methodology. Retrieved on 20 March 2020 from www1.nseindia.com/content/indices/Method_Nifty_50.pdf.

APPENDIX

1. Calculation of proportional component of transaction cost

In calculation of the transaction charges per unit of capital. Assume the following :-
n - number of stocks in tracker.
l - number of stocks to be dropped/added after rebalancing window
$\frac{1}{n}$ - Initial investment in each stock.
X_i - Final value of the investment in stock i.
R_i - Return on stock i ($\frac{X_i}{\frac{1}{n}}$).
$\sum_{i=1}^{n} X_i$ - Total value of investment at the end of rebalancing window.
p - proportional rebalancing cost per unit MAD - Mean Absolute Deviation

Now, the amount that should be present in each stock after the rebalancing period is :-
$\frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$
Hence, transaction cost will be given as:-
$Cost = p \sum_{i=1}^{n} |X_i - \bar{X}| = pnMAD(X_i)$

But we will have to reduce from this value the contribution of those stocks which will be replaced after the transaction period. Therefore,

$Cost = pnMAD(X_i) - \sum_j |X_j - \bar{X}|$, where j denotes the set of stocks to be dropped

Further, we add the transaction cost of dropping those stocks and adding new stocks to the portfolio.

Cost of dropping $= p \sum_j X_j$

Cost of adding $= p \sum_1^l \bar{X}$

Hence, total proportional cost is given by :-

$Cost = pnMAD(X_i) - \sum_j |X_j - \bar{X}| + p \sum_j (X_j + \bar{X})$
$= pnMAD(X_i) + 2p \sum_j min(X_j, \bar{X})$ by identity $(a + b - |a - b| = 2min(a, b))$

$= pMAD(R_i) + 2p(\frac{l}{n})(\frac{1}{l} \sum_j min(R_j, \bar{R}))$

$= pMAD(R_i) + 2p(turnover)(\frac{1}{l} \sum_j min(R_j, \bar{R}))$

Special Proceeding of the 22 nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02–04, 2020; pp 85–101

Statistical Modelling and Calibration of a Dynamic Computer Simulator

Pritam Ranjan
Operations Management and Quantitative Techniques Area
Indian Institute of Management Indore, M.P., India 453556

Received: 05 July 2020; Revised: 13 July 2020; Accepted: 15 July 2020

Abstract

Over the last two decades, the advancement in computing power has led to a significant rise in the usage of computer simulators (or simulation models) for real-life applications where data collection via physical observation or experimentation is infeasible or too expensive. For complex real-life processes, realistic accurate simulation models are also time-consuming to run, and subsequently, statistical surrogate models are often used to emulate the simulator outputs. In this article, we will discuss a surrogate model for emulating dynamic simulator outputs (*i.e.*, a simulator which yields time series response).

For a scalar-valued deterministic simulator, Gaussian process model was first introduced as a statistical surrogate, which is till date the most popular emulator in the computer experiment literature. For a dynamic simulator, this article presents a singular value decomposition (SVD) based Gaussian process (GP) model for the emulation. We will also present an efficient approach for estimating the inverse solution from a dynamic computer model, where the objective is to find the optimal set of inputs that produces outputs matching the target response as close as possible. This is also referred to as the calibration of the computer simulation model. The performance of this innovative approach will be demonstrated via several simulated examples and a real-life application.

Key words: Computer experiments; Expected improvement; Gaussian process regression; Inverse problem; Saddlepoint approximation.

1. Introduction

When the physical processes are too expensive to observe and experiment with, mathematical models are often used to mimic the underlying phenomena. These mathematical models are coded/implemented via some software programming language like C, C++, Java, R, Python, *etc.* The experimentation with such computer codes are referred to as computer experiments. There is a plethora of real-life examples of computer simulators that range from pharmaceutical industry, aviation sector, cosmology, manufacturing, agriculture, re-

Correponding Author: Pritam Ranjan
Email: pritamr@iimidr.ac.in

newable energy, *etc.* We start with briefly presenting a few real-life applications of computer simulators.

Application 1. The data collection for the impact assessment of a car crash due to different types of collision on passengers sitting in the front and back row is really expensive. Though crash testing is a must in some countries, it is not done everywhere and certainly not thoroughly. A popular cheaper alternative is to use simulation models for detailed experimentation. Figure 1 depicts a car-crash experimentation process.

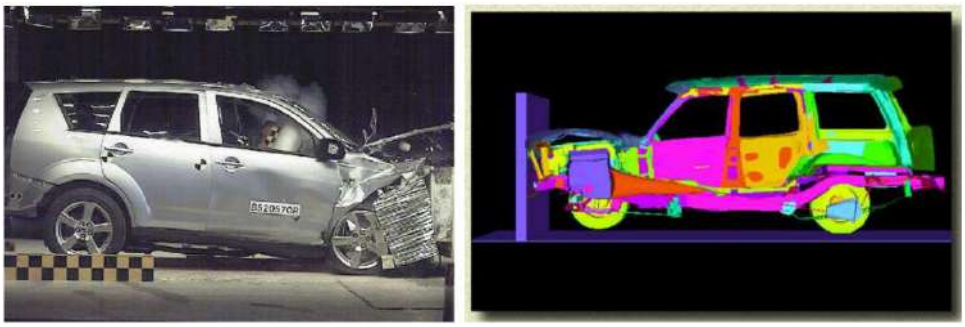


Figure 1: Impact assessment of a car-crash via both physical and computer experiments.

Application 2. Harnessing tidal energy by putting in-stream turbines is an exciting and yet a challenging problem for a variety of researchers around the globe. The Bay of Fundy, located between New Brunswick and Nova Scotia, Canada is world famous for its high tides. In some regions of the Bay of Fundy, the difference in the water level between high tide and low tide can be as much as 17 meters (see Figure 2). The incredible energy in these tides has significant potential for extracting tidal power (Karsten et al., 2008).

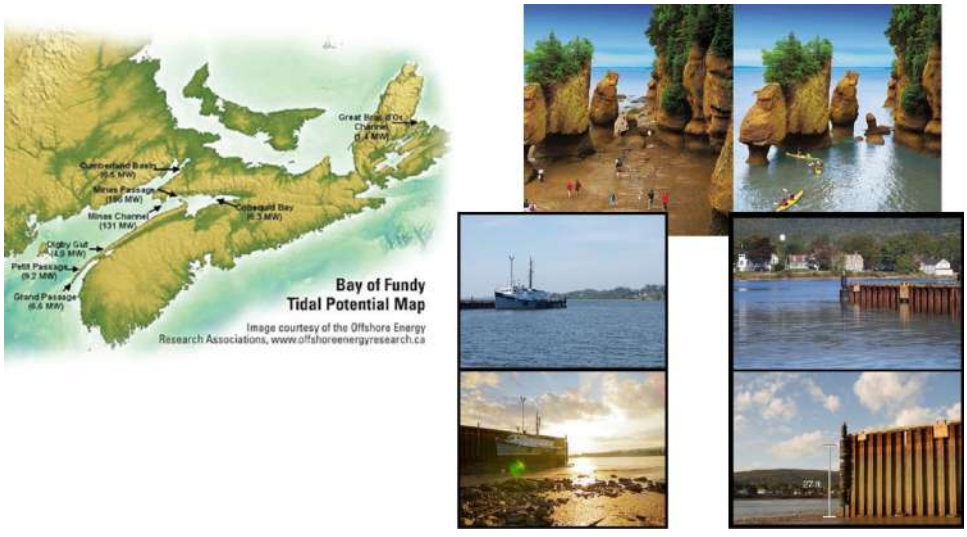


Figure 2: The map of Nova Scotia focussing on the Bay of Fundy. Hopewell Rocks and a port near Wolfville with high tide and low tide.

Preliminary analyses revealed that one can safely extract gigawatts of power by putting

a host of in-stream tidal turbines in the Minas Passage - a narrow passage in the Bay of Fundy where the tidal currents are strongest (see Figure 3). One of the interesting research questions is where exactly should the turbines be placed. Identification of the optimal locations cannot be done via physical experimentation, and we must rely of good computer model based experiments. Currently, researchers have been studying different aspects of the problem using versions of the Finite-Volume Coastal Ocean Model (FVCOM). Details of FVCOM can be found in Chen et al. (2006).

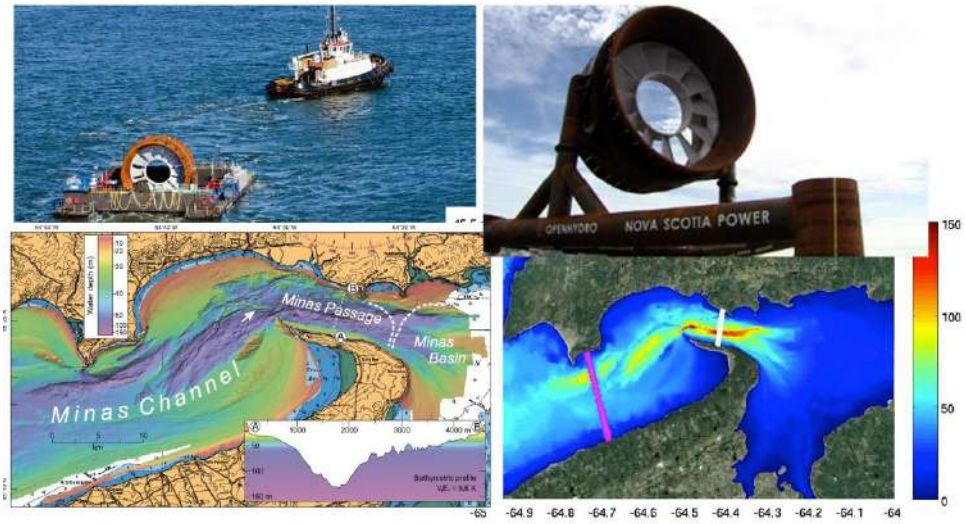


Figure 3: Turbine placement in the Minas Passage.

Application 3. European red mites (ERM) infests on apple leaves and diminish the quality of crop, which inflicts heavy financial loss in the apple industry (see Figure 4). Therefore, the monitoring and subsequent intervention of ERM population dynamics is of vital importance for apple orchards management.

Figure 4: European Red Mites (ERM) infesting on apple leaves.

It is worth noting that data collection for different stages of ERM population, eggs, juveniles and adults, would require counting the number of mites, on randomly sampled orchards, trees, branches and leaves, using a magnifying glass over a period of time. This is undoubtedly expensive (see black solid curve in Figure 5 for a field data). Alternatively, one

can use a computer simulator to mimic the population growth of the three stages. The two-delay blowfly (TDB) model (Teismann et al. (2009)) simulates ERM population dynamics under predator-prey interactions via numerically solving the Nicholson's blowfly differential equation (Gurney et al. (1980)). One of the prime objectives here includes the calibration of this simulator to mimic the reality as much as possible. The TDB model takes eleven input variables (*e.g.*, death rates for different stages, fecundity, hatching time, survival rates, and so on) and returns the time series (at 28 time points) of ERM population evolution at three stages, *i.e.*, eggs, juveniles and adults (see Ranjan et al. (2016) for details). Figure 5 shows the model output at five randomly chosen input points.

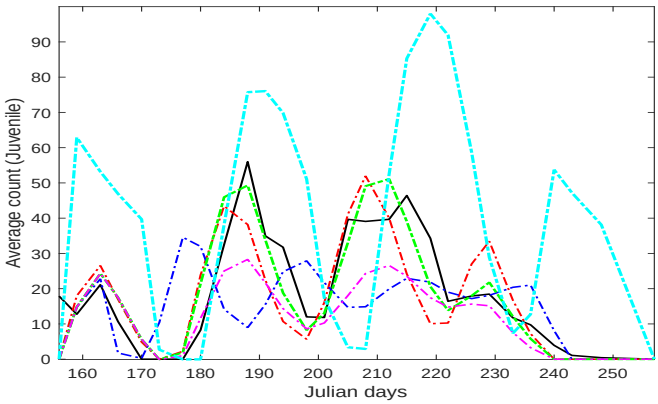


Figure 5: TDB outputs for a few randomly chosen inputs (coloured curved), and the field data (in solid black).

In this article, we first review the popular statistical surrogate called the GP model for emulating the simulator response (Sacks et al., 1989). Subsequently, we discuss the singular value decomposition (SVD) based GP surrogate for emulating the dynamic computer simulator outputs (Higdon et al., 2008; Zhang et al., 2018). Then we focus on the inverse problem for dynamic computer models - which was the key objective of our motivating application from the apple farming industry. This part of the methodology requires a quick recap of the popular expected improvement (EI) based sequential design approach for calibrating the deterministic, expensive to evaluate, scalar-valued simulator outputs (Jones et al., 1998; Ranjan et al., 2008; Bingham et al., 2014), and then present a generalization for dynamic simulators (Zhang et al., 2019). Finally, we compare the performance of this innovative method with the naive approach for several test functions and the TDB model.

2. Statistical Surrogates

The choice of statistical metamodel varies with the characteristics of the computer simulator. For instance, if the output is scalar/vector/functional, the process is stationary/non-stationary, deterministic/stochastic, input space is convex/non-convex, inputs are continuous/discrete, *etc.* We first present the most popular surrogate model called GP model for scalar-valued deterministic simulator, and then outline an extension of this GP model for dynamic simulators.

2.1. Scalar-valued simulator

Let the training data consist of d-dimensional input and 1-dimensional output, denoted by $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ and $y_i = y(x_i)$, respectively. Then, the GP model is written as

$$y_i = \mu + z(x_i), \quad i = 1, 2, \ldots, n, \tag{1}$$

where μ is the overall mean, and $\{z(x), x \in [0,1]^d\} \sim GP(0, \sigma_z^2 R(,))$ with $E(z(x)) = 0$, $Var(z(x)) = \sigma_z^2$, and $Cov(z(x_i), z(x_j)) = \sigma_z^2 R(x_i, x_j)$ where $R(,)$ is a positive definite correlation function. That is, $Y = (y_1, y_2, \ldots, y_n)^T \sim MVN(\mu 1_n, \sigma_z^2 R_n)$, where 1_n is an $n \times 1$ vector of all 1's, and R_n is an $n \times n$ correlation matrix with (i,j)-th element given by $R(x_i, x_j)$ (see Sacks et al. (1989); Santner et al. (2003); Rasmussen and Williams (2006) for more details).

The model described in (1) is typically fitted by either maximizing the likelihood or via Bayesian algorithms like Markov chain Monte Carlo (MCMC). As a result, the predicted response $\hat{y}(x_0)$ is the same as the conditional mean:

$$E(y(x_0)|Y) = \mu + r(x_0)^T R_n^{-1}(Y - 1_n \mu), \tag{2}$$

and the associate prediction uncertainty estimate (denoted by $s^2(x_0)$) can be quantified by the conditional variance:

$$Var(y(x_0)|Y) = \sigma_z^2(1 - r^T(x_0) R_n^{-1} r(x_0)). \tag{3}$$

The most crucial component of such a GP model is the spatial correlation structure, $R(,)$, which dictates the 'smoothness' of the interpolator that passes through the observations. By definition, any positive definite correlation structure would suffice, but the most popular choice is the power-exponential correlation family given by

$$R(x_i, x_j) = \prod_{k=1}^{d} \exp\{-\theta_k |x_{ik} - x_{jk}|^{p_k}\}, \tag{4}$$

where θ_k and p_k controls the wiggliness of the surrogate in the k-th coordinate. A special case with $p_k = 2$ for all $k = 1, 2, ..., d$, represents the most popular Gaussian correlation also known as radial basis kernel in Machine Learning literature.

The parameter estimation for μ, σ_z^2 and θ is often a computationally intensive optimization problem. There are a number of R packages that can provide the GP model fitting, for example, **mlegp**, **GPfit**, **DiceKriging**, **tgp**, **RobustGaSP** and **SAVE** (Dancik, 2013; MacDonald et al., 2015; Roustant et al., 2012; Gramacy, 2007; Gu et al., 2016; Palomo et al., 2015). These R packages are somewhat different in terms of computational efficiency and stability. For the reason of stability, we use the R package **GPfit** in this article. The GP model can be fitted using the following code:

```
GPmodel = GPfit::GP_fit(X, Y, corr = list(type="exponential", power=2))
```

The GPfit object `GPmodel` contains the parameter estimates, which can be further passed on for generating the predictions along with uncertainty estimates on a test set. For example, suppose the simulator output is generated by a one-dimensional test function $f(x) = \log(x + 0.1) + \sin(5\pi x)$, and the input points $\{x_1, ..., x_7\}$ are randomly generated as per a space-filling Latin hypercube design method (McKay et al., 1979). Then, Figure 6 shows the fitted surrogate, prediction uncertainty and the true simulator response curves.

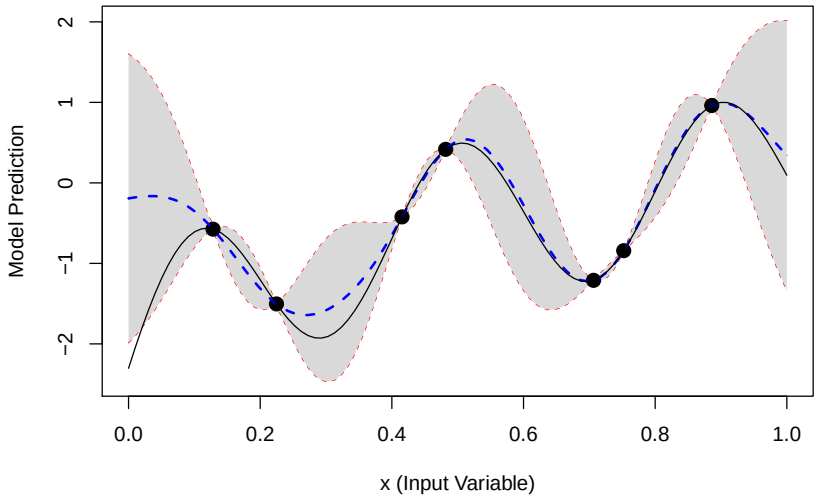


Figure 6: GP regression: The response is generated using the one-dimensional test function $f(x) = \log(x + 0.1) + \sin(5\pi x)$. black solid curve shows the true simulator response, blue dashed curve presents the fitted surrogate and the gray region depicts the uncertainty band.

2.2. Dynamic simulator

A simulator that produces time-series/functional response is referred to as the dynamic computer model. Experimentation via dynamic simulators arise in various applications, for example, rainfall-runoff model (Conti et al., 2009), vehicle suspension system (Bayarri et al., 2007), and the population growth model for European red mites (Zhang et al., 2018) briefly outlined in Section 1.

The time-series structure in the response makes the emulation substantially more challenging as compared to the standard GP model. Recently, a few attempts have been made in this regard. For example, Liu and West (2009) and Farah et al. (2014) proposed time varying autoregressive (TVAR) models. Another clever approach is to represent the time series outputs as linear combinations of a fixed set of basis such as singular vectors (Higdon et al., 2008) and wavelet basis (Bayarri et al., 2007). Zhang et al. (2018) developed an empirical Bayesian approach for the singular value decomposition (SVD) based methodology and generalized it further for large-scale data. We briefly discuss the basic version of SVD-based GP model by Higdon et al. (2008).

Let the simulator inputs and outputs be stored in the $N \times q$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$, and $L \times N$ matrix and $\boldsymbol{Y} = [\boldsymbol{y}(\boldsymbol{x}_1), \ldots, \boldsymbol{y}(\boldsymbol{x}_N)]$, respectively. Then the SVD of $\boldsymbol{Y}$ gives

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$ is an $L \times k$ column-orthogonal matrix, $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_k)$ is a $k \times k$ diagonal matrix of singular values sorted in decreasing order, $\boldsymbol{V}$ is an $N \times k$ column-orthogonal matrix of right singular vectors, and $k = \min\{N, L\}$. Then the simulator response is modelled as

$$\boldsymbol{y}(\boldsymbol{x}) = \sum_{i=1}^{p} c_i(\boldsymbol{x})\boldsymbol{b}_i + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{x} \in \mathbb{R}^q$, and $\boldsymbol{b}_i = d_i\boldsymbol{u}_i \in \mathbb{R}^L$, for $i = 1, \ldots, p$ represent the orthogonal basis. The coefficients c_i's in (5) are assumed to be independent Gaussian processes, *i.e.*, $c_i \sim \mathcal{GP}(0, \sigma_i^2 K_i(\cdot, \cdot; \boldsymbol{\theta_i}))$ for $i = 1, \ldots, p$, where K_i's are correlation functions. We use the popular Gaussian correlation (4) for $K(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta_i})$. The residual term $\boldsymbol{\epsilon}$ in (5) is assumed to be independent $\mathcal{N}(0, \sigma_L^2)$. The number of significant singular values, p, in (5), is determined empirically by the cumulative percentage criterion $p = \min\{m : (\sum_{i=1}^{m} d_i)/(\sum_{i=1}^{k} d_i) > \gamma\}$, where γ is a threshold of the explained variation. With carefully chosen priors, the maximum a posteriori (MAP) method gives closed form predictor along with the uncertainty estimates. We follow the empirical Bayesian implementation by Zhang et al. (2018).

R library called *DynamicGP* (Zhang et al., 2020) provides user-friendly functions for quick usage. The most important function is `svdGP`, and its usage is illustrated as follows:

```
svdGP(design, resp, frac=0.95, nthread=1, clutype="PSOCK", ...)
```

where `design` is the input design matrix, `resp` is the output response matrix, `frac` specifies $\gamma = 95\%$, and `nthread` and `clutype` controls the parallelization of the implementation. See Zhang et al. (2020) for details on the implementation.

Suppose the time-series valued response is generated using the following test function (Forrester et al., 2008) with three-dimensional inputs,

$$f(\boldsymbol{x}, t) = (x_1 t - 2)^2 \sin(x_2 t - x_3), \tag{6}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)^T \in [4, 10] \times [4, 20] \times [1, 7]$, and $t \in [1, 2]$ is on a 200-point equidistant time-grid. Figure 7 illustrates the implementation, by first fitting the svdGP model to a training set of 20 input points randomly generated via maximin Latin hypercube design in the three-dimensional hyper-rectangle $[4, 10] \times [4, 20] \times [1, 7]$, and then predicting the time-series valued simulator output using `svdGP()` function.

From Figure 7, it is clear that the fitted surrogate model predictions are reasonable approximations of the simulator outputs at the design points. We fitted svdGP model using the default settings of *DynamicGP* package. Of course, one can play around with other arguments to obtain better (more accurate) predictions.

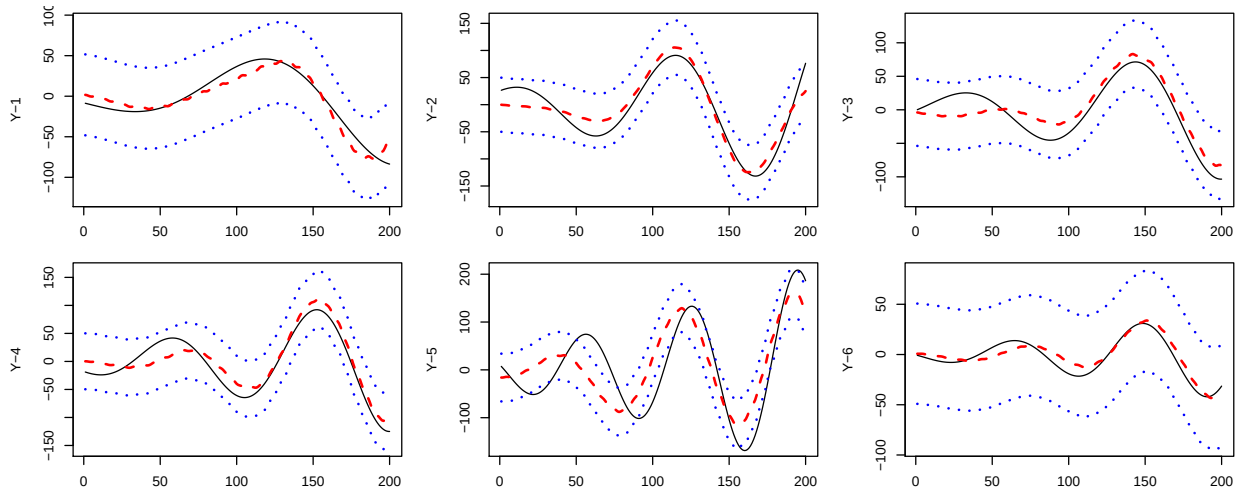


Figure 7: Model prediction for six randomly chosen inputs. Each panel shows the true simulator response (black solid curve), the mean predicted svdGP fit (dashed red curve), and the uncertainty bounds (blue dotted curves).

3. Calibration / Inverse Problem

Calibration of computer simulation models is typically one of the key objectives of computer experiments. For instance, the overall intent of the TDB model application was to ensure that the simulator outputs were as realistic as the observed physical response. This notion of calibration is also referred to as the inverse problem, *i.e.*, the estimation of the set of inputs that yield a pre-specified process value (also called the contour / iso-surface / threshold estimation). Mathematically, the objective is to

$$\text{find } x \in \chi,\ s.t.\ y(x) \approx \xi, \quad \text{or, equivalently,} \quad \underset{x \in \chi}{minimize}\ \|\xi - y(x)\|,$$

where $y(x)$ is the simulator response, ξ is pre-defined target to be estimated, and the objective is to estimate $S(\xi) = \{x \in \chi\ :\ y(x) \approx \xi\}$.

Since realistic simulators of complex processes are often computationally expensive, the number of evaluations of the computer simulator is limited which subsequently makes the inverse problem a lot more challenging. A popular efficient strategy in such a situation is to use sequential design approach which starts with an initial design and adds one point or a batch of points at-a-time iteratively until a tolerance based stopping criterion is met or a pre-specified budget is exhausted. The steps are summarized as follows.

- Step 1. Choose an initial design of run size n_0. Let $n = n_0$.
- Step 2. Build a statistical surrogate model with $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$.
- Step 3. Choose the next design point $\boldsymbol{x}_{n+1}$ by optimizing a merit-based criterion. Run the simulator at $\boldsymbol{x}_{n+1}$ and obtain y_{n+1}.

- Step 4. Let $n = n + 1$ and repeat Steps 2 and 3 until it reaches the budget (N) or satisfies the stopping criterion.

3.1. Calibration of a scalar-valued simulator

Obviously, $c = 1$ is a terrible choice for the player. For then, she must also choose $k = 1$ toss (since there is no opportunity to toss after capturing one node with the first toss); and she will earn 10, 2, 1, 11 nickels with probability 1/4 each. Therefore, per play she will pay 51 cents; she will earn, on average, $5(10+2+1+11)/4 = 30$ cents; and lose 21 cents—a whopping 41.2% loss!

Suppose that the simulator produces a scalar-valued response, and the objective is to estimate $\xi = a$. Ranjan et al. (2008) developed an expected improvement (EI) criterion (for Step 3) under the GP model in Step 2. The idea is to choose $\boldsymbol{x}_{n+1}$ by maximizing

$$\begin{aligned} \mathrm{E}(I(\boldsymbol{x})) &= \int_{v_1(\boldsymbol{x})}^{v_2(\boldsymbol{x})} [\epsilon^2(\boldsymbol{x}) - (t-a)^2] \phi\left(\frac{t-\hat{y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right) dt \\ &= [\epsilon(\boldsymbol{x})^2 - (\hat{y}(\boldsymbol{x})-a)^2 - s^2(\boldsymbol{x})](\Phi(u_2) - \Phi(u_1)) + s^2(\boldsymbol{x})(u_2\phi(u_2) - u_1\phi(u_1)) \\ &\quad +2(\hat{y}(\boldsymbol{x})-a)s(\boldsymbol{x})(\phi(u_2) - \phi(u_1)), \end{aligned} \tag{7}$$

where $u_1 = [a - \epsilon(\boldsymbol{x}) - \hat{y}(\boldsymbol{x})]/s(\boldsymbol{x})$, $u_2 = [a + \epsilon(\boldsymbol{x}) - \hat{y}(\boldsymbol{x})]/s(\boldsymbol{x})$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution function of a standard normal random variable, respectively, and $\epsilon(\boldsymbol{x}) = \alpha s(\boldsymbol{x})$ for a positive constant α. Ranjan et al. (2008) used $\alpha = 1.96$ which is in-sync with 95% confidence interval under the normality assumption of the responses. This criterion is simply obtained by computing the expectation of a carefully designed improvement function $I(\boldsymbol{x}) = \epsilon^2(\boldsymbol{x}) - \min\left\{(y(\boldsymbol{x}) - a)^2, \epsilon^2(\boldsymbol{x})\right\}$ with respect to the predictive distribution $y(\boldsymbol{x}) \sim N(\hat{y}(\boldsymbol{x}), s^2(\boldsymbol{x}))$.

Figure 8 depicts a quick illustration of the contour estimation procedure via the EI criterion for a simulator response generated by a two-dimensional test function.

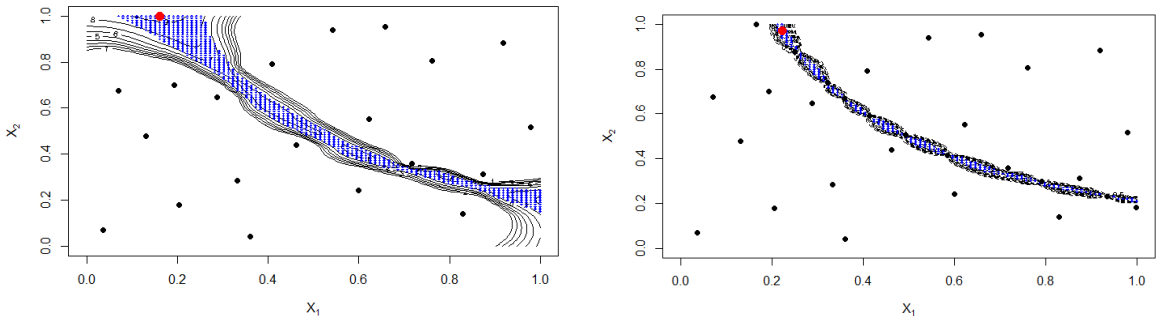


Figure 8: Contour estimation for a two-dimensional test function, $y(x_1, x_2) = (1 + (x_1 + x_2 + 1))(3 + 12x_1x_2)$ **at the target** $a = 300$**, with** $n_0 = 20$ **and** $N = 30$**.**

3.2. Calibration of a dynamic simulator

When we discuss the estimation of the inverse solution for a dynamic computer model, the problem becomes lot more complicated. Note that the target is now a time series, $\xi = \{\xi_1, ..., \xi_L\}$, and the equality or approximation of $y(\boldsymbol{x})$ and ξ have to be formulated more carefully. For a quick reference, the objective of the TDB model application is to find the set of inputs that would produce model outputs closer to the black solid curve shown in Figure 5. An intuitive definition of the inverse problem is to find

$$x^* = \underset{x \in \chi}{\operatorname{argmin}}\ \delta(x), \qquad \text{where} \quad \delta(x) = \|\xi - y(x)\|_2^2.$$

3.2.1. A naive approach

Ranjan et al. (2016) suggested treating the mean discrepancy between the model output and the target, $\omega(\boldsymbol{x}) = \sqrt{\delta(\boldsymbol{x})/L}$, as the scalarized simulator output, then use a GP to emulate $\omega(\boldsymbol{x})$, and adopt the EI-criterion in Jones et al. (1998) for finding the global minimum.

Suppose we consider the following three-dimensional test function with inputs $x = (x_1, x_2, x_3) \in [0, 1]^3$ to generate the time-series response,

$$g(x, t) = \frac{\sin\left(10\pi t^{(2x_3)}\right)}{(1 + 2x_1)t} + |t - 1|^{(2+4x_2)}, \tag{8}$$

where the true field data correspond to $x_0 = (0.5, 0.5, 0.5)$. Figure 9 shows the true field data (solid red curve) and a few simulator outputs (blue dotted curves).

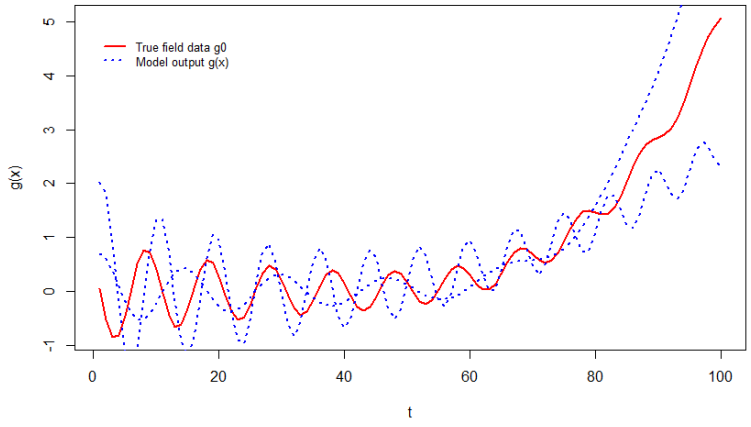
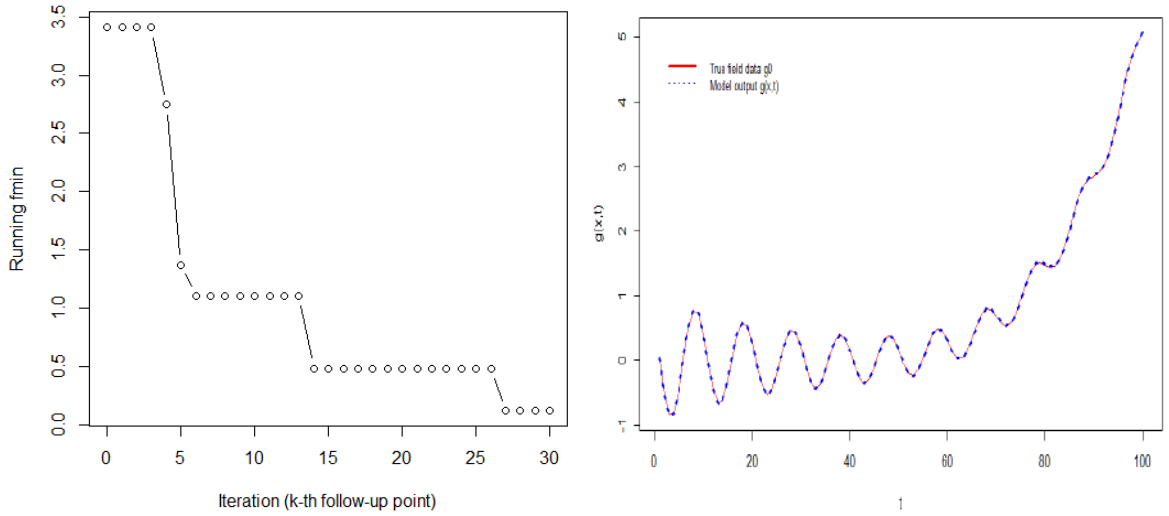


Figure 9: A few computer model outputs and the true data for a three-dimensional test function based dynamic simulator.

Figure 10 shows the implementation of this naive approach for finding the inverse solution using $n_0 = 20$ and $N = 50$.

It turns out that the final estimate of x_{opt} is $(0.4827, 0.4979, 0.4991)$. Though the inverse solution obtained is quite good, there are several theoretical oversights. For instance,



(a) Running estimate of ω_{min} (b) Final inverse solution

Figure 10: Naive approach for solving the inverse problem for a 3-dimensional test function based dynamic simulator.

no efforts are made in Ranjan et al. (2016) for modeling the dynamic computer simulators.

3.2.2. Zhang et al. (2019) approach

Given that the main objective is to efficiently minimize $\delta(\boldsymbol{x}) = \|\boldsymbol{\xi} - \boldsymbol{y}(\boldsymbol{x})\|_2^2$, Zhang et al. (2019) follows the sequential design framework in Jones et al. (1998) for finding the global minimum, and define the improvement function as

$$I(\boldsymbol{x}) = \Big(\delta_{\min} - \delta(\boldsymbol{x})\Big)_+, \tag{9}$$

where $(u)_+ = \max\{0, u\}$ for $u \in \mathbb{R}$ and $\delta_{\min} = \min\{\delta(\boldsymbol{x}_i), i = 1, 2, ..., n\}$. The corresponding EI criterion would look like

$$E[I(\boldsymbol{x})] = E\Big[\Big(\delta_{\min} - \delta(\boldsymbol{x})\Big)_+ \Big| \boldsymbol{Y}\Big], \tag{10}$$

where the expectation is taken with respect to the predictive distribution of $\delta(\boldsymbol{x})$ given $\boldsymbol{Y}$. The distribution of $\delta(\boldsymbol{x})$ is not GP, which differentiates this EI with the standard minimization problem in Jones et al. (1998).

The SVD-based GP surrogate model in Section 2.2 presents the distribution of $y(\boldsymbol{x})$ given $\boldsymbol{Y}$, but not the distribution of $\delta(\boldsymbol{x})$ given $\boldsymbol{Y}$. Zhang et al. (2019) used the saddlepoint approximation method to evaluate (10). The method begins by finding a solution of the derivative equation,

$$\kappa_\delta^{(1)}(s) = \delta_{\min}, \tag{11}$$

where $\kappa_\delta^{(1)}(s)$ is the first order derivative of the cumulant generating function of $\delta(\boldsymbol{x})$ with respect to s. Of course there is no closed form solution of (11)), and they used Broyden's

method (Broyden, 1965) implemented in the R package **nleqslv** (Hasselman, 2010) for numerically solving (11). Let s_0 be the solution of (11), then, for $s_0 > 0, s_0 < 0, s_0 = 0$, Zhang et al. (2019) derived three expressions of saEI (saddlepoint approximation of the expected improvement) which require numerical calculations. The R package **DynamicGP** (Zhang et al., 2020) contains built-in function called `saEI` which facilitates easy computation of this criterion for choosing follow-up design points.

4. Examples and Applications

In this section, we present a few examples to illustrate the performance comparison of saEI approach with other competitors like the naive method outlined in Section 3.2.1. We used a realistic example (TDB model) and some test functions for generating dynamic computer simulator outputs. The performance is measured via $\|x^* - \hat{x^*}\|$ and

$$D_\xi = \frac{\|\xi - y(\hat{x}^*)\|_2^2}{\|\xi - \bar{\xi} 1_L\|_2^2},$$

where x^* is the truth corresponding to the target response ξ, $\hat{x^*}$ is the estimated optimal inverse solution, $\bar{\xi} = \sum_{i=1}^{L} \xi_{t_i}/L$, and 1_L is an L-dimensional vector of ones.

Example 4 of Zhang et al. (2019): Suppose the outputs of the dynamic simulator obey

$$y_t(\boldsymbol{x}) = \exp(3x_1 t + t)\cos(6x_2 t + 2t - 8x_3 - 6), \tag{12}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)^T \in [0,1]^3$ and $t \in [0,1]$ is on a 200-point equidistant time-grid (Harari and Steinberg, 2014). The input producing the target (or, equivalently, the field observation) is randomly generated as $\boldsymbol{x}^* = [0.522, 0.950, 0.427]^T$. Here, the initial design is of size $n_0 = 18$, and $N - n_0 = 36$ points were chosen sequentially one at-a-time as per the individual design criterion (*e.g.*, using $saEI(\boldsymbol{x})$). Figure 11 summarizes the simulation results over 50 replications.

Example 5 of Zhang et al. (2019): The environmental model by Bliznyuk et al. (2008) simulates a pollutant spill at two locations (0 and L) caused by a chemical accident. The simulator outputs are generated using the following model that captures concentration at space-time point (s,t),

$$\begin{aligned} y_t(\boldsymbol{x}) &= C(s,t;M,D,L,\tau) \\ &= \frac{M}{\sqrt{Dt}} \exp\left(\frac{-s^2}{4Dt}\right) + \frac{M}{\sqrt{D(t-\tau)}} \exp\left(-\frac{(s-L)^2}{4D(t-\tau)}\right) I(\tau < t), \end{aligned} \tag{13}$$

where $\boldsymbol{x} = (M, D, L, \tau, s)^T$, M denotes the mass of pollutant spilled at each location, D is the diffusion rate in the chemical channel, and 0 and τ are the time of the two spills. The input domain is $\boldsymbol{x} \in [7,13] \times [0.02, 0.12] \times [0.01, 3] \times [30.01, 30.304] \times [0, 3]$, and $t \in [35.3, 95]$ lies on a regular 200-point equidistant time-grid. In this example, the randomly chosen input that produces the field observation is $\boldsymbol{x}^* = [9.640, 0.059, 1.445, 30.277, 2.520]^T$. Zhang et al. (2019) used a 30-point initial design and 60 follow-up points to estimate the inverse solution. Figure 12 summarizes the results of 50 simulations.

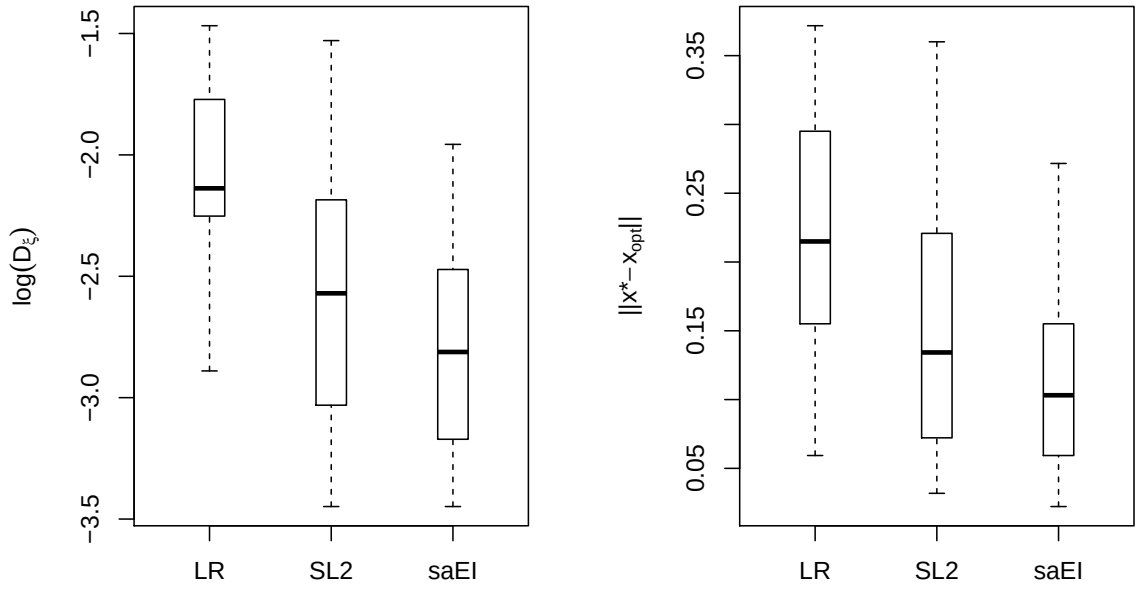


Figure 11: Performance comparison of saEI with SL2 (discussed in Section 3.2.1) and LR method (Pratola et al., 2013). The dynamic responses are generated using a test function from Harari and Steinberg (2014). See Example 4 of Zhang et al. (2019) for details.

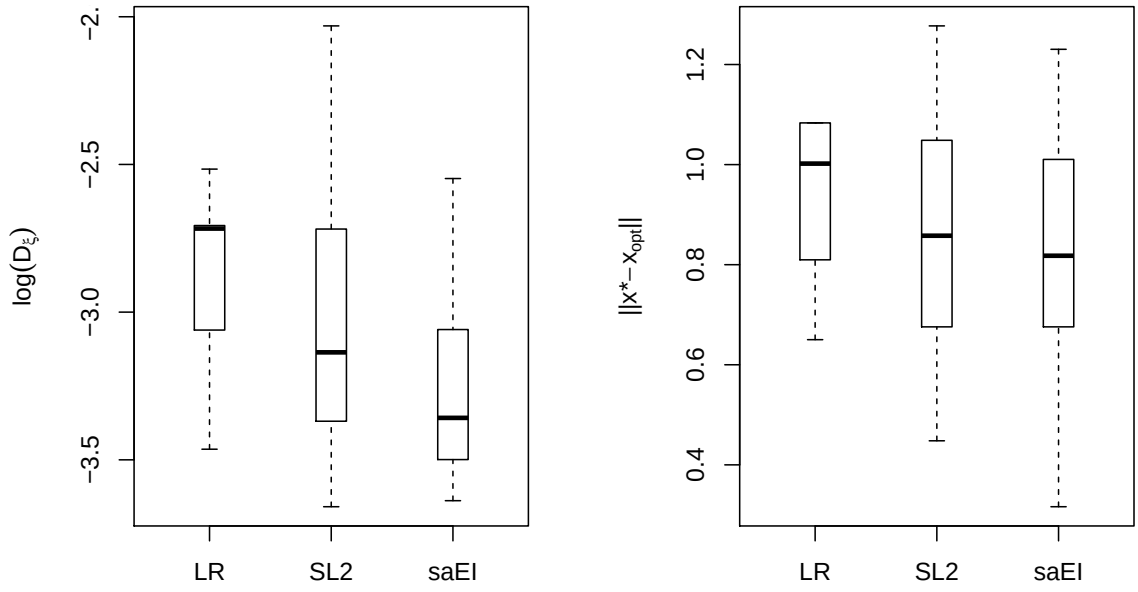


Figure 12: Performance comparison of saEI with SL2 (discussed in Section 3.2.1) and LR method (Pratola et al., 2013). The dynamic responses are generated using a test function from Bliznyuk et al. (2008). See Example 5 of Zhang et al. (2019) for details.

TDB example (Section 4.3 of Zhang et al. (2019)): The target response corresponds to the average count of juvenile population over a 28-day sample during 156-th to 257-th

Julian days (the period during which the apple farming takes place in the Annapolis Valley, Canada). The data was collected twice a week at regular intervals. The version of TDB model used here assumed the following six input variables:

- μ_4 – adult death rate,
- β – maximum fecundity (eggs laid per day),
- ν – non-linear crowding parameter,
- τ_1 – first delay - hatching time of summer eggs,
- τ_2 – second delay - time to maturation of recently hatched eggs,
- Season – average number of days on which adults switch to laying winter eggs.

The performance comparison of the three methods in solving the inverse problem are illustrated in Figure 13. Zhang et al. (2019) used $n_0 = 36$ and $n_{new} = 72$ for finding the inverse solution.

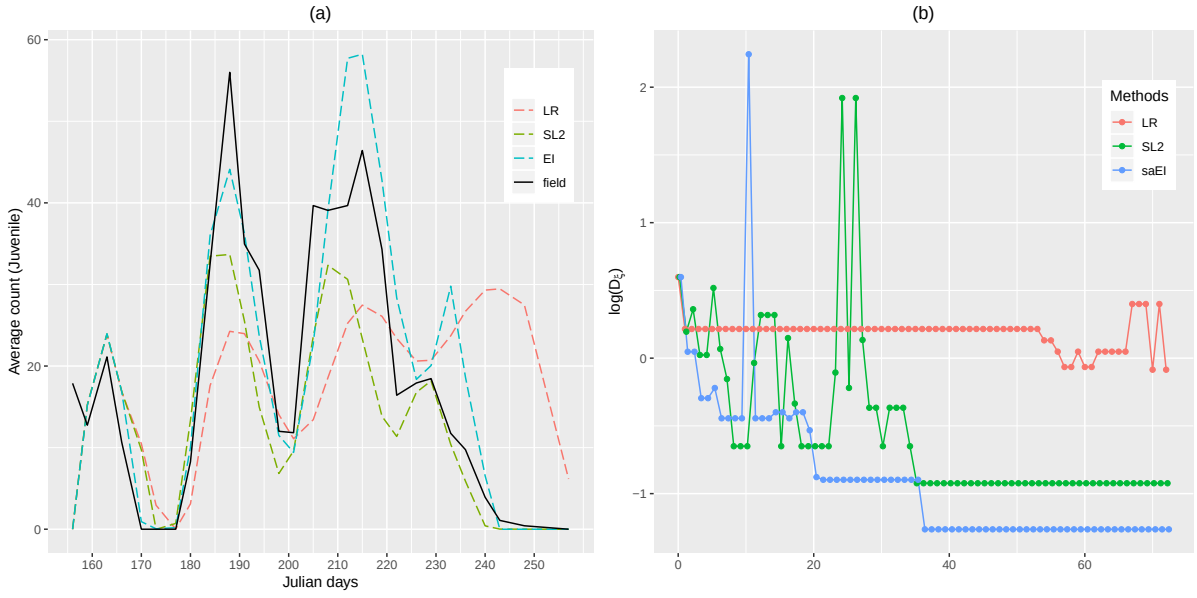


Figure 13: TDB example (Section 4.3 of Zhang et al. (2019)): (a) The field observation of the juvenile ERM population evolution and the located simulator outputs that match the field observation produced by the three sequential design methods. (b) The traces of the $\log(D_\xi)$ criterion values of the 72 iterations for the three methods.

On average, the saEI approach clearly outperforms both the competitors (LR by Pratola et al. (2013) and SL2 by Ranjan et al. (2016)) at least with respect to the examples applications considered here.

5. Concluding Remarks

In this article, we presented the popular statistical metamodel called the GP model for emulating scalar-valued deterministic simulator outputs. This was further generalized to SVD-based GP model for efficient emulation of the dynamic computer simulator response. We also discussed the inverse problem for both scalar and dynamic response simulators that are expensive to evaluate. Several test functions and real-life examples are presented to demonstrate the performance of the methodologies. We also highlight the freely available R libraries for easy implementation of these methodologies. This is particularly helpful for practitioners and young researchers in this area.

There are numerous active research problems in this area such as uncertainty quantification, design and analysis of spatio-temporal simulator data, stochastic simulator data, and in particular how to analyze BIG data coming from computer simulation models.

Acknowledgments This work is primarily based on the Ph.D. thesis of Ru Zhang co-supervised by me and Dr. Chunfang Devon Lin from Queens University, Canada. The key papers published from the thesis are Zhang et al. (2018) and Zhang et al. (2019). I would also like to thank Dr. Vinod Gupta for inviting me to this wonderful conference.

References

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J. and Walsh, D. (2007). Computer model validation with functional output. *The Annals of Statistics*, **35**(**5**), 1874–1906.

Bingham, D., Ranjan, P. and Welch, W. J. (2014). Sequential design of computer experiments for optimization, estimating contours, and related objectives. In J. F. Lawless (Ed.), *Statistics in Action: A Canadian Outlook* (pp. 109–124). Boca Raton, FL: Chapman & Hall/CRC.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S. and Mugunthan, P. (2008). Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, **17**(**2**), 270–294.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, **19**(**92**), 577–593.

Chen, C., Beardsley, R. C. and Cowles, G. (2006). An unstructured grid, finite-volume coastal ocean model (fvcom) system. *Oceanography*, **19**, 78–89.

Conti, S., Gosling, J. P., Oakley, J. E. and O'Hagan, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika*, **96**(**3**), 663–676.

Dancik, G. M. (2013). *mlegp: maximum likelihood estimates of Gaussian processes*. R package version 3.1.4.

Farah, M., Birrell, P., Conti, S. and Angelis, D. D. (2014). Bayesian emulation and calibration

of a dynamic epidemic model for a/h1n1 influenza. *Journal of the American Statistical Association*, **109**(**508**), 1398–1411.

Forrester, A., Sobester, A. and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Gramacy, R. B. (2007). tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software*, **19**(**9**), 1–46.

Gu, M., Palomo, J. and Berger., J. O. (2016). *RobustGaSP: Robust Gaussian Stochastic Process Emulation*. R package version 0.5.7.

Gurney, W., Blythe, S. and Nisbet, R. (1980). Nicholson's blowflies revisited. *Nature*, **287**, 17–21.

Harari, O. and Steinberg, D. M. (2014). Convex combination of Gaussian processes for Bayesian analysis of deterministic computer experiments. *Technometrics*, **56**(**4**), 443–454.

Hasselman, B. (2010). *nleqslv: Solve Systems of Non linear Equations, 2010*. R package version 3.3.2.

Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, **103**(**482**), 570–583.

Jones, D. R., Schonlau, M. and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**(**4**), 455–492.

Karsten, R., McMillan, J., Lickley, M. and Haynes, R. (2008). Assessment of tidal current energy for the minas passage, bay of fundy. In *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy* (pp. 493–507).

Liu, F. and West, M. (2009). A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Analysis*, **4**(**2**), 393–411.

MacDonald, B., Ranjan, P., Chipman, H., et al. (2015). GPfit: an R package for fitting a Gaussian process model to deterministic simulator outputs. *Journal of Statistical Software*, **64**(**12**), 1–23.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **42**(**1**), 55–61.

Palomo, J., Paulo, R. and Garcia-Donato, G. (2015). Save: An r package for the statistical analysis ofcomputer models. *Journal of Statistical Software*, **64**(**13**), 1–23.

Pratola, M. T., Sain, S. R., Bingham, D., Wiltberger, M. and Rigler, E. J. (2013). Fast sequential computer model calibration of large nonstationary spatial-temporal processes. *Technometrics*, **55**(**2**), 232–242.

Ranjan, P., Bingham, D. and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, **50**(**4**), 527–541.

Ranjan, P., Thomas, M., Teismann, H. and Mukhoti, S. (2016). Inverse problem for a time-series valued computer simulator via scalarization. *Open Journal of Statistics*, **6**(**3**), 528–544.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Roustant, O., Ginsbourger, D. and Deville, Y. (2012). Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, **51**(**1**), 1–55.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, **4**(**4**), 409–423.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The design and analysis of computer experiments*. New York: Springer-Verlag.

Teismann, H., Karsten, R., Hammond, R., Hardman, J. and Franklin, J. (2009). On the possibility of counter-productive intervention: the population mean for blowflies models can be an increasing function of the death rate. *Journal of Biological Systems*, **17**(**4**), 739–757.

Zhang, R., Lin, C. D. and Ranjan, P. (2018). Local Gaussian process model for large-scale dynamic computer experiments. *Journal of Computational and Graphical Statistics,*, **27**(**4**), 798–807.

Zhang, R., Lin, C. D. and Ranjan, P. (2019). A sequential design approach for calibrating a dynamic population growth model. *SIAM/ASA Journal on Uncertainty Quantification*, **7**(**4**), 1245–1274.

Zhang, R., Lin, C. D. and Ranjan, P. (2020). *DynamicGP: Modelling and Analysis of Dynamic Computer Experiments*. R package version 1.1-6.

Special Proceedings of the 22nd Annual Conference of SSCA held at S. P. Pune University, during January 02–04, 2020; pp 103–116

Optimal Decisions Under Partial Refunds: A Reward-Earning Random Walk on a Parity Dial

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

Received: 20 July 2020; Revised: 31 July 2020; Accepted: 03 August 2020

Abstract

We solve an unsolved problem posed in Sarkar (2020), which proposed a reward-earning binary random walk game on a parity dial whose twelve nodes, when read clockwise, are labeled as $(1, 11, 3, 9, 5, 7, 6, 8, 4, 10, 2, 0)$. Starting from Node 0, at each step the player tosses a fair coin and moves one step clockwise (if heads) or counterclockwise (if tails). The player pays $25c + k$ cents if she intends to capture c nodes and toss the coin k times. When the c non-zero nodes are captured or when the k tosses are over the game ends; and the player earns as many nickels as the sum of the labels of the captured nodes. The player's objective is to determine (c, k) to minimize the expected percentage loss.

Here we consider a more complex game in which the player is offered several options for a partial refund on each unused toss on payment of an additional upfront overhead fee. Which partial refund offer should she choose? Having chosen the refund option, how should she determine (c, k) to minimize the expected percentage loss?

Under partial refund offers, the player may choose a higher c and a higher k compared to those in the no refund scenario. The optimal choice is discovered through computer simulation, leaving open the theoretical development. Lessons learned from such games empower all parties engaged in the marketplace to determine when to intervene and how to make decisions to benefit from an opportunity and/or prevent a catastrophe.

Key words: Bernoulli variable; Reward random walk; Stopping time; Guaranteed refund.

AMS Subject Classifications: 60G50, 05C81

1. Introduction

It ought to be a common knowledge that when a casino offers you a game of chance and you agree to play, *on average* you should expect to lose money: For otherwise, the Casino would simply toss the game out. You willingly accept this anticipated loss in exchange for deriving some entertainment pleasure and experiencing the excitement of winning a big windfall (although that would happen only rarely). The casino must make money even after paying windfalls, costs, staff salaries, subsidies and taxes. The lure of a game is irresistible

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

when the game *appears to be* in favor of the player for then the casino can entice more players play it more often, and earn more profit for itself. The casino, of course, knows the exact long run prospects of each game it offers. Sarkar (2020) proposed and analyzed such a game, but left as unsolved a more realistic, and more complex, problem of how to choose among several refund policies. Here we take up that generalized problem and discover the optimal choice for the player.

Both the original game and the generalized game serve as models for entrepreneurial decisions and consumer choices. In repeated plays of the game, the optimal choice for each party may be discovered by utilizing the theory of stochastic processes. We direct interested readers to Ross (1996) and Medhi (1982) to encounter the general theory of stochastic processes, to Lovasz (1993) to learn about random walks on graphs, and to Maiti and Sarkar (2019) to study symmetric random walks on paths and cycles. However, to communicate better with researchers outside mathematical sciences, here we rely on computer simulations to discover the optimal choices. Lessons learned from the game will empower all parties engaged in the marketplace to determine when and how to intervene in order to maximally benefit from an opportunity and/or prevent a catastrophe.

In Section 2, we describe the original game proposed in Sarkar (2020) and summarize the optimal choice for the player. In Section 3, we describe the modified game and an expedited search algorithm to conduct the simulation study. In Section 4, we study each refund option and discover the optimal choice of (c, k) through simulation. Section 5 gives the properties of the optimal game within the optimal refund option. Section 6 translates the lessons learnt from this generalized game of reward-earning binary random walk to decision making in the marketplace.

All computations are done using the freeware R, and codes are given in the Annexure.

2. The Original Game

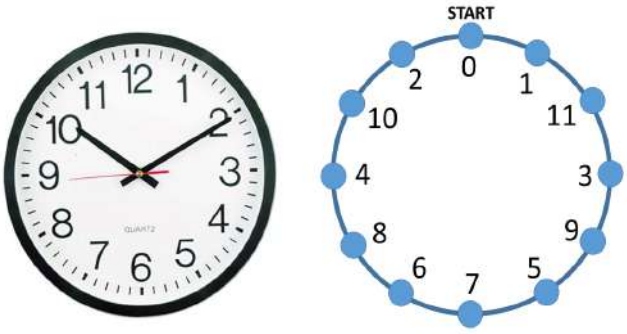


Figure 1: The usual dial of a clock and the parity dial

On a circle there are twelve nodes labeled $(1, 11, 3, 9, 5, 7, 6, 8, 4, 10, 2, 0)$ going clockwise, as shown in Figure 1. The labels are obtained from the usual dial of a clock by changing the top node from 12 to 0 and by interchanging nodes within pairs $(2, 11)$, $(4, 9)$, and $(6, 7)$. Thus, all odd values are on the right half while all even values are on the left half of the dial. Hence, this dial is called the parity dial.

To understand the nature and properties of a binary random walk on any dial, we refer the interested reader to Sarkar (2006). If this random walk also produces earned rewards when a specific node is visited it becomes a reward-earning random walk. Such a reward-earning binary random walk on the parity dial studied in Sarkar (2020): A player pays an admission price of $25c + k$ cents, where c is the number of nodes she intends to visit and capture and k is the number of times she wishes to toss a fair coin. The player begins at Node 0. After each toss, the player moves one step clockwise if the outcome is heads (with probability half), or one step counterclockwise if the outcome is tails; and she captures a node on the first visit to it. The game ends as soon as either c nodes (other than Node 0) are captured or k tosses are over. Then the player will earn as many nickels as the sum of the labels of the captured nodes. How should the player determine c and k?

The choice of $c = 1$ is immediately ruled out because then the player must choose $k = 1$, and pay 26 cents per play. She will earn either one or two nickels with probability half each, or on average, $(5 + 10)/2 = 7.5$ cents. Therefore, she will lose 18.5 cents—a whopping 71.2% loss! Next, for $c = 2$, Sarkar (2020) proves that the optimal choice of k is 6; and in that case, the player stands to lose about 11.7 cents or 21% of her wager of 56 cents. Thereafter, for $3 \leq c \leq 11$, he obtains the optimal k via simulation (based on 10^5 iterations). We summarize his results in Table 1.

Table 1: For the game with no refund, the optimal k's for each $3 \leq c \leq 11$, determine the optimal choice of (c, k) as $(6, 28)$ with an expected loss of 9.69% (marked by a †).

		cents			
c	k	price	E[rew]	E[loss]	E[%loss]
3	10	85	70.13	14.87	17.50
4	16	116	102.27	13.73	11.84
5	22	147	131.53	15.47	10.53
† 6	28	178	160.76	17.24	† 9.69
7	36	211	189.72	21.28	10.09
8	44	244	220.29	23.71	9.72
9	54	279	248.65	30.35	10.88
10	64	314	279.96	34.04	10.84
11	72	347	304.72	42.28	12.18

Based on Table 1, we learn that the gambler's best choice game is $(c = 6, k = 28)$; and with this choice, she faces a 9.69% expected loss. A gambler with a tolerance limit of 10% loss can play this game. The only other choice within her tolerance limit is $(c = 8, k = 44)$ with a 9.72% loss. It is somewhat perplexing that $(c = 7, k = 36)$ results in a higher expected loss of 10.09% than either $(c = 6, k = 28)$ or $(c = 8, k = 44)$. But this can be explained by noting that the node labels an odd distance away from Node 0 are typically smaller than node labels at an even distance.

3. The Generalized Game with Refund Options

To understand the need and the nature of the generalized game, let me paraphrase a very productive conversation I had with one attentive listener when I gave a talk on the above-stated original game at the 2020 Pune Conference of the SSCA. During high tea after the conference ended, among other things, this compassionate (to a gambler who would play this game) and extravagantly helpful (to me) gentleman spoke to me thus:

> "When the gambler captures c nodes, she is happy. But wouldn't she feel poorly about forfeiting all the unused tosses she already has paid for?"
>
> "Indeed, she would. That is the very essence of the decision-making problem in choosing both c and k optimally. Having chosen a c, if the gambler picks too small a k, she will likely not have captured all c nodes when she has tossed k times. On the other hand, if she picks too large a k, by the time she has captured c nodes she would have many unused tosses which she would forfeit. She must choose k cautiously."
>
> "You mentioned offering a guaranteed refund in exchange for the unused tosses. If you refund a full penny for every unused toss, the game surely becomes more attractive to the player. It may even become favorable to the gambler! Is that something the Gambling House will allow?"
>
> "A guaranteed refund does not mean a full refund. If it did, then the gambler would simply pay for $k = 1000$ (or a large number thereabouts) knowing that there is no risk of losing the excess payment. She would recover it all as soon as she captures c nodes, which will happen with almost certainty. For all practical purposes, one can think as if the player pays $25c$ at the start of the game and then pays one penny before each toss until c nodes are captured, requiring a random number of tosses K_c. Alternatively, the payment of $25c + K_c$ can be determined when the game ends. In either case, the problem changes to choosing c alone. Furthermore, the Gambling House will likely charge the player an upfront fee to purchase this option to get a 100% refund."

I had not calculated the optimal c under the full refund scenario with or without any fee since a 100% refund option was not on my mind prior to this conversation. Therefore, I could not talk about the optimal choice of c, except to say that it is likely to be an even number 6 or more, and to reiterate that that is not what I meant by a guaranteed refund.

For the benefit of my readers, I have since then carried out that missing simulation. R codes are given in the Annexure. Table 2 summarizes the expected percentage loss for various choices of c under 100% refund at overhead fees $0, 5, 10, \ldots, 30$ cents. I should point out that in this scenario, the (random) number of tosses is not right truncated by a predetermined k as in the original game; hence, the price paid is genuinely random (and it is determined when the game ends with c nodes captured). Hence, the expected percentage loss is calculated only after the game ends using the formula

$$\text{E[\%loss]} = 100 \times \frac{\text{E[loss] + overhead fee}}{\text{E[price] + overhead fee}}$$

Bear in mind that the very definition of expected loss has changed! One could pretend as if 100 extra tosses are paid for and the extra payment is recovered as refund, in which

Table 2: Determining optimal c (marked by a *) under a 100% refund option bought upfront on payment of some overhead fee

	cents			E[%loss] when overhead fee is (in cents)						
c	E[price]	E[rew]	E[loss]	0	5	10	15	20	25	30
3	81.00	72.51	8.48	10.48	15.67	20.31	24.46	28.20	31.58	34.67
4	110.02	105.97	4.05	3.68	7.87	11.71	15.24	18.50	21.52	24.31
5	139.99	137.51	2.48	1.77	5.16	8.32	11.28	14.05	16.66	19.11
6	171.03	170.02	1.01	0.59	3.41	6.08	8.61	11.00	13.27	15.43
7	202.94	201.24	1.71	0.84	3.23	5.50	7.67	9.74	11.72	13.61
8	235.90	235.06	0.83	*0.35	*2.42	*4.40	6.31	8.14	9.90	11.59
9	270.10	265.52	4.58	1.70	3.48	5.21	6.87	8.47	10.02	11.52
10	305.01	299.98	5.03	1.65	3.24	4.77	*6.26	*7.70	*9.10	*10.46
11	341.19	330.00	11.19	3.28	4.68	6.03	7.35	8.64	9.88	11.10

case the percentage loss can be reduced artificially, since the denominator increases by 100 but the numerator remains the same. More extremely, if 1000 tosses are paid for then the percentage loss is driven down to almost zero! Notwithstanding, to compare different values of c, invoking monotonic relation, our adopted definition of expected percentage loss works just fine. For the 100% refund option with an overhead fee of 14 cents or less (details are not shown), the best choice is $c = 8$ (and a very large k), but for a fee of 15 cents or more, it is $c = 10$. As anticipated, as the overhead fee increases, so does the player's percentage loss.

An astute reader can anticipate how our post-conference conversation ended:

> "If not a 100% refund of the price of the unused tosses, what then do you mean by a guaranteed refund?"
>
> "A guaranteed refund means a percentage of the purchase price of the unused tosses will be refunded if the player had bought this option by paying an additional overhead fee at the very outset of the game. For instance, in the original game, the guaranteed refund is 0% for an overhead fee of 0 cents: The gambler gets nothing back on the unused tosses; and pays no extra fee. The Gambling House could offer several options: (1) 50% refund for a fee of 5 cents; (2) 60% refund for a fee of 7 cents; (3) 70% refund for a fee of 10 cents; (4) 80% refund for a fee of 15 cents. In each case, we would ask what is the optimum choice of (c, k)? When we answer these questions, we can determine which of the four offers of guaranteed percentage refund is optimum."

In this paper, I will answer the optimal choices in the modified game that offers a partial refund of unused tosses for a modest fee upfront. A player who was intending to play the original $(c = 6, k = 28)$ game, when offered the modified game with partial refund, has some incentive to pay for a few extra tosses at the outset in hope of improving her chance of capturing all $c = 6$ nodes; and yet should she capture them early, she can recover a percentage of her wager. What should be her best choice now? If this offer were available at no overhead fee, the player would lower her expected percentage loss below that in the original game (where it was 9.69%). But the presence of an overhead fee makes it challenging to anticipate the expected percentage loss without studying the process in more details.

Moreover, the offer of a partial refund may cause the player to rethink how many nodes she should set out to capture. This paper is devoted to answering the optimal choice among the four percentage refunds—50%, 60%, 70%, 80%—with associated overhead fees 5, 7, 10, 15 cents, respectively—and the corresponding optimal choice of (c, k).

4. Optimal Games Under Different Percentage Refunds

Suppose that the Gambling House offers the gambler for an upfront payment of 5 cents, a 50% refund on the purchase price of all unused tosses by the time the player captures c nodes. How should the player determine (c, k)? We leave to the reader to check that, as it was in the original game, choosing $c = 1$ or $c = 2$ is not good for the gambler.

Proceeding in a routine manner, for every fixed $3 \leq c \leq 11$, one can simulate the expected loss for various choices of $k \geq c$. However, a smarter search algorithm can be implemented: Sarkar (2020) argued that for each contemplated c, the player is better off choosing an *even* $k \geq c$. Roughly speaking, this is because on the parity dial nodes at an odd distance away from Node 0 have smaller labels compared to nodes at an even distance away. Below we exhibit the simulation results for the choice of $c = 8$, and $50 \leq k \leq 60$, demonstrating that indeed k ought to be chosen an even number because for each odd k, the expected percentage loss is lower at both of its even neighbors.

Table 3: For 50% refund at 5 cents, expected reward and expected loss for $c = 8$ and $50 \leq k \leq 60$, exhibit that even values of k are preferable.

		cents			E[%loss]	
c	k	price	E[rew]	E[loss]	k odd	k even
8	50	255	233.54	21.46		8.42
8	51	256	234.29	21.71	8.48	
8	52	257	235.54	21.46		8.35
8	53	258	236.26	21.74	8.43	
8	54	259	237.55	21.45		8.28
8	55	260	238.25	21.75	8.37	
8	56	261	239.47	21.53		*8.25
8	57	262	239.90	22.10	8.44	
8	58	263	240.82	22.18		8.43
8	59	264	241.64	22.36	8.47	
8	60	265	242.62	22.38		8.44

Moreover, having found the optimal k for a specific c, say k_c, the search for the optimal k for $(c + 1)$ can be expedited by taking k even, and not just larger than $(c + 1)$ but larger than k_c. Henceforth, for all percentage refund options, we shall only look at even $k > k_{c-1}$ corresponding to each contemplated c. In Tables 4 and 5, for the partial refund options (1)–(4) we document the expected percentage loss corresponding to each $c \in \{3, 4, \ldots, 11\}$ and selected k's that help us determine the optimal (c, k). Finally, using Tables 4 and 5, we choose the best among the four positive refund options (1)–(4).

Table 4: Expected percentage losses for $3 \le c \le 11$ and selected even k's determine the optimal (c, k), under refund options (1) and (2). For each refund option, min E[%loss] within c is marked by *, and the minimum overall by #.

		(1) 50% refund at 5 cents						(2) 60% refund at 7 cents			
c	k	price	E[rew]	E[loss]	E[%loss]	c	k	price	E[rew]	E[loss]	E[%loss]
3	8	88	68.88	19.12	21.73	3	12	94	74.92	19.08	20.30
3	10	90	72.14	17.86	19.85	3	14	96	76.75	19.25	*20.05
3	12	92	74.35	17.65	*19.19	3	16	98	78.24	19.76	20.17
3	14	94	75.91	18.09	19.25	3	18	100	79.53	20.47	20.47
3	16	96	77.18	18.82	19.60	3	20	102	80.84	21.16	20.74
4	16	121	105.48	15.52	12.83	4	18	125	108.51	16.49	13.20
4	18	123	107.75	15.25	12.40	4	20	127	110.50	16.50	*12.99
4	20	125	109.53	15.47	*12.38	4	22	129	112.22	16.78	13.01
4	22	127	110.95	16.05	12.64	4	24	131	113.76	17.24	13.16
4	24	129	112.47	16.53	12.81	4	26	133	115.18	17.82	13.40
5	22	152	135.54	16.46	10.83	5	24	156	138.87	17.13	10.98
5	24	154	137.99	16.01	10.39	5	26	158	141.06	16.94	10.72
5	26	156	139.92	16.08	*10.31	5	28	160	143.17	16.83	*10.52
5	28	158	141.62	16.38	10.37	5	30	162	144.88	17.12	10.57
5	30	160	143.32	16.68	10.43	5	32	164	146.44	17.56	10.71
6	32	187	170.11	16.89	9.03	6	34	191	173.54	17.46	9.14
6	34	189	172.28	16.72	* 8.85	6	36	193	175.55	17.45	9.04
6	36	191	174.07	16.93	8.86	6	38	195	177.52	17.48	8.96
6	40	195	177.21	17.79	9.12	6	40	197	179.16	17.84	* 9.06
6	38	193	175.61	17.39	9.01	6	42	199	180.77	18.23	9.16
7	40	220	200.09	19.91	9.05	7	42	224	203.88	20.12	8.98
7	42	222	202.28	19.72	* 8.88	7	44	226	205.81	20.19	8.93
7	44	224	204.09	19.91	8.89	7	46	228	207.88	20.12	* 8.82
7	46	226	205.88	20.12	8.90	7	48	230	209.66	20.34	8.84
7	48	228	207.56	20.44	8.96	7	50	232	211.35	20.65	8.90
8	50	255	233.57	21.43	8.40	8	54	261	239.55	21.45	8.22
8	52	257	235.58	21.42	8.33	8	56	263	241.40	21.60	8.21
8	54	259	237.50	21.50	8.30	8	58	265	243.29	21.71	8.19
8	56	261	239.40	21.60	#* 8.28	8	60	267	245.20	21.80	#* **8.16**
8	58	263	240.85	22.15	8.42	8	62	269	246.81	22.19	8.25
9	60	290	263.10	26.90	9.28	9	66	298	271.11	26.89	9.02
9	62	292	265.04	26.96	9.23	9	68	300	273.11	26.89	* 8.96
9	64	294	266.98	27.02	* 9.19	9	70	302	274.80	27.20	9.01
9	66	296	268.71	27.29	9.22	9	72	304	276.72	27.28	8.97
9	68	298	270.42	27.58	9.26	9	74	306	278.33	27.67	9.04
10	72	327	297.21	29.79	9.11	10	76	333	303.69	29.31	8.80
10	74	329	299.15	29.85	9.07	10	78	335	305.75	29.25	* 8.73
10	76	331	301.05	29.95	* 9.05	10	80	337	307.38	29.62	8.79
10	78	333	302.88	30.12	9.05	10	82	339	309.38	29.62	8.74
10	80	335	304.48	30.52	9.11	10	84	341	311.00	30.00	8.8
11	86	366	328.17	37.83	10.34	11	90	372	335.15	36.85	9.91
11	88	368	330.38	37.62	10.22	11	92	374	336.85	37.15	9.93
11	90	370	332.22	37.78	*10.21	11	94	376	338.76	37.24	9.91
11	92	372	333.74	38.26	10.29	11	96	378	340.62	37.38	* 9.89
11	94	374	335.34	38.66	10.34	11	98	380	342.31	37.69	9.92

Table 5: Expected percentage losses for $3 \leq c \leq 11$ and selected even k's determine the optimal (c, k), under refund options (3) and (4). For each refund option, min E[%loss] within c is marked by *, and the minimum overall by #.

		(3) 70% refund at 10 cents						(4) 80% refund at 15 cents			
c	k	price	E[rew]	E[loss]	E[%loss]	c	k	price	E[rew]	E[loss]	E[%loss]
3	14	99	77.55	21.45	21.67	3	22	112	85.29	26.71	23.85
3	16	101	79.20	21.80	21.59	3	24	114	86.88	27.12	23.79
3	18	103	80.77	22.23	*21.58	3	26	116	88.52	27.48	23.69
3	20	105	82.24	22.76	21.68	3	28	118	90.12	27.88	23.63
3	22	107	83.63	23.37	21.84	3	30	120	91.68	28.32	*23.60
4	20	130	111.54	18.46	14.20	4	26	141	118.38	22.62	16.04
4	22	132	113.54	18.46	*13.99	4	28	143	120.15	22.85	15.98
4	24	134	115.11	18.89	14.10	4	30	145	121.85	23.15	*15.97
4	26	136	116.87	19.13	14.07	4	32	147	123.49	23.51	15.99
4	28	138	118.35	19.65	14.24	4	34	149	125.07	23.93	16.06
5	28	163	144.46	18.54	11.38	5	34	174	151.79	22.21	12.76
5	30	165	146.39	18.61	*11.28	5	36	176	153.71	22.29	12.67
5	32	167	148.13	18.87	11.30	5	38	178	155.50	22.50	*12.64
5	34	169	149.89	19.11	11.31	5	40	180	157.04	22.96	12.75
5	36	171	151.51	19.49	11.40	5	42	182	158.81	23.19	12.74
6	38	198	179.28	18.72	9.45	6	46	211	188.85	22.15	10.50
6	40	200	181.21	18.79	* 9.40	6	48	213	190.64	22.36	10.50
6	42	202	182.94	19.06	9.44	6	50	215	192.51	22.49	*10.46
6	44	204	184.77	19.23	9.43	6	52	217	194.24	22.76	10.49
6	46	206	186.40	19.60	9.51	6	54	219	195.95	23.05	10.53
7	48	233	211.84	21.16	9.08	7	58	248	223.73	24.27	9.79
7	50	235	213.64	21.36	9.09	7	60	250	225.54	24.46	9.79
7	52	237	215.57	21.43	* 9.04	7	62	252	227.48	24.52	* 9.73
7	54	239	217.34	21.66	9.06	7	64	254	229.07	24.93	9.82
7	56	241	219.15	21.85	9.07	7	66	256	230.89	25.11	9.81
8	58	268	245.87	22.13	8.26	8	70	285	260.11	24.89	8.74
8	60	270	247.84	22.16	8.21	8	72	287	262.02	24.98	8.70
8	62	272	249.71	22.29	#* 8.20	8	74	289	263.87	25.13	* 8.69
8	64	274	251.38	22.62	8.19	8	76	291	265.68	25.32	8.70
8	66	276	253.39	22.61	8.26	8	78	293	267.24	25.76	8.76
9	72	307	279.68	27.32	8.90	9	82	322	292.56	29.44	9.14
9	74	309	281.39	27.61	8.94	9	84	324	294.30	29.70	9.17
9	76	311	283.46	27.54	* 8.85	9	86	326	296.18	29.82	* 9.15
9	78	313	285.11	27.89	8.91	9	88	328	297.96	30.04	9.16
9	80	315	286.95	28.05	8.90	9	90	330	299.63	30.37	9.20
10	84	344	314.41	29.59	8.60	10	100	365	333.37	31.63	8.67
10	86	346	316.45	29.55	8.54	10	102	367	335.23	31.77	8.66
10	88	348	318.38	29.62	* 8.51	10	104	369	337.33	31.67	#* 8.58
10	90	350	320.02	29.98	8.57	10	106	371	338.65	32.35	8.72
10	92	352	321.70	30.30	8.61	10	108	373	340.42	32.58	8.73
11	98	383	345.96	37.04	9.67	11	112	402	363.42	38.58	9.60
11	100	385	347.67	37.33	9.70	11	114	404	365.38	38.62	9.56
11	102	387	349.84	37.16	* 9.60	11	116	406	367.17	38.83	* 9.56
11	104	389	351.32	37.68	9.69	11	118	408	368.85	39.15	9.59
11	106	391	353.13	37.87	9.68	11	120	410	370.66	39.34	9.60

From Tables 4 and 5, we note that for the partial refund options of 50%, 60%, 70% and 80% on payment of 5, 7, 10, 15 cents, respectively, the optimal (c, k) are $(8, 56)$, $(8, 60)$ $(8, 62)$ and $(10, 104)$. Recall that in the original game with no refund the optimal game was $(6, 28)$. Thus, as the refund percentage increases, the gambler not only may choose to buy more tosses, but also commit to capturing more nodes! The safety net of getting a partial refund on prepaid excess tosses, makes the player more inclined to targeting a higher c and choosing a higher k.

5. Properties of the Optimal Game Under the Optimal Refund Option

In the previous section, we learned that among the various refund options offered to the gambler, the best is Option (2): 60% refund on payment of 7 cents. For this case, the optimum (c, k) is $(8, 60)$. This optimal game has admission price $25 \times 8 + 60 + 7 = 267$ cents. For this optimum game, we exhibit some characteristics such as the number of tosses until the game ends, the probability distribution of the number of nodes captured, the probability distribution of the farthest node captured going clockwise from Node 0, and the probability distribution of the reward earned (plus refund).

Based on a simulation of 10^6 (one million=ten lakhs) iterations of game $(8, 60)$, we can estimate the number of tosses until the game ends by capturing all 8 nodes using Figure 2.

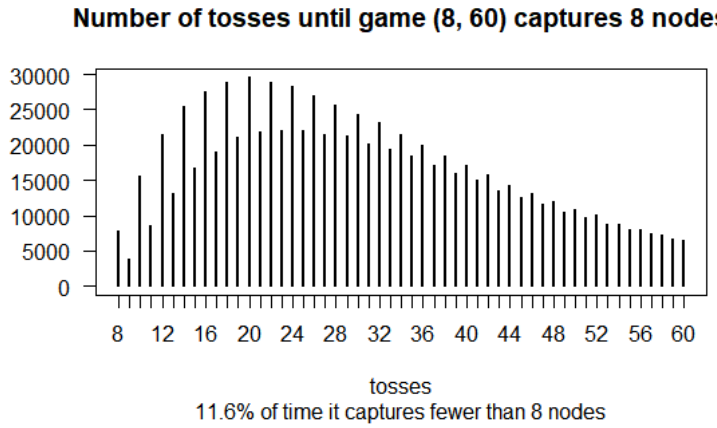


Figure 2: The number of tosses until game $(8, 60)$ ends by capturing 8 nodes. About 11.6% of times the game captures fewer than 8 nodes in 60 tosses.

The probability of capturing 8 nodes in 8, 9 or 10 tosses are respectively $2^{-7}, 2^{-8}, 2^{-6}$. This is supported by the simulation where out of 10^6 iterations the frequencies of 8, 9, 10 tosses are respectively 7790, 3896, 15632 (*P-values* of chi-square tests with one degree of freedom are respectively .8027, .8758, .9582). We leave it to the inquisitive reader to explain a curious feature observed in Figure 2: The relative frequencies for odd number of tosses are smaller when compared to those of their two immediate neighboring even values!

The number of tosses until game (8, 60) ends with 8 nodes captured is shown in Figure 2. In 115,872 more iterations (not shown in Figure 2) fewer than 8 nodes are captured in 60 tosses. For these iterations, how many nodes are actually captured? The answer is

given in Table 6, which also shows the corresponding probabilities, which are correct to three decimal places because their 95% confidence intervals are at most 0.001 wide (see Devore and Berk, 2007, for example). All 8 vertices are captured 88.4% of times; 7 nodes 7.2% of times and 6 or fewer nodes 4.4% of times.

Table 6: The simulated distribution of the number of nodes captured by game $(8, 60)$ shows about 11.6% of time not all 8 nodes are captured.

# nodes	1	2	3	4	5	6	7	8	sum
frequency	0	0	9	841	8823	34224	71975	884128	1000000
probability	.000000	.000000	.000009	.000841	.008823	.034224	.071975	.884128	1.000000

Using the information in Figure 2 and Table 6, the number of tosses until game (8, 60) ends has the following summary statistics (see R code in the Annexure):

N=10^6, Min=8, Q1=21, Q2=31, Mean=33.88, Q3=46, Max=60, SD=15.50

The probability distribution of the number of tosses left over when the game ends is obtained simply by subtracting from 60 the number of tosses needed to capture 8 nodes (and adding 0.115872 to the probability that no toss is left over). Thereafter, one can construct the probability distribution of the refund amount by multiplying the number of leftover tosses by the refund percentage.

Also, based on this same simulation, and using the built-in kernel density estimator in R (see Silverman, 1986), the estimated probability density function of the reward earned (plus refund) is shown in Figure 3, with its summary statistics given by

N=10^6, Min=64.4, Q1=235.2, Q2=248.2, Mean=245.2, Q3=264.2, Max=286.2, SD=27.82

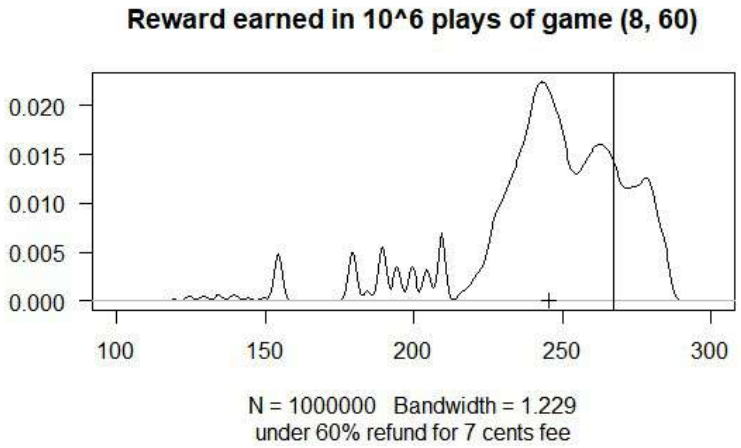


Figure 3: The reward (plus refund) distribution in game $(8, 60)$ exhibits mean=245.2 (the + sign), SD=27.82, a 20.4% chance of winning (reward > 267 cents (the vertical line)), and an expected percentage loss of 8.18.

Using the estimated density function given in Figure 3, we can infer (see Rohatgi, 2003, for example) some probabilities the gambler would like to know. While on average the gambler loses 8.18% of her wager of 267 cents per game, about 20.4% of the time she earns more reward (plus refund) than the wager. Thus, the game does not always end in a loss for the player. Moreover, a gambler who has a tolerance limit of 10% loss, actually earns back more than 90% of her wager (or 240.3 cents) about 66.1% of times, making the game quite attractive to her!

Finally, note that Option (3), which offers a 70% refund on payment of 10 cents overhead fee, is a close second best when the player chooses $(c = 8, k = 62)$, with an expected loss of 8.20%. We leave it to the interested reader to study its properties by adapting the R codes given in the Annexure.

6. Lessons Learnt From The Game

The problem studied in Sarkar (2020) was formulated in response to an invitation to deliver a keynote speech at the 2nd International Conference on "Frontiers of Operations Research & Business Studies" held during 27–28 December 2019, at the Calcutta Business School. The mission of FORBS (see FORBS, 2019) is described as follows:

> Most often organizations are confronted with questions like how to make a good decision? What is really a good decision? What constitutes a poor decision? Is there any pattern in the decisions made? In the quest for finding answers to these questions, the contributions of several disciplines like statistics, mathematics, sociology, economics, information technology, operations research and behavioral science need to be acknowledged. In other words "Decision Sciences incorporate an economic framework—a consistent, rational and objective system to "price" each possible outcome, taking into account risks and rewards."

Sarkar (2020) demonstrated the essential elements of optimal decision making in a rather simple model: Choose (c, k) to minimize the expected percentage loss in the reward-earning binary random walk game on the parity dial. Here we have expanded that problem to incorporate one more layer of complexity: First choose the refund policy offered at several different options with associated overhead fees, and then choose (c, k) to minimize the expected percentage loss.

We showed that when the gambler judiciously buys the optimal option for partial refund, the game may become more favorable to the gambler than playing the original game with no refund. However, we will be remiss if we did not mention that the Gambling House still has the last laugh: It can, for instance, raise the overhead fee for each refund option by, say, 5 cents. Then the best option for the gambler will be the original game with an expected percentage loss of 9.79. The other refund options (1)–(4) has expected percentage loss of 10.00, 9.853, 9.852, 9.805, respectively. To calculate these percentage losses, simply take the ratio of expected loss and price after adding to both quantities the change in the overhead fee, as in Table 2. See R codes in the Annexure. However, to keep the gambler playing, the Gambling House cannot remain totally adversarial; it must keep the overhead price in check. It is precisely this tension that keeps decision making exciting and intriguing.

The reward-earning random walk game, translates in the marketplace into a decision about investing sufficient resources to ensure a good chance of fulfilling the mission of a venture. However, to avoid letting the unused part of the investment go to waste, the entrepreneur will act prudently by purchasing an insurance to protect the resource. Thus, the refund option can be thought of as an insurance policy. Should the entrepreneur accomplish the goal of the venture and still have some resources left over, she will at least get back a predetermined percentage as refund. The insurance company that underwrites such an insurance plan likely has a market where they can resell the leftover resources and pass on (part of) the proceeds to the insuree.

The central message of the reward-earning random walk games is that while facing uncertainty of outcomes, an entrepreneur can and should make the best decision based on the information available, and adjust the decision should the conditions change. A careful and adequate planning and flexibility in decision making are necessary to maximize the expected return from a venture.

Acknowledgment

I thank the participants of the 22nd Annual Conference of the SSCA, especially the discussant during high tea, for their inquisitive questions which inspired me to solve the problem I had presented there as unsolved.

References

Devore, J. L. and Berk, K. N. (2007). *Modern Mathematical Statistics*. Thomson-Brooks / Cole, Belmont, California.

FORBS (2019). *Mission Statement*. https://calcuttabusinessschool.org/FORBS2019/

Lovasz, L. (1993). Random walks on graphs: A survey. In: *Combinatorics, Paul Erdös is Eighty*, Bolyai Society Mathematical Studies, 2 Keszthely (Hungary), 11–46.

Maiti, S. I. and Sarkar, J. (2019). Random walks on paths and cycles. *Mathematics Magazine*, **92(4)**, 252–268. DOI: 10.1080/0025570X.2019.1611166

Medhi, J. (1982). *Stochastic Processes*. A Halsted Press Book, John Wiley & Sons, New York.

Rohatgi, V. K. (2003). *Statistical Inference*. Second Edition, Dover, New York.

Ross, S. M. (1996). *Stochastic Processes*. Second Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Sarkar, J. (2006). Random walk on a polygon. In: *Recent Developments in Nonparametric Inference and Probability*, IMS Lecture Notes **50**, 31–43. Institute of Mathematical Statistics, Beachwood, Ohio.

Sarkar, J. (2020). Rules, resolutions, randomness and reward: Lessons from a game. *Proceedings of the 2nd International Conference on "Frontiers of Operations Research & Business Studies" (FORBS)*, 27–28 December 2019, Calcutta Business School.

Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall, London, UK.

ANNEXURE

We document the R codes used to prepare Tables 2–6 and draw Figures 2 and 3.

```
### Table 2: Simulate expected nb of tosses, price reward, loss and %loss
## with 0% refund at 0 cents overhead fee
k=1000  # a very large number of tosses
reward=function(c,k){    # c=vertices to capture, k=tosses allowed
  rf=c(1,11,3,9,5,7,6,8,4,10,2,0)
  ber=2*rbinom(k,1,1/2)-1; cber=c(0,cumsum(ber));
  (nv=cummax(cber)-cummin(cber)); l=sum(nv<c); l1=l+1
    if(l<=k){cber=cber[1:l1]}
  maxv=max(cber); vv=seq(min(cber),max(cber))
  visited=vv*(vv>0)+(vv+12)*(vv<=0);
  nvv=length(visited); rew=5*sum(rf[visited])
  c(nvv,maxv,rew,l) }

el=matrix(0,9,6)   # initialize E[%loss] matrix
for (i in 1:9){
  data=replicate(10^5,reward(i+2,k))
  price=0+25*(i+2) + mean(data[4,])   # no fee yet
  rew=mean(data[3,]);  loss=price-rew
  el[i,]=(c( round(i+2,0), round(k,0), round(price,2),
     round(rew,2),round(loss,2), round(100*loss/price,2) ) }
el # calculate more columns as (E[loss]+fee)/(E[price]+fee)

### Tables 3, 4 and 5: How much is the random reward?
reward=function(c,k){ # c=vertices to capture, k=tosses allowed
  rf=c(1,11,3,9,5,7,6,8,4,10,2,0) # nickels at the nodes
  ber=2*rbinom(k,1,1/2)-1; cber=c(0,cumsum(ber));
  (nv=cummax(cber)-cummin(cber)) # nb of non-zero vertices
  l=sum(nv<c)  # nb tosses until capture c nodes
  l1=l+1; if(l<=k){cber=cber[1:l1]} # cber has an initial 0
  maxv=max(cber); vv=seq(min(cber),max(cber)) # vertices captured
  visited=vv*(vv>0)+(vv+12)*(vv<=0) # recode vertices captured
  nvv=length(visited) # nb of nodes visited (includes Node 0)
  rew=5*sum(rf[visited]) + (k-l)*0.60 # add refund (% of excess tosses)
  c(nvv-1,maxv,rew,l) } # outputs

## (refund %, payment)=(.50, 5), (.60, 7), (.70, 10), (.80, 15)
## simulate expected nb of tosses, price, reward, loss and %loss
c=8  # 2, 3, 4, ..., 11
for (k in seq(58,62,2)){     # try a range of values of k (even)
  price=10+25*c+k  # overhead fee + admission
  data=replicate(10^5,reward(c,k))
  rew=mean(data[3,]); loss=price-rew
```

```
  print(c( round(c,0), round(k,0), round(price,2),
   round(rew,2),round(loss,2), round(100*loss/price,2) )) }

### Figures 2, 3 and Table 6: Properties of the optimum game
##  (refund 60%, overhead 7 cents) (c=8, k=60)
c=8; k=60
data=replicate(10^6,reward(c,k))
summary(data[4,])   # nb of tosses=61 means < 8 nodes captured

# Figure 2.
plot(table(data[4,])[1:53], las=1, ylab='', xlab='tosses',
   main="Number of tosses until game (8, 60) captures 8 nodes",
   sub="11.6% of time 60 tosses capture fewer than 8 nodes")
prop.test(7790,10^6,1/128) # test P{ntoss=8)=2/2^8=1/2^7
prop.test(3896,10^6,1/256) # test P{ntoss=9)=2/2^9=1/2^8
prop.test(15632,10^6,1/64) # test P{ntoss=10)=2*8/2^10=1/2^6
nbtoss=data[4,]-(data[4,]==61) # nb tosses at game end (60 replaces 61)
summary(nbtoss); sd(nbtoss)

# Table 6.
summary(data[1,]); table(data[1,]) # nb of nodes captured

# Calculate E[%loss]
summary(data[3,]) # reward earned (plus refund)
price=25*c+k+7    # include overhead fee
rew=mean(data[3,]); loss=price-rew
print(c( round(c,0), round(k,0), round(price,2),
   round(rew,2),round(loss,2), round(100*loss/price,2) ))

# Figure 3. Kernel Density Plot
d <- density(data[3,])   # returns density data
plot(d, las=1, xlim=c(100,300), ylab='',
   main="Reward earned in 10^6 plays of game (8, 60)",
   sub="under 60% refund for 7 cents fee; price=267") # plots the results
abline(v=267); points(245.2,0, pch=3)   # reference price and mean reward

sum(data[3,]>price)/10^6 # prob of winning above the price
sum(data[3,]>0.90*price)/10^6 # prob of earning above the 10% loss threshold

### What if the overhead fees change?
# Simply revise the expected % loss
eloss=c(21.60, 21.80, 22.29, 31.67)
price=c(261, 267, 272, 369)
for (i in -5:8){print((eloss+i)/(price+i))}
```

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02–04, 2020; pp 117-131

NAFINDEX: Measure of Financial Inclusion based on NABARD All India Rural Financial Inclusion Survey (NAFIS) Data

K.J.S. Satyasai and Ashutosh Kumar
National Bank for Agriculture and Rural Development, Mumbai

Received: 28 July 2020; Revised: 12 August 2020; Accepted: 14 August 2020

Abstract

Financial inclusion (FI) is a multi-dimensional phenomenon unlike its pre-cursor concepts of access to credit or access to savings bank account which define financial inclusion in a narrow sense. Hence, measuring financial inclusion is complicated and requires developing a suitable index. Several scholars developed FI index mostly following methodology of Human Development Index. Sharma (2008), Mehrotra (2009), Ambarkhane *et al.* (2012), Gupte *et al.* (2012), Goel and Sharma (2017) are a few of them. CRISIL's Inclusix is an index at district level. Department of Financial Services (DFS), Ministry of Finance, Government of India also is constructing an index of financial inclusion to help monitoring over the years. These indices covered different dimensions. All these indices are constructed using data from secondary sources and measure supply side access. That is, they mainly represent the access an individual can have. Actual use of a financial service by an individual or household is not reflected in these indices. World Bank's Findex is one index developed based on survey data of individuals. We recommend that a FI index should manifest the actual usage of financial services in terms of breadth, intensity and extent of digital penetration. We, therefore, propose NAFINDEX, based on state-wise household level access to financial services based on data from NABARD All India Rural Financial Inclusion survey (NAFIS). Based on the field level data collected through NAFIS 2016-17, NAFINDEX has been constructed for different states of India. Three dimensions, traditional banking products, modern banking products, and payment systems, are considered for constructing the index. The average value of index at all India is 0.337. There are variations across states in the value of NAFINDEX and dimension indices. Interestingly, many states which saw lower penetration of traditional banking products as reflected in the respective dimension index, the modern banking products and payment mechanisms showed higher values. This underlines the direction for the future banking expansion in hither to unreached states.

Key words: Financial inclusion; Index; NAFINDEX

1. Introduction

Financial inclusion is increasingly being recognized world over as a key driver of economic growth and poverty alleviation. Apart from these benefits, financial inclusion (FI) imparts formal identity, provides access to the payments system and to savings safety net like deposit insurance, and enables the poor to receive direct benefit transferred in a leak-proof

Corresponding Author: K.J.S. Satyasai
Email: satyasaik@outlook.com

manner. At a macro level, greater FI is considered crucial for sustainable and inclusive socio economic growth for all. However, the FI is not an end in itself as it is only a means to reach higher levels of development. The potential for development in the various sectors of the economy such as primary sector (agriculture and allied sectors) and Micro, Small and Medium Enterprises (MSME) sector is enormous. However, the limited access to affordable financial services such as savings, loan, remittance and insurance services by the vast majority of the population in the rural areas and unorganised sector is believed to be acting as a major constraint to the growth impetus in these sectors. It is widely believed that access to affordable financial services - especially credit and insurance - enlarges livelihood opportunities and empowers the poor to take charge of their lives. Such empowerment also adds to social and political stability in the economy.

2. What is Financial Inclusion?

With an objective to extend such financial services to a sizeable majority of population particularly who continue to remain excluded from the opportunities and services provided by the financial sector, a Committee on Financial Inclusion (CFI) was set up by the Govt. of India under the Chairmanship of Dr. C. Rangarajan in 2006. This Committee on Financial Inclusion (Rangarajan, 2008) defined Financial Inclusion as:

> "*process of ensuring access to financial services and timely and adequate credit where needed by vulnerable groups such as weaker sections and low income groups at affordable costs."*

The report identified demand and supply sides of financial services and emphasised on improving human and physical resource endowments. Subsequently, Planning Commission, Govt. of India (2009) in a Report of the Committee on Financial Sector Reforms mentioned:

> "*Financial Inclusion is not only about credit but involves a wide range of Financial Services including savings accounts, insurance and remittance products. Moreover, credit provision without adequate measures to create livelihood opportunities and enhance credit absorption amongst poor will not yield desired results.*"

Emphasizing the importance of those financial products, the report recommended that access to safe and remunerative methods of savings, remittances, insurance and pension need to be expanded. They suggested crop insurance for farmers and health insurance for the poor as vulnerability reducing instruments.

The recent developments in banking technology have transformed banking from the traditional brick-and-mortar infrastructure like staffed branches to a system supplemented by other channels like automated teller machines (ATM), credit/debit cards, internet banking, online money transfers, mobile money, UPI, *etc*. The moot point, however, is that access to such technology is restricted only to certain segments of the society. Indeed, some trends, such as increasingly sophisticated customer segmentation technology – allowing, for example, more accurate targeting of certain sections of the market – have led to restricted access to financial services for some groups. There has been a growing divide, with an increased range of personal finance options for a segment of high and upper middle-income population on one hand and a significantly large section of the population who lack access to even the most basic banking services on the other. This is termed "financial exclusion". These people, particularly, those living on low incomes, cannot access mainstream financial

products such as bank accounts, credit, remittances and payment services, financial advisory services, insurance facilities, etc. The essence of financial inclusion is in trying to ensure that a range of appropriate financial services is available to every individual and enabling them to understand and access those services. Total financial inclusion or "Sampoorn Viteeyea Samaveshan" (SVS) envisaged to cover six broad areas, *viz*., (1) Ensuring every district with 1,000-5,000 households had access to banking services within 5 kms by March 2016; (2) Provide financial literacy; (3) Provide basic banking for all beneficiaries of government schemes by March 2016; (4) An overdraft of Rs. 5,000; (5) Micro insurance; and (6) Pension scheme for the unorganized sector (Mehta and Shah, 2014). Now the question arises, how do we get to know the level of financial inclusion of a population in a geography? This necessitates measurement of level of financial inclusion through an objective tool say financial inclusion index.

3. Why Financial Inclusion Should be Measured?

Financial inclusion is a key policy area and the central banks world over has an interest in it. Greater financial inclusion is essential for sustained economic welfare and for reducing poverty. It also supports economic, monetary and financial stability, by making saving and investment decisions more efficient, enhancing the effectiveness of monetary policy instruments, and facilitating the functioning of the economy (IFC Bulletin No. 38, Bank for International Settlements). In turn, economic stability helps to develop and strengthen a smoothly functioning financial system that can support financial inclusion. Therefore, it is very essential to measure financial inclusion objectively.

4. How to Measure Financial Inclusion?

Now, the question arises, how we measure financial inclusion. Financial inclusion is a multi-dimensional phenomenon and hence, its measurement remains inadequate if crucial dimensions are not included. Further, data on various indicators of financial inclusion raise important issues. Well-founded data frameworks are essential while developing financial services for the poor, in both formal and informal markets. Appropriate indicators in adequate number are a precondition for good financial inclusion measurement. They ensure that financial inclusion is properly assessed and that policies aimed at it are adequately implemented, monitored, and adjusted as required. Good statistics can also help to strike a fine balance between encouraging innovation and the growth of financial services on the one hand, and ensuring that financial stability is preserved, on the other.

5. Developing Indices of Financial Inclusion

Measurement of financial inclusion could be done through developing a suitable financial inclusion index (FII). A composite financial inclusion index, provides scope for multiple dimensions of financial inclusion to be reduced to a single one, making it simpler for analysts and policymakers alike. In general, such indices have no units and are constructed by making all the measured dimensions comparable. Such an index can be a valuable instrument to diagnose the financial inclusion situation for a specific geographic location, and to facilitate spatial and temporal comparisons. In turn, the index based on a set of identified key performance indicators can be established as a benchmark and used to identify best practices. Nevertheless, FII cannot be considered as a universal or exclusive policy tool. In fact, developing composite index is not a goal in itself. The quality of underlying data, however, is crucial.

Once we construct an index and measure financial inclusion, we can strive to achieve beyond its benchmark level. However, how do we measure financial inclusion depends on how we define it. In India, there have been several attempts to measure financial inclusion based on proportion of adult population having access to formal banking system, proportion of adults having bank account, bank accounts per 1000 adult population, ATMs per 1000 sq. km, population being serviced per branch, etc. However, all these are supply-side factors determining the status of financial inclusion. Similarly, there are demand-side factors such as income level, credit absorption capacity, financial awareness and literacy level of people, availability of livelihood opportunities in the area, etc. which determine the level of financial inclusion of a particular geography. Further, the quality of financial services being supplied and availed is another dimension determining the quality of financial inclusion. Thus, financial inclusion is a multi-dimensional phenomenon that can be better summarised by a composite index. The financial inclusion index (FII) should be such that: (i) it represents the true situation as far as possible; (ii) it is simple and easy to compute so that it is amenable for comparison; (iii) it should have meaningful bounds (say 0 and 1); and (iv) it should have monotonicity (higher values indicating higher level of financial inclusion).

There are 4 Steps to be followed in constructing FII. Develop a clear theoretical framework, to begin with, to have a sound basis for selecting the individual indicators of interest. Second, define precisely the data content, analysis, weighting and aggregation scheme for the selected indicators. Third, conduct sensitivity and robustness analysis to ensure quality. For instance, the indicator should not change dramatically if one of the individual components is excluded, or if a different scheme of weights is used. Lastly, create a framework for representing and communicating information provided by an FII, especially when making cross-country comparisons on the overall performance of the index, and the contribution of the various indicators to it.

6. Various Approaches to Construction of Index of FI

FI index is constructed using multi-dimensional framework representing demand and supply factors. Beck, Kunt and Peria (2007) make a clear distinction between (1) access and the possibility of use, and (2) the actual use of financial services. Honohan (2005) included contribution of financial access to household wellbeing and firm productivity on demand side, while product/service design (usefulness for the poor), cost and information barriers to access on supply side. And, they used following financial access indicators:

1. Payments: Inland and international remittances –crucial for the families dependent on migrant income.
2. Savings mobilization (deposit services).
3. Monitoring of users of funds (mechanisms for building credit worthiness)
4. Transforming Risk (Insurance etc.)

Sarma (2010) considered 3 dimensions: penetration (number of bank accounts per adult population), availability (number of banking outlets (branches and ATMs) per 1000 population), and usage (volume of credit and deposit as proportion of GDP). Arora (2010) considered outreach, ease, and cost. Outreach is measured by branch and ATM penetration per area and population. Ease is measured by (a) minimum amount to open saving account; (b) minimum amount to maintain saving account; and (c) number of documents required to open bank account. Cost includes fees for different services offered by the bank. Here again, all the dimensions are related with banks only, and other financial services are left out in the

process. Gupte, Venkataramani and Gupta (2012) considered four dimensions, namely, outreach, usage, ease, and cost of transaction, which are combined taking geometric mean. Kunt and Klapper (2012) also measured financial inclusion using four indicators, *viz*., (1) formal accounts; (2) savings behavior; (3) sources of borrowing, purposes of borrowing, and use of credit cards; and (4) use of insurance products. Rahman (2013) considered four indicators, namely, convenient accessibility, take up rate, responsible usage, and satisfaction level. All are assigned equal weights adding to unity. Yorulmz (2013) followed the method suggested by Sarma (2008) and used multi-dimensional approach. Normalized inverse Euclidean distance from the ideal point for three dimensions, access, availability, and usage are considered. A summary of various Researcher/Social Scientists and variables used by them are presented below (Table 1).

Table 1: Summary of various researcher/social scientists and variables used

Researchers	Variables used
Beck, Kunt and Peria (2007)	(1) Access and possibility of use; and (2) Actual use
Honohan (2005)	1. Payments, 2. Savings mobilization, 3. Monitoring of users of funds and 4. Transforming Risk
The Consultative Group to Assist the Poor (2009)	1. Savings, 2. Payments, 3. Credit and 4. Delivery
Sarma (2010)	1. Penetration, 2. Availability and 3. Usage
Arora (2010)	1. Outreach, 2. Ease and 3. Cost
Rahman (2013)	1. Convenient Accessibility, 2. Take Up Rate, 3. Responsible Usage and 4. Satisfaction level.
Gupte, Venkataramani and Gupta (2012)	1. Penetration, 2. Availability, 3. Usage 4. Ease and 5. Cost.
Kunt, Klapper (2012)	1. Formal accounts (a. the mechanics of the use, b. purpose, c. barriers, d. alternatives to formal accounts, e. penetration and f. receipt of payments), 2. Savings behavior (a. use of accounts, b. use of community-based savings methods and c. the prevalence of savings goals), 3. Sources of borrowing, purposes of borrowing, and use of credit cards and 4. Use of insurance products
Yorulmz (2013)	1. Access, 2. Availability and 3. Usage
Credit Rating and Information System of Indian Ltd. (CRISIL) (2013)	1. Branch penetration, 2. Credit
Amberkhane *et al*. (2014)	Drag factors besides demand, supply and infrastructure

Amberkhane *et al.* (2014) considered drag factors besides demand, supply, infrastructure dimensions to construct FI index. On demand and supply sides the indicators are related banks, NBFCs and insurance with 50% weight to banks and 25% weight for each of the other two. On supply side, the indicators are about spread of branches or outlets. On demand side, the indicators are related to deposits, loans, remittances, density of SHGs, insurance penetration, etc. Infrastructure indicators are on irrigation, transport, power, literacy, and health. Drag factors considered are population growth, law and order situation, and corruption. The values of all indicators are normalized to converts values of indicators between 0 and 1 using formula below:

$$d_i = (A_i - m)/(M - m) \quad (1)$$

where, for i^{th} State

d_i is the normalized value of indicator.
A_i is the actual value of indicator.
M is maximum value of indicator.
m is the minimum value of indicator.

Then FI index is the Euclidean distance measured by using displaced ideal (D.I.) method. Financial inclusion index for r^{th} state was obtained by inverse normalized distance from the ideal as given below:

$$1 - \frac{\sqrt{\{ (1 - s_r)^2 + (1 - d_r)^2 + (1 - i_r)^2 \}}}{\sqrt{3}} \quad (2)$$

The term $\sqrt{\{ (1 - s_r)^2 + (1 - d_r)^2 + (1 - i_r)^2 \}}$ in (2) above is the Euclidean distance of the point (s_r, d_r, i_r), *i.e.* position of rth State from the ideal (1, 1, 1) (This is the distance between the points S_r and P in Figure 1) and dividing it by $\sqrt{3}$ normalizes it, in three dimensional space. Further subtracting this from 1 (*i.e.* normalized distance of ideal point from the origin) gives inverse normalized distance, which is the index. This Index satisfies all intuitive properties of an index suggested by Nathan, Mishra, and Reddy (2008); namely Normalization, Anonymity, Monotony, Proximity, Uniformity, and Signalling.

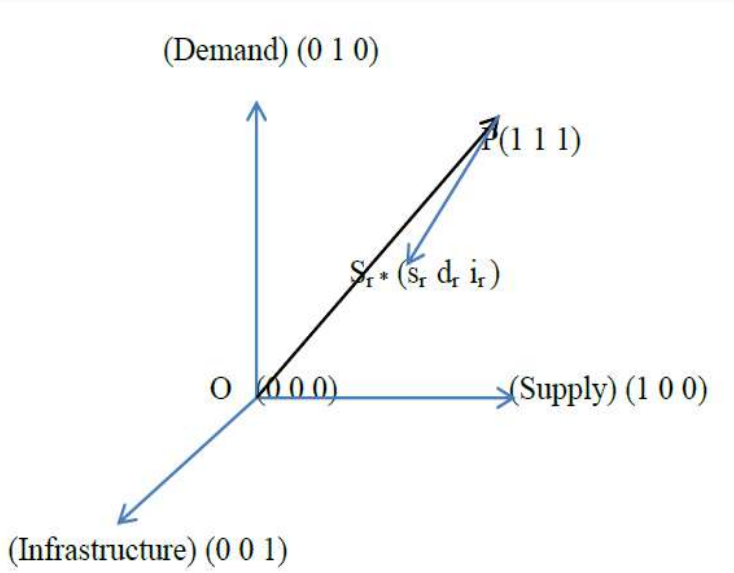


Figure 1: Diagrammatic representation of Euclidean distance method
Source: Nathan, Mishra and Reddy (2008)

The final Financial Inclusion Index is derived after applying the drag factors. Suppose D_r is the drag index for r^{th} state, the impact factor is taken as

$$1/(1+ D_r) \quad (3)$$

Comprehensive Final Inclusion Index is arrived at by multiplying impact factor as above with Financial Inclusion Index obtained earlier.

Points of difference with other methodologies

(1) Method is like UNDP approach following multidimensional approach
(2) This method uses Linear Averaging method for calculation of indices of Demand, Supply and Infrastructure dimensions, whereas Displaced Average Method is used for combining three indices.
(3) Signalling characteristics of D.I. Method is considered more suitable to proposed index as it indicates unique optimal path to reach higher value. Moreover, the signalling characteristic implies that an improvement in a dimension that has lower value is more important than an equivalent improvement in a dimension that has a higher value. Methodology suggested is flexible as any other relevant factor or indicator identified can be added to any of the dimensions or in drag.

However, it is questionable if an index to measure a phenomenon shall include explanatory factors such as infrastructure related or drag factors. For instance, higher corruption may inhibit inclusion because of which the financial inclusion index may be lower in a state. We consider it inappropriate to multiply with drag factor. Even if inclusion of infrastructure dimension is justified, some indicators therein are about physical aspects of infrastructure. Certain others are proxies. For instance, female literacy cannot be taken as proxy for educational infrastructure. This is a methodological issue.

Goel and Sharma (2017) have used following parameters for constructing index:

- Banking Penetration (D1) - demographic branch penetration *i.e*., number of accounts (deposits and loans) per 1,000 populations with different financial institutions (d1).
- Availability (D_2) of banking services – number of ATMs per 1,00,000 population (d_2).
- Number of bank branches per 1,00,000 population (d_3),
- Number of ATMs per 1,000 sq. km (d_4)
- Number of scheduled commercial banks per 1,000 sq. km (d_5).
- Access to Insurance (D3) – number of life insurance (LIC) offices (d_6).

The indicators are normalised, and indices are constructed using weights. The FII is measured as the simple average of two indices, X1 and X2, measured, respectively, based on distance from zero, and the ideal point, *w*, for each indicator.

$$d_d = W_d * (A_d - m_d)/(M_d - m_d) \quad (4)$$

where,

w_d = Weight attached to the dimension *d*, $1 \geq w_d \geq 0$;
A_d = Actual value of dimension *d*;
m_d = Minimum value of dimension *d*;
M_d = Maximum value of dimension *d*;
d_d = Dimensions of financial inclusion *d*.

$$X_1 = \frac{\sqrt{d_1^2+d_2^2+d_3^2+\cdots+d_n^2}}{\sqrt{w_1^2+w_2^2+w_3^2+\cdots+w_n^2}} \tag{5}$$

$$X_2 = 1 - \frac{\sqrt{(w_1-d_1)^2+(w_2-d_2)^2+(w_3-d_3)^2+\cdots+(w_n-d_n)^2}}{\sqrt{w_1^2+w_2^2+w_3^2+\cdots+w_n^2}} \tag{6}$$

$$\text{FII} = \frac{1}{2}\,(\text{X}_1 + \text{X}_2) \tag{7}$$

Depending on the value of FII, the time period under study has been categorized as

1. $0 \leq FII \leq 0.4$; indicates low financial inclusion, LFI.
2. $0.4 < FII \leq 0.6$; indicates medium financial inclusion, MFI.
3. $0.6 < FII \leq 1$; indicates high financial inclusion, HFI.

From the computation of *FII* across a time period of twelve years, India can be categorized under low financial inclusion during 2005 to 2012. During this time period, the value of *FII* ranged between 0 - 0.4. During 2013, condition of financial inclusion improved, and India fell under medium financial inclusion with *FII* from 0.4 to 0.6. The objective of inclusive growth was achieved further during 2014-2015 and India fall under high financial inclusion range in this time period. The value of *FII* ranged from 0.6 to 1. Unlike earlier studies where only indices such as banking penetration, availability of banking services and usage of banking system were used, Goel and Sharma (2017), included indicators such as access to savings and access to insurance also. Also, *FII* is constructed for a longer period of twelve years.

Sriram and Sundaram (2015) measured *FII* using 3 dimensions, *viz*., access, availability, and usage. These dimensions are assigned weights of 1 for access and 0.5 each for the remaining two. Access is measured through number of bank accounts in the area, availability, through number of access points (branches, ATMs, banking correspondents) in the area, and, usage, through number of accounts (savings, deposits, loan and credit) held by respondents. These dimensions are combined to compute FII as the difference of square root of Euclidean distance with reference to the ideal (*i.e.* weight) from unity. The formula is as below:

$$FII = 1 - sqrt\{[(1 - P_i)^2 + (0.5 - A_i)^2 + (0.5 - U_i)^2] / 1.5\} \tag{8}$$

where,

FII = Financial Inclusion Index
P_i = Access
A_i = Availability
U_i = Usage

Sarma (2008) has computed the values of IFI for 54 countries using the three basic dimensions of financial inclusion–accessibility, availability and usage of banking services. Accessibility has been measured by the penetration of the banking system proxied by the number of bank A/C per 1000 population. Availability has been measured by the number of bank branches and number of ATMs per 100,000 people. The proxy used for the usage dimension is the volume of credit plus deposit relative to the GDP. Gupte et al. (2012) considered 4 dimensions, outreach, usage, ease and cost. Outreach has two sub-dimensions, penetration, and availability. Ease too has two, directly related, and inversely related. Total 5 indicators were included, and no indicators were considered for ease and cost dimensions.

Demirguc-Kunt and Klapper (2012) delineated the methodology and computing Global Findex which measured the use of financial services (demand) as opposed to the access (supply) to them. Several indicators were used for computing Findex. The first set of indicators focuses on formal accounts; the mechanics of the use of these accounts (frequency of use, mode of access); the purpose of these accounts (personal or business, receipt of payments from work, government, or family); barriers to account use; and alternatives to formal accounts (mobile money). The account penetration indicator measures individual or joint ownership of formal accounts—accounts at a formal financial institution such as a bank, credit union, co-operative, post office, or microfinance institution. It includes those who report having a debit or ATM card tied to an account. Indicators relating to the receipt of payments measure the use of formal accounts to receive wages (payments for work or from selling goods), payments or money from the government, and family remittances (money from family members living elsewhere). The second set of indicators focuses on savings behaviour. This relates to the use of accounts, as people often save at formal financial institutions. Other indicators explore the use of community-based savings methods and the prevalence of savings goals. The third set focuses on sources of borrowing (formal and informal); purposes of borrowing (mortgage, emergency or health purposes, and the like); and use of credit cards. The fourth focuses on use of insurance products for health care and agriculture.

Most of the above are at country level and one or two are there for states. Credit Rating and Information Services of Indian Ltd (CRISIL) (2013) calculated index, Inclusix, available at district level. It considered three dimensions, namely branch penetration, credit penetration and deposit penetration. However, it has serious limitations in terms of coverage of dimensions and indicators. It is only in terms of number of accounts and not amount. Further, it covers scheduled commercial banks data only.

Mehrotra *et al*. (2009) also built up an index for financial inclusion using similar kind of aggregate indicators like number of rural offices, number of rural deposit accounts, volume of rural deposit and credit from banking data for sixteen major states of India. Here also, Financial Inclusion Index is estimated at the district level in India.

8. Issues/Limitations with Existing Measures

Whatever existing measures were used so far, they had some issue or the other to be resolved/improved upon. Some of them are as under:

- Mostly based on secondary and administrative data though ‘demand side’ and ‘usage’ based indicators are incorporated. Exceptions like Findex exist

- Large scale survey-based data not used. Again, Findex is an exception. There too, the sample for any individual country is limited and representative at country level only.
- Mix up between indicators and their drivers is an issue. The question is should the financial inclusion index should combine indicators of access and/or use as well as factors influencing these indicators, together.
- Many dimensions and large number of indicators are available. Should there be any standardization and consensus on indicators?

9. Index Based on NAFIS data

To take care of some of the limitations of the existing measures, we have tried to build an index, NAFINDEX, based on NABARD Rural Financial Inclusion Survey (**NAFIS)** 2016-17 data. NAFIS was undertaken by NABARD pan-India during 2016-17 covering both financial and livelihood aspects of 40000 sample households across 29 states. The survey covered all aspects of financial inclusion from a household perspective, *viz*., savings, borrowing, investment, remittances and payments, and insurance. Besides, the survey also covered financial literacy and experience of households with payment mechanisms.

The index is generated at all India and state-level based on the field level data collected from households. For constructing NAFINDEX, we covered three dimensions – traditional banking products (*T*), modern banking services (*M*), and payment mechanisms (P). Traditional banking products covered savings, investments, loans, and others (insurance & pension); modern banking services included usage level of ATMs, internet banking, and mobile banking; and, payment mechanisms covered usage of cheque and credit/debit card as well as ease of using them. The indicators used and weights assigned for this Index are given in Table 2.

Table 2: Indicators used for constructing NAFINDEX

Dimension	Service/sub-dimension	Indicator	Symbol of normalised indicator	Weight
Traditional Banking Products	Savings	% households that made any saving in the last 1 yr	*T*11	0.125
		mean savings (with all agencies) per household in the last 1 year [base: saver household who reported their saving amount]	*T*12	0.125
	Investment	% households that made any investment in the last one year	*T*21	0.125
		mean investment in all assets for household reporting any investment in the last one year	*T*22	0.125
	Loans	incidence of indebtedness	*T*31	0.125
		average outstanding debt per indebted household (rs.)	*T*32	0.125
	Others	% households with at least one member having any insurance	*T*41	0.125
		% households having pension	*T*42	0.125

Modern Banking services	usage	% ATM users	*M*11	0.167
		% internet banking users	*M*12	0.167
		% mobile banking users	*M*13	0.167
	Ease in using	% users having ease of using ATM	*M*21	0.167
		% users having ease of using internet banking	*M*22	0.167
		% users having ease of using mobile banking	*M*23	0.167
Payment Mechanisms	usage	% users of cheque	*P*11	0.25
		% users of debit/credit card	*P*12	0.25
	Ease in using	% users having ease in using cheque	*P*21	0.25
		% users having ease in using debit/credit card	*P*22	0.25

The indicators are combined to form dimension indices which are in turn combined into NAFINDEX. The values of all indicators are normalized to scale down values of indicators between 0 and 1 using formula at (1). Individual dimension indices are computed as below:

$T_n = \sum(W_{ij} * T_{ij})$
$M_n = \sum(W_{ij} * M_{ij})$
$P_n = \sum(W_{ij} * P_{ij})$

where,
T_n is the dimension index for traditional banking products for nth state;
M_n is the dimension index for modern banking services; and,
P_n is the dimension index for payment mechanisms.

Subscripts *i* and *j* stand for sub-dimension and indicator, respectively.

$\text{NAFINDEX} = \sqrt[3]{(T_n * M_n * P_n)}$

We have fitted a linear regression model to understand the explanatory factors for variation of NAFINDEX across states.

Dependent variable: NAFINDEX = Financial Inclusion Index

Independent variables:

Mf-membership = index of per cent HH having membership with microfinance institutions
% trained = proportion of HH received training
income index = index of HH income
% institutional loan = share of institutional loan in total

The regression is worked for agricultural households, non-agricultural households and all rural households.

10. State-wise NAFINDEX Values

The state wise Index of *FI* calculated based on NAFIS data are given in Table 3. The NAFINDEX for all India is 0.337 in a scale of 0 to 1. The value of the index for banking products dimension is 0.307. The value for the payment mechanisms dimension is the highest at 0.370 followed by 0.345 for banking services. Punjab, Kerala, and Karnataka ranked top three states in banking products dimension while Bihar, Chhattisgarhi, and Madhya Pradesh are at the last three positions. Goa, Manipur, and Nagaland are at the top for banking services dimension and Jharkhand, Madhya Pradesh, and Meghalaya are at the bottom. For the payment mechanisms dimension, top ranking states are Goa, Assam, Manipur, and Tripura while Uttarakhand, Rajasthan, and Chhattisgarhi are at the bottom.

Table 3: NAFINDEX values for different states and all India

State	Banking products	Rank	Banking Services	Rank	Payment mechanism	Rank	NAFINDEX	Rank
Goa	0.472	5	0.946	1	0.761	1	0.600	1
Punjab	0.617	1	0.473	12	0.383	19	0.486	2
Karnataka	0.533	3	0.430	14	0.438	13	0.483	3
Telangana	0.482	4	0.563	8	0.478	8	0.480	4
Andhra Pradesh	0.424	7	0.703	4	0.529	5	0.473	5
Kerala	0.609	2	0.446	13	0.362	21	0.470	6
Manipur	0.385	12	0.791	2	0.558	3	0.464	7
Tripura	0.366	14	0.523	10	0.558	3	0.452	8
Jammu & Kashmir	0.420	8	0.427	15	0.450	12	0.435	9
Odisha	0.379	13	0.381	24	0.477	9	0.425	10
Haryana	0.409	10	0.328	26	0.423	14	0.416	11
Mizoram	0.322	16	0.580	6	0.476	10	0.392	12
Assam	0.237	21	0.482	11	0.625	2	0.385	13
Himachal Pradesh	0.460	6	0.565	7	0.310	23	0.377	14
Meghalaya	0.318	17	0.240	29	0.403	17	0.358	15
Arunachal Pradesh	0.337	15	0.353	25	0.374	20	0.355	16
Sikkim	0.253	20	0.678	5	0.486	7	0.351	17
Nagaland	0.318	17	0.734	3	0.325	22	0.322	18
West Bengal	0.202	25	0.419	16	0.507	6	0.320	19
Maharashtra	0.224	22	0.416	18	0.416	16	0.305	20
Jharkhand	0.200	26	0.321	27	0.451	11	0.301	21
Gujarat	0.215	24	0.531	9	0.420	15	0.300	22
Uttar Pradesh	0.217	23	0.417	17	0.397	18	0.294	23
Tamil Nadu	0.387	11	0.404	20	0.208	25	0.284	24
Uttarakhand	0.420	8	0.401	21	0.189	27	0.281	25
Bihar	0.198	27	0.387	23	0.264	24	0.229	26

State	Banking products	Rank	Banking Services	Rank	Payment mechanism	Rank	NAFINDEX	Rank
Rajasthan	0.276	19	0.398	22	0.178	28	0.222	27
Madhya Pradesh	0.141	29	0.266	28	0.195	26	0.166	28
Chhattisgarh	0.160	28	0.411	19	0.055	29	0.094	29
All India	0.307		0.345		0.370		0.337	

Table 4 gives results of linear regression model estimated to explain the variation in NAFINDEX. Of the four variables included in the model two variables Mf-membership and income index are significant for agricultural, non-agricultural and overall rural households. Proportion of households trained has significant effect on NAFINDEX. That is, states where the penetration of microfinancing institution is higher and where households reported higher income, the financial inclusion index is also higher. The NAFINDEX among non-agricultural households is higher in states with higher proportion of households with trained households. The explanatory of power the regression is 48 to 55 per cent and is statistically significant.

Table 4: Factors explaining variation in NAFINDEX

Variable/description	Particular	Agri HH	Non-Ag HH	Rural HH
Constant	*Coefficient*	**0.252247** ***	**0.140434** ***	**0.211313** ***
	std error	0.0518705	0.0486343	0.046158
	p - value	<0.0001	0.0081	0.0001
Mf-membership *(index of per cent HH having membership with microfinance institutions)*	*Coefficient*	**0.204269** ***	**0.232292** ***	**0.219316** ***
	std error	0.0573099	0.0708419	0.0646751
	p - value	0.0016	0.0032	0.0024
% trained (*proportion of HH received training*)	*Coefficient*	**−0.00254901**	**0.186124** ***	**0.0588672**
	std error	0.0683906	0.0720409	0.0693977
	p - value	0.9706	0.0163	0.4047
income index (*index of HH income*)	*Coefficient*	**0.243671** ***	**0.165032** **	**0.294322** ***
	std error	0.0768699	0.0841741	0.0831651
	p - value	0.0041	0.0616	0.0017
% institutional loan (*share of institutional loan in total*)	*Coefficient*	**−0.0352766**	**0.102054**	**−0.0385928**
	std error	0.0826278	0.073567	0.0792173
	p - value	0.6732	0.1781	0.6306
Note: ***, ** significant at 1% and 5%, respectively				
Mean dependent var		0.346811	0.365558	0.362676
Sum squared residual		0.136775	0.208367	0.168062
R-squared		0.478384	0.546915	0.499992
F(4, 24)		5.502705	7.242553	5.999799
Log-likelihood		36.52308	30.41914	33.5362

11. Conclusion

Based on the field level data collected through NAFIS 2016-17, NAFINDEX has been constructed for different states of India. Three dimensions, traditional banking products, modern banking products, and payment systems, are considered for constructing the index. The average value of index at all India is 0.337. There are variations across states in the value of NAFINDEX and dimension indices. Interestingly, many states which saw lower penetration of traditional banking products as reflected in the respective dimension index, the modern banking products and payment mechanisms showed higher values. This underlines the direction for the future banking expansion in hither to unreached states.

Acknowledgements

The views in this paper are those of the authors alone. We acknowledge the suggestions of the participants in the Conference and the reviewer on the earlier draft.

References

Ambarkhane, D., Singh, Ardhendu S., Venkataramani, B. (2014). Developing a comprehensive financial inclusion index. Symbiosis School of Banking and Finance, Symbiosis International University. Retrieved from http://ssrn.com/abstract=2485774 or http://dx.doi.org/10.2139/ssrn.24857749 (accessed on 15 June 2014).

Ambarkhane, D., Singh, A. S. and Venkitaraman, B. (2016). Measuring comprehensive financial inclusion index of Indian states. *Indian Journal of Rural Management*, IRMA, Anand.

Arora, R. (2010). Measuring financial access. *Discussion Paper - Economics.* Griffith University.

Beck, T., Kunt, A. D. and Peria, M. S. M. (2007). Banking services for everyone? Barriers to bank access and use around the world. *Working Paper Series,* Policy Research Working Paper, World Bank.

Credit Rating and Information System of India Ltd. (2013). CRISIL *Inclusix.* Mumbai.

Goel, S. and Sharma, R. (2017). Developing a financial inclusion index for India. *Procedia Computer Science*, **122**, 949-956.

Govt. of India (2009). *Report of the Committee on Financial Sector Reforms.* Planning Commission, New Delhi.

Gupte, R., Venkataramani, B. and Gupta, D. (2012). Computation of financial inclusion index for India. *International Journal of Procedia - Social and Behavioral Sciences*, **37**, 133 –149.

Honohan, P. (2005). *Measuring Microfinance Access: Building on Existing Cross-Country Data*. The World Bank.

Kunt, A. D. and Klapper, L. (2012). Measuring financial inclusion: the global findex database. *Policy Research Working Paper 6025*, World Bank, Washington, DC.

Mehrotra, N., Puhazhendhi, V., Nair G. G. and Sahoo, B. B. (2009). Financial Inclusion: an overview. *Occasional Paper No.* **48**, NABARD, Mumbai.

Mehta, S. and Shah, P. (2014). Modi government woos India's unbanked population, to offer OD bonanza for financial inclusion. *The Economic Times, Jul*, *9*, 2014.

NABARD (2018). *NABARD All India Financial Inclusion Survey (NAFIS)* 2016-17.

Nathan, H. S. K., Mishra, S. and Reddy, S. (2008). An alternative approach to measure HDI. *Working Paper*, IGIDR, Mumbai.

Rahman, Z. A. (2013). Developing a financial inclusion index. *Central Bank. Retrieved from www.Centralbanking.com/digital_assets/6715/CB23. 4 May2013 Financial inclusion Rahman New.pdf (accessed on 10 June 2014)*.

Rangarajan, C. (2008). Report of the committee on financial inclusion. *Ministry of Finance, Government of India*.

Sarma, M. (2008). Index of financial inclusion. *Working Paper* 215. Indian Council for Research on International Economic Relations (ICRIER), New Delhi.

Sarma, M. and Pais, J. (2008). Financial inclusion and development: a cross country analysis. ICRIER, New Delhi.

Sarma, M. (2010). Index of financial inclusion. *Discussion Paper 10-05*, Centre for International Trade and Development, School of International Studies, Jawahar Lal University, New Delhi.

Satyasai K. J. S. (2011). Innovation for financial inclusion of rural poor and women in India. In: (Ed.) Suresh Pal, *Agriculture for Inclusive Growth,* IARI, New Delhi (164-180)

Singh, N. (2017). Financial inclusion: Concepts, issues and policies for India. *Issues and Policies for India (July 4, 2017)*.

Sriram, M. and Sundaram, N. (2015). Financial inclusion index: a customized regional model with reference to economically most backward districts of Tamil Nadu, India. *Mediterranean Journal of Social Sciences*, **6**(**6**), 209.

The Consultative Group To Assist the Poor (2009). Measuring access to financial services around the world, World Bank.

The Global Findex Database (2017). *Measuring Financial Inclusion and the Fintech Revolution.* The World Bank Group, Washington.

Yorulmz, R. (2013). Construction of a regional financial inclusion index in Turkey. University of Sheffield, Department of Economics.

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 133-144

Large Scale Assessment Survey to Evaluate Learning Level of Students

Vishal D. Pajankar
National Council of Education Research and Training, New Delhi, India

Received: 17 July 2020; Revised: 12 August 2020; Accepted: 16 August 2020

Abstract

Many countries have started assessing their students through different assessment surveys to know the learning level of students, to know what they can do and what they know? In India, Large scale educational assessment survey named, National Achievement Survey (NAS) is being conducted at grades 3, 5, and 8 of the elementary stage under the flagship program of the Government of India. In 2017, the first time the NAS was conducted for grades 3, 5, and 8 on a single day i.e. November 13, 2017, in all 701 districts of 36 states and Union territories in India. About 2.2 million students from 1.2 lacs schools across the country were participated (NCERT, 2020). Schools in each distract were sampled using a stratified sampling procedure and section and students by NCERT designed a random sampling process. In this round, the district was the reporting unit of the study. NAS test items were constructed based on different competencies at different grades and linked them with learning outcomes instead of content-based. In this paper, we discussed how the NAS was implemented in the country and how the students' performed in the NAS. This paper highlights the performance of the students in different grades and different subjects. The comparisons between gender, school location, and school management are also discussed in the paper.

Key words: National achievement survey; IRT; Learning outcomes; Test items; Cohen's D.

1. Introduction

Large scale Assessment in education is one such tool that obtains information to assess the health of education systems and try to know whether the students meet curricular standards. Since the mid-1980s, the interest of measuring, comparing, and monitoring educational standards is growing in almost all countries. So, in the global countries have started assessing their students through different assessment surveys to know the learning level of students, to know what they can do and what they know? Many countries are taking participating in intercountry assessment surveys (large scale); for example PISA (Programme for International Students Assessment), TIMSS (the Trends in International Mathematics and Science Study), and PIRLS (the Progress in International Reading Literacy Study), *etc*. Some countries are conducting their assessment surveys to judge educational standards against national expectations. (NCERT, 2015a). The Sustainable Development Goal for education (SDG 4) is also called for an increased focus on learning outcomes, with five of the ten targets highlighting the learning skills and outcomes of children and adults (UNESCO, 2018).

In India, the large scale assessment survey is being conducted periodically since the seventies under different schemes and program of Govt. of India. In 2001, it is named as National Achievement Survey (NAS). The main aim of NAS is to provide reliable information about the achievement of students in the different grades of education in government and government aided schools. The NAS report gives a national and state-level picture, rather than scores for individual students or schools. The purpose of this survey is to obtain an overall picture of what students in specific classes know and can do and to use these findings to identify gaps and diagnose areas that need improvement. This information can then be used to impact policies and interventions for improving children learning in the country.

The data from National Achievement Survey gives the policymakers, curriculum specialists, researchers, and, most importantly, school principals and teachers a 'snapshot' of what students are achieving in key subjects at a particular point in time. By repeating such measurements at regular intervals, trends can be explored providing an invaluable perspective from which to consider educational reform and improvement. It does not give scores to individual students or schools (Pajankar, 2019).

2. History of National Achievement Survey (NAS)

India has a long history of conduct achievement (or assessment) surveys. The first notable survey was conducted by Kulkarni (1970) to know the achievement of students at different stages of school education in Mathematics. Another important study was untaken by Dave (1988) in NCERT under the project of Primary Education Curriculum Renewal (PECR) in 22 States at the primary stage in Language, Mathematics, and Environmental Studies. The third major study at the primary stage in Language and Mathematics was initiated by Shukla (1990) in NCERT and was completed in 1994 in 22 States and UTs. This was followed by district-specific surveys in primary classes under the District Primary Education Programme (DPEP) as the baseline, midterm, and terminal cycles (Dave, 1988 and NCERT, 2011).

Under the Sarva Shiksha Abhiyan (SSA) flagship program of Govt. of India, the survey is restructured and then named as National Achievement Survey (NAS). From 2001, NAS has been conducted in the different cycles in the country. The level was class III, class V, and class VIII in 2-3 years. Till 2017, 4 cycles of each grade have been conducted. In these cycles, the reporting unit was State and districts were sampled from each state. So NAS reported the learning level status of state only (Pajankar, 2019). The time-line of the conduct of NAS is given in table 1.

In 2017, the structure and nature of NAS was again changed. In 2017, the test was conducted on a single day November 13, 2017, in class III, V, and VIII. For NAS 2017, schools sample drawn through the Population Proportionate to Size (PPS) procedure includes nearly 2.2 million children from 1,10,000 schools spread across all districts in India. The salient features of this NAS 2017 were as below (NCERT, 2017):

- National Achievement Survey was linked to the learning outcomes;
- Assessment was being conducted for classes 3, 5 and 8 on a single day across the country;
- District was the unit for reporting;
- Automated reports were generated at the district level;
- Pedagogical interventions were provided in the same academic year.

National Achievement Survey 2017 was achieved by administering standardized tests to students of classes III, V, and VIII. NAS 2017 has contributed several new elements and gave remarkable momentum to the development of competency-based assessment. One of the main virtues of NAS 2017 is that it is embedded in an extremely rich system of background variables. The results help to accurately discover the students' performance in different learning outcomes vis-à-vis the contextual variables. The synthesis of the results of the national level provides a rich repository of evidence for developing and designing the future course of action for the Indian education system (Pajankar, 2019). Internationally accepted technical standards and practices were being adhered to while planning, designing, and implementing the NAS to ensure its robustness and sustainability (NCERT, 2017).

Table 1: The time period of the conduct of National Achievement Survey

Survey Cycle	**Class V**	**Class VIII**	**Class III**
Cycle I	2001-02	2002-03	2003-04
Cycle II	2005-06	2007-08	2007-08
Cycle III	2009-11	2010-13	2012-13
Cycle IV	2013-15	2014-16	2014-16
Subjects Tested	Mathematics Language Environmental-Studies	Mathematics Language Science Social Science	Mathematics Language Environmental-Studies
Background Questionnaires	Pupil Questionnaire (PQ)	Teacher Questionnaire (TQ)	School Questionnaire (SQ)

3. Sampling Design

In this NAS 2017, the target population was the students from classes III, V, and VIII from Government and Government aided schools. In earlier surveys, the state was reporting unit and 'n' number of districts was sampled. In NAS 2017, the district was reporting unit. So, all districts from 36 states/Union Territories were taken into consideration. 703 districts were listed in the sampling frame. But due to political reasons, 2 districts could not participate in the survey. Finally, the NAS was conducted in 701 districts in the country.

Sampling was done in three stages; first stage: school sampling; second stage: section sampling and third stage: students sapling. At the first stage, schools from each district were sampled using the Probability Proportional to Size (PPS) sampling procedure. Two types of stratification were used namely (i) explicit stratification (for the district) and (ii) implicit stratification (for block, area, school management, type of schools and medium of instructions). This exercise was conducted for each class III, V, and VIII, separately. The target sample size was 61 schools for class III and V and 51 schools for class VIII. Two additional lists of sampled schools, parallel to the original sampled list was prepared to replace the schools only if the original school does not exist or enrolment less than 5 or destroyed in natural calamities/ Naxal attack.

A second stage, one section of class III, V, and VIII was selected from each sampled school. If the sampled school has the only section of either class then the section was considered as selected. In the third stage, 30 students were selected from each selected section of sampled schools. The maximum target of students was 30. If in a class total attendance is less than 30 then all students were considered. The selection of sections in sampled schools and students from the selected section was conducted by a simple random procedure designed by the NCERT team.

Accuracy of a sample statistic as an estimate of an unknown population parameter is assessed through standard errors. Standard errors are computed through the following formula:

$$\sigma_{(\hat{\theta})} = \sqrt{\sigma^2_{(\hat{\theta})}}.$$

This formula assumes the use of Simple Random Sampling (SRS). Large scale assessments including NAS use complex sampling procedures. To ensure unbiased estimates of Standard Errors (SE) are generated, SEs are computed using the Jackknife Repeated Replication technique (JRR) for ability θ (NCERT, 2020).

4. Methodology

In this National Achievement survey 2017 (NAS 2017), schools were sampled using probability proportional to size procedure. In this process about 1,10,000 government and government aided schools were sampled from 701 districts of all states and union territories in the country. On November 13, 2017; the test was administered in all over the country About 2.2 million students from these sampled schools participated in NAS 2017. The students of classes III, V, and VIII were tested in different subjects through two sets of test booklets as shown below in Table 2.

Table 2: Class wise test booklets along with subjects and number of items

Class	Subjects	No. of items	Total items in Test Booklet	Number of Sets
III	Language* Mathematics Environmental Science	15 15 15	45	Two Sets 31 & 32
V	Language* Mathematics Environmental Science	15 15 15	45	Two Sets 51 & 52
VIII	Language* Mathematics Science Social Science	15 15 15 15	60	Two Sets 81 & 82

** Language used in a state as a local or regional language*

In NAS 2017, all subjects were tested through two test booklets for each class. Each subject had 15 items. So, 45 items in class III and class V and 60 items in class VIII. To

maximize the coverage of the test, two sets of test booklets were constructed. To establish a link between test booklets and to put them in a common platform, 05 items of each subject were common in both sets. Items were constructed with different competencies and linked with the learning outcomes (LOs) developed by the NCERT at elementary stage (classes I to VIII) in 2016. The items were piloted and removed all non-functioning items before finalizing the test booklets. The test booklets were then translated into 20 modern Indian languages. For quality check, translation was verified by experts and by back translation activity (with limited items).

Other booklets: three questionnaires *i.e.* Student Questionnaire (PQ), Teacher Questionnaire (TQ), and School Questionnaire (SQ) were also prepared for this NAS 2017. The objective of these questionnaires was to analyse the associations between the achievement and the background variables.

4.1. Analysis procedure

The data was collected through two sources; one was test booklets i.e. achievement data and another was questionnaires *i.e.* information of background factors collected from school heads, teachers, and pupils through interview mode. The achievement data of the students was analysed by classical test theory and item response theory. However, questionnaire data were analysed by classical test theory. Two different approaches were used to analyse and for reporting at a different level.

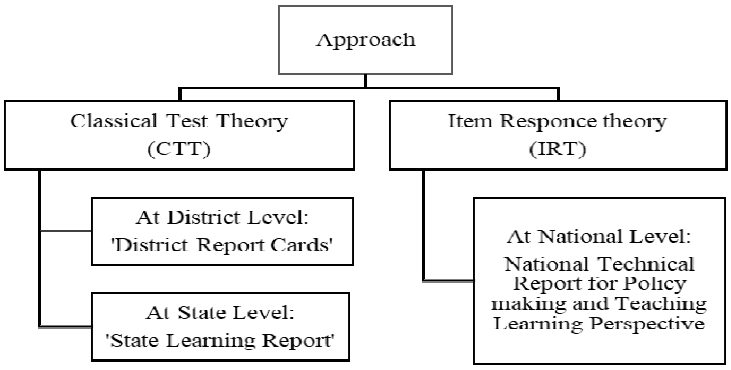


Figure 1: Different approaches used for the analysis at district level, state level and national level

Under the classical test theory, raw percentages of correct responses were used to measure students' abilities and item difficulties. With the classical test theory, the district report cards (DRCs) and state learning reports (SLRs) were generated for all districts and states/union territories with a record period of 2.5 months and 5 months respectively from the date of NAS 2017 administered. The district report cards were prepared in such a way that it can be easily read and understood by a layperson. The main objective behind it was that every parent/guardian can understand the learning level of his/her child. 10 independent report cards for each subject of each class of each district were generated. All DRCs and SLRs are available in the public domain at the NCERT web portal at link http://www.ncert.nic.in/programmes/NAS/NAS.html.

Table 3: List of district report cards for a district

Sl. No.	Class	Subject	No. of pages*
1	Class III	Language	2
2	Class III	Mathematics	2
3	Class III	Environmental Sciences (EVS)	2
4	Class V	Language	2
5	Class V	Mathematics	2
6	Class V	Environmental Sciences (EVS)	2
7	Class VIII	Language	2
8	Class VIII	Mathematics	2
9	Class VIII	Sciences (Sci)	2
10	Class VIII	Social Sciences (SSc)	2

**Few reports may have 3 pages*

Item Response Theory (IRT) approach was used in NAS 2017. Major large scale assessment studies conducted at international levels such as PISA (Programme for International Students Assessment), TIMSS (the Trends in International Mathematics and Science Study) and PIRLS (the Progress in International Reading Literacy Study), etc., are also using Item Response Theory (IRT). IRT measures the learning ability of students by calculating the probability of a student to respond to an item correctly. IRT analysis places students and test items on the same numerical scale and this helps us to create meaningful 'maps' of item difficulties and student abilities. In IRT, the difficulty of an item does not depend on the group of test-takers. Multiple test booklets can be used in IRT to increase the measurement points in any subject and the booklets can also be linked (NCERT, 2020).

IRT uses mathematical models that ensure the statistical connection between the difficulty level of the test item, the ability of the student, and the probability of that student being successful on a particular item. For example, students with higher ability scores are more likely to succeed on any item than their peers of lower ability Therefore, analysis in IRT is more complex than traditional methods like CTT. IRT uses the concept of an Item Characteristic Curve (ICC) to show the relationship between students' ability and performance on an item (NCERT, 2015b).

The two-parameter model (2PL) to the items was applied to analyse the data. The 2PL model associates student's ability to both item difficulty and item discrimination. The model includes difficulty (b) and discrimination (a) of the item. The expression for P_{ij}, the probability of the i^{th} examinee, ability θ_i, being successful on the j^{th} item, difficulty b_j is given by Thissen and Wainer (2001)

$$P_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + exp[a_j(\theta_i - b_j)]}$$

$$P_{ij} = \frac{1}{1 + exp[-a_j(\theta_i - b_j)]}$$

where, P_{ij} is the probability of the i^{th} examinee, ability θ_i, being successful on the j^{th} item, difficulty b_j.

Test Reliability was estimated using the following formula

$$\bar{\rho} = \frac{\sigma_\theta^2 - \sigma_e^2}{\sigma_\theta^2}.$$

where, σ_θ^2 is the variance of the test score scale in the sample, and σ_e^2 is the mean error variance of scores. The values of both were estimated from BILOG software (Zimowski *et. al.*, 1996). At item and tests level, quality of achievement indices (or instruments) such as Item difficulty indices (*p*-value), Item discrimination indices (DI), Options analysis or Distractor analysis (DE), Differential Item Functioning (DIF) and Test reliability; were conducted using Classical Test Theory (CTT) and Item Response Theory (IRT) approach.

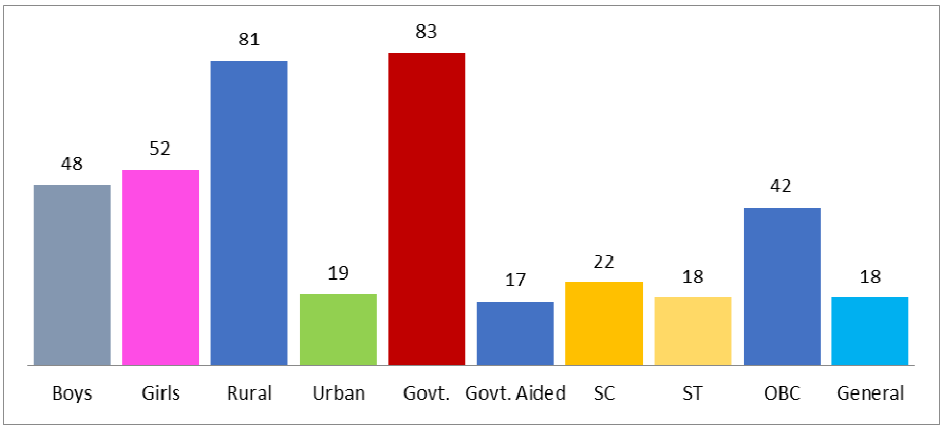
4.2. The reporting scale

IRT approach uses scale scores for reporting the results. In scaling, raw scores were transformed into a new set of scores by using either linear or nonlinear methods. The converted scores called Scaled Scores. The IRT scores were initially generated in the logit metrics, and then they were linearly converted into a scale that facilitates score interpretation. The reporting scale was set to the range of 100 - 500 with a mean of 300 and standard deviation of 50. Thus, the linear transformation from ability estimates expressed on the logit scale to the reporting scale scores was conducted using the expression: *Scale Score = Logit Score* * 50 + 300. Scaled scores were computed by statistically adjusting and converting raw scores into a common scale to account for differences in difficulty across different test forms (NCERT 2020 and 2014a).

5. Major Findings and Discussion

The National Achievement Survey (NAS) 2017 was conducted in India on dated November 13, 2017, in class III, V, and VIII. It was the first time when NAS for different classes administered on a single day. About 2.2 million from 1,10,000 government and government aided schools participated in this mega event. It may be the first kind of mega activities conducted on the globe in such a large magnitude. It includes participation from different sections: gender, location, management of schools, and social groups. Figure 2 shows the participation statistics.

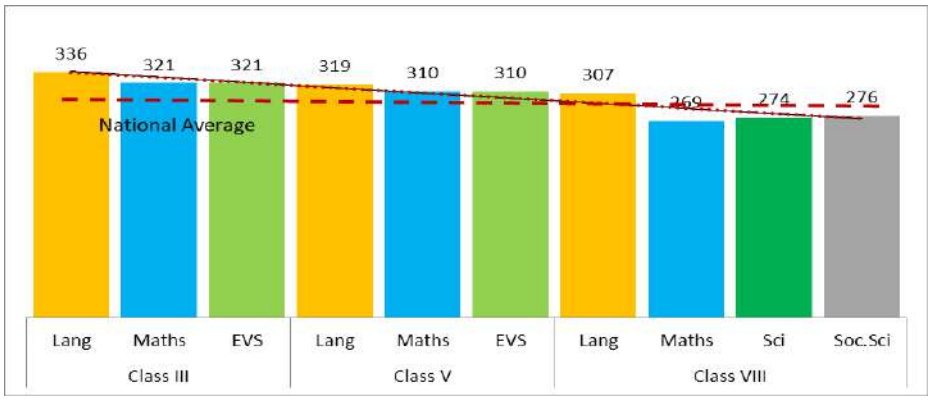
For gender, the participation of boys and girls was almost equal. Whereas, participation from rural-urban and government – government aided schools were very unequal. Participation was a cumulative representation in major social groups SC, ST, OBC and General as 22 %, 18 %, 42 % and 18 % respectively.



Source: NCERT (2020)

Figure 2: Participation by gender, location, school management and social groups in classes III, V, and VIII (figures in percentage)

Figure 3 shows the performance of students in classes III, V and VIII in the different subjects. The performance in different subjects was given in scale score values. It shows that the average achievement of students in class III was 326, the average achievement of students in class V was 313 and the average achievement of students in class VIII was 282. The overall national average of 300. Class III preference was much better than national performance and class V performance was close to the national average. However, the performance of class VIII was much below the national average. From the figure, it is concluded that with higher classes the performance of the students was decreasing.



Source: NCERT (2020)

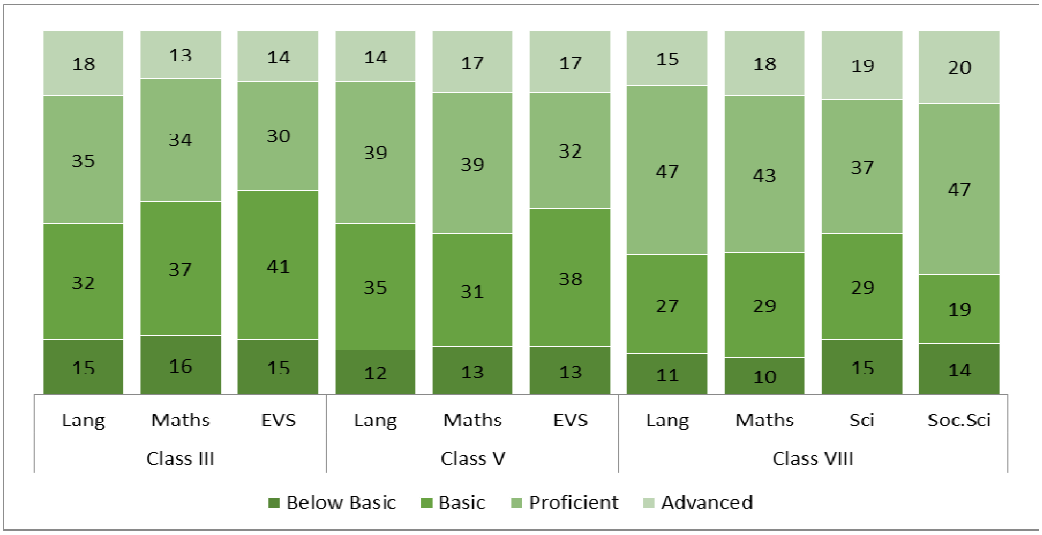
Figure 3: Students' performance at national level with class wise and subject wise and national average

Table 4 shown below presents the cut scores for each class III, V, and VIII for NAS. In Figure 4, students' performance at respective classes were given in percentage.

Table 4: Final cut scores for National Achievement Survey (NAS) tests

Test	Basic	Proficient	Advanced
Class III Language	268	315	370
Class III Mathematics	285	339	395
Class III Environmental Studies	263	315	375
Class V Language	260	306	370
Class V Mathematics	264	320	383
Class V Environmental Studies	261	315	375
Class VIII Language	255	320	370
Class VIII Mathematics	225	275	340
Class VIII Science	228	275	333
Class VIII Social Science	236	298	338

Source: NCERT (2020)



Source: NCERT (2020)

Figure 4: Percentage of students in each performance level (National Results)

Table 5: Performance of students in class III by gender, area and school management

Class III		Lang	Sig.	Cohen's D	Maths	Sig.	Cohen's D	EVS	Sig.	Cohen's D
Gender	Boys	335	**	-0.06	321	**	-0.01	320	**	-0.04
	Girls	338			321			322		
Area	Rural	336	**	-0.05	321	**	-0.02	320	**	-0.07
	Urban	336			322			324		
Management	Govt.	335	**	-0.12	320	**	-0.12	319	**	-0.22
	Govt. Aided	342			326			330		

Source: NCERT (2020)

*Note: Lang – Language, Maths – Mathematics and EVS – Environmental Science . * Statistically significant at p< 0.05; ** Statistically significant at p<0.01. The sizes of statistically significant differences are expressed by Cohen's D (Cohen, 1988). The size of the difference that is lesser than D=0.20 is considered small and practically irrelevant.*

Table 6: Performance of students in class V by gender, area and school management

Class V		Lang	Sig.	Cohen's D	Maths	Sig.	Cohen's D	EVS	Sig.	Cohen's D
Gender	Boys	317	**	-0.06	310	**	-0.04	309	**	-0.03
	Girls	320			311			310		
Area	Rural	318	**	-0.06	312	**	0.10	311	**	0.08
	Urban	321			306			306		
Management	Govt.	317	**	-0.12	311	**	0.07	311	**	-0.06
	Govt.Aided	324			308			307		

Source: NCERT (2020)

** Statistically significant at p< 0.05; ** Statistically significant at p<0.01. The sizes of statistically significant differences are expressed by Cohen's D (Cohen, 1988). The size of the difference that is lesser than D = 0.20 is considered small and practically irrelevant.*

Table 7: Performance of students in class VIII by gender, area and school management

Class VIII		Lang	Sig.	Cohen D	Maths	Sig.	Cohen D	Sci	Sig.	Cohen D	S.Sci	Sig.	Cohen D
Gender	Boys	306	**	-0.04	269	**	-0.01	275	**	0.02	278	**	0.01
	Girls	308			269			274			279		
Area	Rural	306	**	-0.09	271	**	0.18	276	**	0.16	280	**	0.13
	Urban	311			262			267			273		
Management	Govt.	305	**	-0.11	271	**	0.11	277	**	0.14	282	**	0.19
	Govt. Aided	311			265			269			271		

Source:NCERT (2020)

** Statistically significant at p< 0.05; ** Statistically significant at p<0.01. The sizes of statistically significant differences are expressed by Cohen's D (Cohen, 1988). The size of the difference that is lesser than D = 0.20 is considered small and practically irrelevant.*

6. Conclusion

From the overall analysis of NAS result, the following result was concluded as –

- Performance of Girls in the National Achievement Survey 2017 (NAS 2017) was slightly higher than boys' performance in almost all the classes.
- In class III, the achievement between urban and rural students was not distinguishable. In class V, urban students were performed statistically higher in language and rural students were performed higher in mathematics and environmental science. In class VIII, rural students were performed statistically higher in mathematics and sciences and urban students were higher in language.
- In class III, the performance of Government aided schools was statistically higher. In class V, Government schools were performed statistically higher in Mathematics and EVS whereas, in Government aided schools performance was higher in language. In class VIII, Government aided schools were performed statistically higher in Language, and Government in Mathematics, Science, and Social Science. However, in every class, the difference between the government and aided schools was very small.
- In class III, the performance of general category students was higher followed by OBC. In class V and VIII, general and OBC groups were performing slightly higher than other two social groups (SC and ST) in all subjects except language. In language, the OBC category was performed better than the general group in both classes. However, the differences were relatively small.
- A final remark on, if we considered the average of scale score achieved in all subjects in each class then there are no significant differences between gender (boys and girls) and areas (rural and urban schools). in all classes except management *i.e.*, between government and government aided schools however different was very small in class VIII. It means that boys and girls, and rural schools and urban schools were equally performed in the NAS 2017. But the performance of government schools and government aided schools was not the same.

The main objective of the study is to know the learning level of the student at different competency levels and to identify the gap in their learning. Ranking the states/Union Territories based on the performance of their students was not the objective of NAS 2017. Only to know the Top and Low performing states/union territories, ten names are: top ten states/union territories are Rajasthan, Karnataka, Chandigarh, Andhra Pradesh, Jharkhand, Dadra and Nagar Haveli, Assam, Gujarat, Kerala and Uttarakhand and low ten states/union territories are Arunachal Pradesh, Delhi, Puducherry, Meghalaya, Lakshadweep, Daman and Diu, Uttar Pradesh, Sikkim, Punjab, and Nagaland.

Acknowledgement

The author is indeed thankful to the Head and faculty members of the Educational Survey Division, NCERT, for providing the report and valuable guidance and support in the preparation of the paper. Thanks are also due to them for making available Figures and Tables from the technical report of NAS 2017 for the preparation of this paper.

References

Dave, P.N. (1988). *Achievement Under Project PECR*. New Delhi: NCERT (Mimeographed).

NCERT (2011). *Introduction of Item Response Theory in National Achievement Surveys Under Sarva Shiksha Abhiyan*. Concept Note. New Delhi, NCERT.

NCERT (2014 a). *National Achievement Survey*: *Class VIII* (*Cycle* 3). New Delhi, NCERT.

NCERT (2014 b). *National Achievement Survey* (*Cycle III*) *CLASS III*: *Achièvement Highlights*. New Delhi, NCERT.

NCERT (2015 a). *What Students Know and can do*: *A summary of National Achievement Survey Class X*. New Delhi, NCERT and MHRD, Government of India.

NCERT (2015 b). *What students of Class V know and can do*: *A Summary of India's National Achievement Survey, Class V* (*Cycle* 4). New Delhi, NCERT.

NCERT (2017). *Operational Guidelines-cum-Training Manual*: *National Achievement Survey - NAS* (2017) *Classes III, V and VIII*. New Delhi, NCERT and MHRD.

NCERT (2018). *Post NAS Intervention*: *Communication and Understanding of Districts Report Cards for NAS 2017*. New Delhi, NCERT.

NCERT (2020). *National Report to Inform Policy, Practices and Teaching Learning*: *National Achievement Survey* 2017 *Grades III, V and VIII. New Delhi*, NCERT and UNICEF (Unpublished).

Pajankar, V. D. (2019). Assessment of students' learning in different subjects at different levels of grades in Maharashtra (India). *International Journal of Research*, **9**(**1**), 73–89.

Thissen, D. and Wainer, H. (2001). *Test Scoring. Mahwah*, New Jersey, Lawrence Erlbaum Associates.

UNESCO (2018). *The Impact of Large-Scale Learning Assessments*. UNESCO and UIS, March 2018.

Zimowski, M., Muraki, E., Mislevy, R. and Bock, D. (1996). BILOG-MG (Computer program). *Available from Scientific Software International*, Inc. 7383 N. Lincoln Avenue, Suite 100 Lincolnwood, IL 60712-1747 USA.

Special Proceedings of the 22nd Annual Conference of SSCA held at S.P. Pune University, Pune, during 02-04 January 2020, 145-152

Selection of Designs for Model Misspecification in Generalized Linear Models: A Review

Siuli Mukhopadhyay[1] **and Ishapathik Das**[2]
[1]*Department of Mathematics, Indian Institute of Technology Bombay, India*
[2]*Department of Mathematics and Statistics, Indian Institute of Technology Tirupati, India*

Received: 10 July 2020; Revised: 20 August 2020; Accepted: 21 August 2020

Abstract

This article provides a brief review on design selection under model misspecification in linear and generalized linear models. Design selection for fitting a hypothesized model is one of the main focuses of response surface methodology. However, if the assumptions regarding the relationship between the response and the covariates are incorrect, then the design based on the assumed model may not provide accurate results. In generalized linear models (GLMs), a certain form of the linear predictor and the link function of the model is usually assumed and the selected designs are based on these assumptions. Model misspecification in GLMs can arise when the form of the linear predictor and/or the link function assumed is not correct. Many researchers have proposed several methods for selecting appropriate designs accounting for model bias and the prediction variance for both linear and generalized linear models. The literature review presented here discusses several existing methods in the literature based on the mean squared error criterion for comparing/selecting designs robust to the possible misspecification in the model. Several papers based on robust designs for GLMs are highlighted here. The method of comparing designs by quantile dispersion graphs (QDGs) approach addressing the linear predictor misspecification problem using an unknown function and the link function misspecification problem using a family of link functions is discussed in detail. A numerical example based on real data is provided to illustrate the QDGs methodology.

Key words: Family of link functions; Mean squared error of prediction; Quantile dispersion graphs; Robust designs.

1. Introduction

One of the main purposes of response surface methodology (RSM) is to choose an appropriate design for fitting a hypothesized model. Usually a low-degree polynomial or a simple linear model is used to explain the complex and possibly non linear relationship between the response variable and the inputs/covariates. Since the simple fitted model may not adequately approximate the unknown functional relationship that depicts the true mean

Corresponding Author: Siuli Mukhopadhyay
Email: siuli@math.iitb.ac.in

response, there is always a chance of estimates being biased. Thus, giving rise to the model misspecification problem. Due to this reason the chosen design should protect against the possibility of a sizeable model bias. Box and Draper (Box & Draper (1959) and Box & Draper (1963)) introduced the so-called integrated mean squared error (IMSE) criterion which accounts for both prediction variance and model bias and advised experimenters to choose designs on the basis of the IMSE. Instead of looking at an overall measure like the average MSE, Giovannitti-Jensen & Myers (1989) and Vining & Myers (1991) used a graphical approach to evaluate how a design performs over every portion of the region of interest in terms of IMSE. More recently, Mukhopadhyay & Khuri (2008) presented the technique of qunatile plots for evaluating and comparing response surface designs on the basis of the mean squared error of prediction (MSEP). Four MSEP-related criteria functions free of any unknown parameters that pertain to the unfitted true model and error variance were proposed. They obtained plots of the quantiles of these criterion functions on concentric spheres within a region of interest. These quantile plots gave complete information concerning the distribution of each criterion function over the selected spheres.

Recently, there has been an increase in interest among researchers to study designs robust to model misspecification in the context of generalized linear models (GLMs). Model misspecification in GLMs is a little more complex than in linear models, since in GLMs along with simple form of the linear predictor, the experimenter also assumes a form for the link function. If the assumptions regarding the functional form of the linear predictor or/and the link function are incorrect, then the inference drawn from the fitted model may not provide accurate results, giving rise to model misspecification problem in GLMs.

Selecting robust designs for GLMs have been studied by Abdelbasit & Butler (2006), Woods et al. (2006) and Dror & Steinberg (2006). In the context of logistic regression models, Adewale & Wiens (2009) used the average mean-squared error criterion to generate designs less sensitive to possible misspecifications in the linear predictors. Their work was extended by Adewale & Xu (2010) where misspecification in both linear predictors and link functions were considered. More recently, Mukhopadhyay & Khuri (2012) used quantile dispersion graphs based on MSEP to compare designs for GLMs in the presence of model misspecification in linear predictors. Their approach accounted for the bias of the fitted model's parameter estimates in addition to their variances.

2. Model Misspecification in GLMs

GLMs are usually specified by three components:

- Distributional component: It is assumed that the data of size n $y_1, \ldots, y_n$, are independent and have the following density function,

$$s(y_j|\theta_j, \phi) = \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j, \phi)\right], \; j = 1, \ldots, n, \tag{1}$$

where $b(\cdot)$, $c(\cdot)$ are known functions and ϕ is the unknown dispersion parameter. The

mean and variance of y_j are, $E(y_j) = \mu_j = \frac{db(\theta_j)}{d\theta_j}$ and Var $(y_j) = \sigma_j^2 = a(\phi)\frac{d^2b(\theta_j)}{d\theta_j^2}$, respectively.

- Linear Predictor: The linear predictor, denoted by $\eta(\mathbf{x})$, is a function of the p control variables $\mathbf{x} = (x_1, \ldots, x_p)^T$.

- Link function: The linear predictor $\eta(\mathbf{x})$ is related to the mean response, $\mu(\mathbf{x})$, through a link function g, where the inverse of g, denoted by h, is assumed to exist. The true relationship between η and $\mathbf{x}$ being usually unknown and of highly nonlinear nature.

As mentioned above, the true relationship between $\eta(\mathbf{x})$ and the vector $\mathbf{x}$ of control variables is usually unknown. The experimenter approximates the unknown relationship by a low-order polynomial model of the form,

$$\eta(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta}, \tag{2}$$

where, $\mathbf{z}^T(\mathbf{x})$ is a known vector function of $\mathbf{x}$ and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. Under the assumed model, the estimated mean response is,

$$\hat{\boldsymbol{\mu}}(\mathbf{x}) = h[\hat{\eta}(\mathbf{x})] = h[\mathbf{z}^T(\mathbf{x})\hat{\boldsymbol{\beta}}], \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimate of $\boldsymbol{\beta}$. However, suppose the true functional form of the linear predictor is different from the fitted form and is actually,

$$\eta_T(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x}), \tag{4}$$

where $f(\mathbf{x})$ is not known and the true mean response is,

$$\mu_T(\mathbf{x}) = h[\eta_T(\mathbf{x})] = h[\mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x})]. \tag{5}$$

The MSEP for the estimated mean response when the linear predictor is misspecified from Mukhopadhyay & Khuri (2012) is given by

$$\begin{aligned} \text{MSEP}[\hat{\mu}(\mathbf{x})] &\doteq \left[\frac{dh[\eta(\mathbf{x})]}{d\eta(\mathbf{x})} + f(\mathbf{x})\frac{d^2h[\eta(\mathbf{x})]}{d\eta^2(\mathbf{x})}\right]^2 \text{Var } [\hat{\eta}(\mathbf{x})] \\ &+ \left\{\text{Bias } [\hat{\eta}(\mathbf{x})]\left[\frac{dh[\eta(\mathbf{x})]}{d\eta(\mathbf{x})} + f(\mathbf{x})\frac{d^2h[\eta(\mathbf{x})]}{d\eta^2(\mathbf{x})}\right]\right\}^2, \end{aligned} \tag{6}$$

where,

$$\begin{aligned} \text{Bias } [\hat{\eta}(\mathbf{x})] &= E[\hat{\eta}(\mathbf{x})] - \eta_T(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})E(\hat{\boldsymbol{\beta}}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} - f(\mathbf{x}) \\ &= \mathbf{z}^T(\mathbf{x}) \text{ Bias } (\hat{\boldsymbol{\beta}}) - f(\mathbf{x}), \end{aligned}$$

and

$$\text{Var } [\hat{\eta}(\mathbf{x})] = \mathbf{z}^T(\mathbf{x}) \text{ Var } (\hat{\boldsymbol{\beta}})\mathbf{z}(\mathbf{x}).$$

The bias and variance of $\hat{\boldsymbol{\beta}}$ under a misspecified linear predictor are

$$\text{Bias } (\hat{\boldsymbol{\beta}}) \doteq \mathbf{H}_n^{-1}\mathbf{b}, \tag{7}$$

and

$$\text{Var } (\hat{\boldsymbol{\beta}}) \doteq \frac{1}{N}\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}, \tag{8}$$

where $\mathbf{H}_n = \mathbf{Z}^T\mathbf{P}\mathbf{W}\mathbf{Z}$, $\tilde{\mathbf{H}}_n = \mathbf{Z}^T\mathbf{P}\mathbf{W}_T\mathbf{Z}$, and $\mathbf{b} = \frac{\mathbf{Z}^T\mathbf{P}(\boldsymbol{\mu}_T - \boldsymbol{\mu})}{a(\phi)}$; $\mathbf{Z}$ is a matrix with rows $\mathbf{z}^T(\mathbf{x}_j)$, $j = 1, \ldots, n$ and $\mathbf{P}$ is an $n \times n$ diagonal matrix with elements $\frac{n_j}{N}$ with N is the total number of observations, i.e., $N = \sum_{j=1}^{n} n_j$. Both, $\mathbf{W}$ and $\mathbf{W}_T$ are $n \times n$ diagonal matrices with elements, $w_j\,(= \frac{d\mu_j/d\eta_j}{a(\phi)})$, $w_{T,j}\,(= \frac{\text{Var } (y_j)}{a^2(\phi)})$, respectively, where Var (y_j) is the true variance of y_j.

A scaled version of the MSEP (SMSEP) was used for design comparison. Their main goal was to select designs with lower values of SMSEP. For comparing two designs say D_1 and D_2, if the SMSEP of D_1 was lower than D_2 then design D_1 was said to have better prediction capability than D_2. Thus, implying that the predictive performance of design D_1 is more robust to misspecification in the linear predictor than D_2. However, two major difficulties in using the SMSEP as a design criterion was its dependency on the unknown model parameters and $f(\mathbf{x})$. Mukhopadhyay & Khuri (2012) addressed the linear predictor misspecification problem by an unknown function which was estimated using parametric empirical kriging at any point in the design region. The dependence of SMSEP on the model parameters was answered by the quantile dispersion graphs (QDGs) approach.

Das et al. (2015) considered robust GLM designs for misspecification in both linear predictors and link functions. To address the possibility of incorrect forms of link functions, they used the works of Prentice (1976); Pregibon (1980); Aranda-Ordaz (1981); Guerrero & Johnson (1982); Stukel (1988); Czado (1989, 1997) on generalized family of link functions for GLMs.

A family of parametric link functions were defined, relating $\eta(\mathbf{x})$ and $\mu(\mathbf{x})$ by $\mu = E(y|\mathbf{x}) = h(\boldsymbol{\alpha}, \eta)$, where $h(\boldsymbol{\alpha}, \cdot)$ is inverse of the parametric link function parameterized by $\boldsymbol{\alpha}$ the link parameter vector (Czado, 1997). Thus, for misspecification in both the linear predictor as well as the link function, the assumed model is

$$\mu(\mathbf{x}) = h[\boldsymbol{\alpha}_0, \eta(\mathbf{x})],$$

where $h(\boldsymbol{\alpha}_0, \cdot)$ is the assumed link function belonging to the family $\boldsymbol{\Lambda} = \{h(\boldsymbol{\alpha}, \cdot) : \boldsymbol{\alpha} \in \boldsymbol{\Omega}\}$, and the true model

$$\mu_T(\mathbf{x}) = h[\boldsymbol{\alpha}_T, \eta_T(\mathbf{x})],$$

is

$$\eta_T(\mathbf{x}) = \mathbf{Z}(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x})$$

is the true linear predictor and $\boldsymbol{\alpha}_T$ the true link parameter. The MSEP of $\hat{\boldsymbol{\mu}}(\mathbf{x})$ from Das

et al. (2015) is given by

$$MSEP[\hat{\boldsymbol{\mu}}(\mathbf{x})] = Var[\hat{\boldsymbol{\mu}}(\mathbf{x})] + Bias[\hat{\boldsymbol{\mu}}(\mathbf{x})][Bias\{\hat{\boldsymbol{\mu}}(\mathbf{x})\}]^T,$$

where

$$Var[\hat{\boldsymbol{\mu}}(\mathbf{x})] = \left[\frac{\partial h}{\partial \eta}\right]_{(\boldsymbol{\alpha}_0, \eta_T(\mathbf{x}))} \mathbf{Z}(\mathbf{x}) Var(\hat{\boldsymbol{\beta}}) \mathbf{Z}^T(\mathbf{x}) \left[\frac{\partial h}{\partial \eta}\right]^T_{(\boldsymbol{\alpha}_0, \eta_T(\mathbf{x}))},$$

and

$$\begin{aligned} Bias[\hat{\boldsymbol{\mu}}(\mathbf{x})] &= \left[\frac{\partial h}{\partial \boldsymbol{\alpha}}\right]_{(\boldsymbol{\alpha}_T, \eta_T(\mathbf{x})} (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_T) \\ &+ \left[\frac{\partial h}{\partial \eta}\right]_{(\boldsymbol{\alpha}_0, \eta_T(\mathbf{x}))} [\mathbf{Z}(\mathbf{x}) Bias(\hat{\boldsymbol{\beta}}) - f(\mathbf{x})], \end{aligned}$$

The asymptotic bias and variance of $\hat{\boldsymbol{\beta}}$ are given by

$$\begin{aligned} Bias(\hat{\boldsymbol{\beta}}) &= \mathbf{H}_n^{-1} b, \text{ and} \\ Var(\hat{\boldsymbol{\beta}}) &= \frac{1}{N} \mathbf{H}_n^{-1} \tilde{\mathbf{H}}_n \mathbf{H}_n^{-1}, \end{aligned}$$

where

$$b = \sum_{i=1}^{n} \frac{1}{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} [Var(y_i)]^{-1} (\mu_{T,i} - \mu_i),$$

$$\tilde{\mathbf{H}}_n = \frac{1}{N} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} [Var(y_i)]^{-1} [Var(y_{T,i})] [Var(y_i)]^{-1} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T},$$

and

$$\mathbf{H}_n = \frac{1}{N} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} [Var(y_i)]^{-1} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} - \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{q} \frac{\partial^2 \theta_{ij}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (y_{ij} - \mu_{ij}) n_i.$$

See Das et al. (2015) for details.

2.1. Example

We consider a real data set (Calandra Granaria data (Adewale & Xu, 2010)) containing information about studying the mortality of grain beetle after exposure to ethylene oxide (C_2H_4O). This same example was considered in Das et al. (2015). The response variable y is the proportion of killed grain beetle after one-hour exposure of 10 different levels of concentrations of C_2H_4O, which is considered as the explanatory variable (x) of the model. Here, we compare three designs: (i) the design D_7 (original design), (ii) the design D_8 ("Naive" design) and (iii) the regular optimal design D_9 (Adewale & Xu, 2010) under a

misspecified linear predictor. The data set and design settings can be found in Table 7 of Das et al. (2015).

We start fitting the data with the linear predictor

$$\eta(x) = \beta_0 + \beta_1 x, \tag{9}$$

and use the logistic link function. After estimating the unknown parameters of the model using maximum likelihood estimation method, we see $\hat{\beta}_0 = -3.4429$ and $\hat{\beta}_1 = 14.4404$, and the deviance of the fitted model is 36.2498 with 8 degrees of freedom. The observed information provides that the P-value is less than 0.0001, showing a lack of fit due to the possible misspecification of the linear predictor of the model. So, we add an unknown function to the previous linear predictor to have the modified linear predictor as

$$\eta(x) = \beta_0 + \beta_1 x + f(x). \tag{10}$$

The values of the unknown function f are first estimated at design points using the method given in Section 4.1 of Das et al. (2015) and then estimated at any points other than the design points using parametric empirical kriging. We see that the deviance of the fitted model after adding an unknown function f is decreased to 4.9778, showing an improvement of the fit by addressing the linear predictor misspecification of the model.

For comparing the performance of three designs concerning the proximity to the center/boundary of the design region, the experimental region $\mathcal{R} = \{x : 0.0330 \leq x \leq 0.3940\}$ is divided into several concentric regions $\mathcal{R}_\nu$ parametrized by some parameter $\nu \in [0.5, 1]$. The designs are compared based on the minimum and maximum quantiles of the estimated MSEP values over randomly selected 1000 samples from $\mathcal{R}_\nu$ and 1000 samples from a 95% confidence region $\mathcal{C}$ of the regression parameter vector $\boldsymbol{\beta}$. The minimum and maximum quantiles of the three designs for $\nu = 0.6, 0.7, 0.8, 0.9$, are shown in Figure 1, which is known as the quantile dispersion graphs (QDGs). From the QDGs, we see that the minimum quantiles are close to each other for all designs. The maximum quantiles of D_9 are larger than D_7 and D_8 if $\mathfrak{p} > 0.5$ for all values of ν. So, the prediction capabilities of D_7 and D_8 are better than the design D_9, while designs D_7 and D_8 have comparable prediction capabilities throughout the region as the maximum quantiles are very close to each other for all values of ν. It can also be noted by observing the differences of maximum and minimum quantiles of the designs that D_7 and D_8 are more robust than D_9 with respect to the changes of the values of $\boldsymbol{\beta}$. More details about this example can be found in Section 5.3 of Das et al. (2015).

3. Some New Directions

Though the topic of model misspecification and its effect on design selection has been discussed by several researchers for single response linear and generalized linear models, very little work has been done in the multivariate response case. However, in many experimental situations, instead of one response, several such responses are recorded for the same subject. This is very common in drug testing experiments where along with efficacy of the drug, toxic effect of the drug are also measured, and the two responses are then modeled using a bivariate distribution. Very recently, Das & Mukhopadhyay (2019) discussed the effect

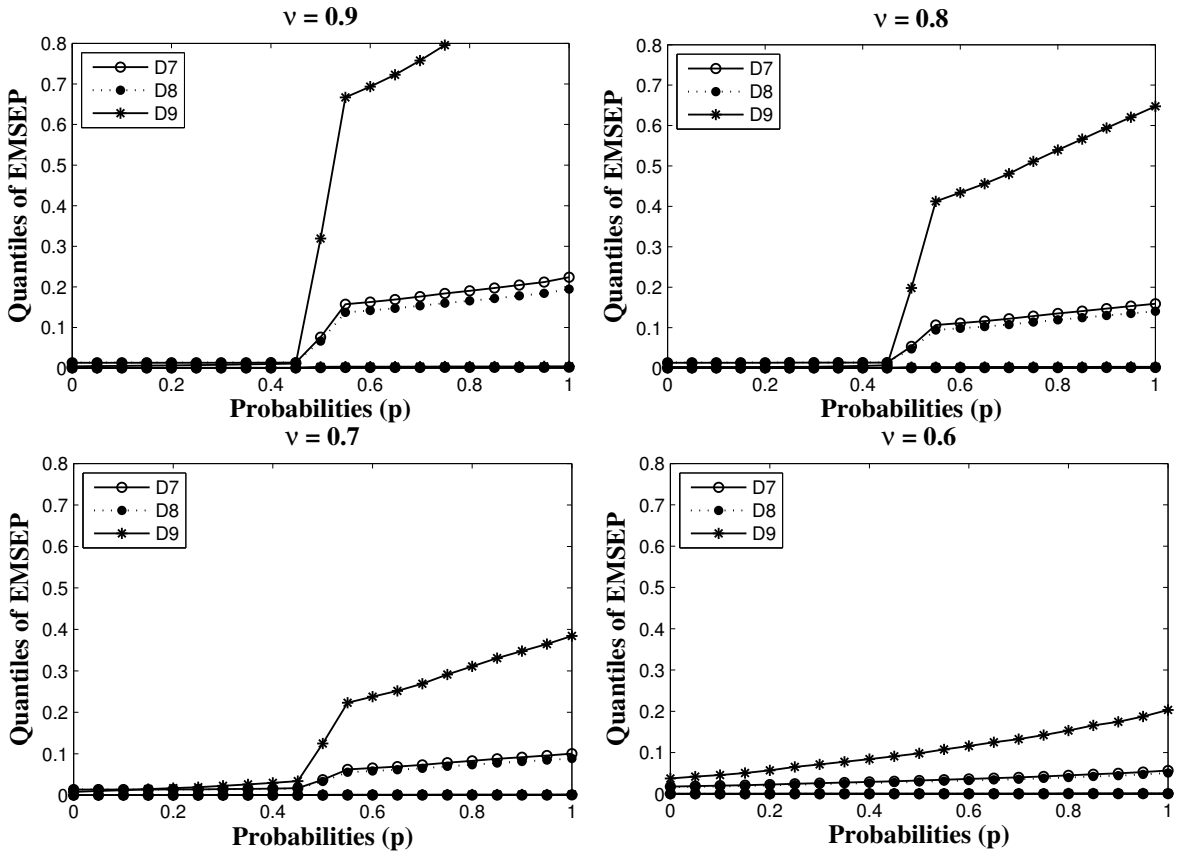


Figure 1: Quantile dispersion graphs for designs D_7, D_8, and D_9. This figure is reproduced from Figure 3 of Das et al. (2015).

of such misspecification on design selection for multinomial GLMs and proposed the use of quantile dispersion graphs to select robust designs. While multivariate kriging was used to tackle the unknown functional relationships between the linear predictors and covariates, a parametric link function family for the multinomial distribution (Das & Mukhopadhyay (2014)) was used for possible link function correction. Compromised exact D- optimal designs which are robust to possible misspecifications in the model and link functions were discussed recently by Singh & Mukhopadhyay (2019) for gene sequence studies modeled by count time series models.

References

Abdelbasit, K. M. and Butler, N. A. (2006). Minimum bias designs for generalized linear models. *Sankhya: The Indian Journal of Statistics (2003-2007)*, **68(4)**, 587–599.

Adewale, A. J. and Wiens, D. P. (2009). Robust designs for misspecified logistic models. *Journal of Statistical Planning and Inference*, **139**, 3–15.

Adewale, A. J. and Xu, X. (2010). Robust designs for generalized linear models with possi-

ble overdispersion and misspecified link functions. *Computational Statistics and Data Analysis*, **54**, 875–890.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, **61**, 357–363.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association*, **54**, 622–654.

Box, G. E. P. and Draper, N. R. (1963). The choice of a second order rotatable design. *Biometrika*, **50**, 335–352.

Czado, C. (1989). *Link misspecification and data selected transformations in binary regression models*. Technical report, Ph.D. Thesis. School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

Czado, C. (1997). On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and inference*, **61**, 125–139.

Das, I., Aggarwal, M. and Mukhopadhyay, S. (2015). Robust designs in generalized linear models: A quantile dispersion graphs approach. *Communications in Statistics-Simulation and Computation*, **44(9)**, 2348–2370.

Das, I. and Mukhopadhyay, S. (2014). On generalized multinomial models and joint percentile estimation. *Journal of Statistical Planning and Inference*, **145**, 190–203.

Das, I. and Mukhopadhyay, S. (2019). Robust designs for multinomial models. *Communications in Statistics-Simulation and Computation*, **48(10)**, 2998–3021.

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, **48**, 520–529.

Giovannitti-Jensen, A. and Myers, R. H. (1989). Graphical assessment of the prediction capability of response surface designs. *Technometrics*, **31**, 159–171.

Guerrero, V. and Johnson, R. (1982). Use of the box cox transformation with binary response models. *Biometrika*, **69**, 309–314.

Mukhopadhyay, S. and Khuri, A. I. (2008). A new graphical approach for comparing response surface designs on the basis of the mean squared error of prediction criterion. *Statistics and Applications*, **6**, 293–324.

Mukhopadhyay, S. and Khuri, A. I. (2012). Comparison of designs for generalized linear models under model misspecification. *Statistical Methodology*, **9**, 285–304.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society*, **29**, 15–24.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, **32**, 761–768.

Singh, R. and Mukhopadhyay, S. (2019). Exact bayesian designs for count time series. *Computational Statistics and Data Analysis*, **134**, 157 – 170.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, **83**, 426–431.

Vining, G. G. and Myers, R. H. (1991). A graphical approach for evaluating response surface designs in terms of the mean squared error of prediction. *Technometrics*, **33**, 315–326.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284–292.

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Ganeshkhind, Pune, during January 02-04, 2000; pp 153-162

On Weighted Distributions and Applications

Saumyadipta Pyne[1,2,3]

[1]*Public Health Dynamics Laboratory, University of Pittsburgh, Pittsburgh, USA*

[2]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA*

[3]*Health Analytics Network, PA, USA*

Received: 15 August 2020; Revised: 22 August 2020; Accepted: 25 August 2020

Abstract

The birth centenary of C.R. Rao in 2020 presents an occasion to not only celebrate the remarkable life and career of a living legend of statistics, but also remember the worldwide immense development of the field over the past century. Here, I discuss about C.R. Rao's pioneering work on a general theory of weighted distributions presented in 1965, which was followed by significant development in that area. I end with discussion of some recent advances in methodology with applications to environmental data fusion.

Keywords: C.R. Rao; Weighted distribution; Density ratio model; Data fusion.

0. About This Paper

This paper is based on the first lecture delivered during the C.R. Rao Birth Centenary Session of the 22nd Annual Conference of Society of Statistics, Computer and Applications and ISGES 2020, held on January 2, 2020, at the Department of Statistics, Savitribai Phule Pune University, Pune. Three more talks were delivered during this session. The session was Chaired by Professor Vinod K. Gupta.

1. Background

Calyampudi Radhakrishna Rao, popularly known as "C.R. Rao", was born on the 10th of September, 1920, in the Madras Presidency of British India. He is widely regarded as a "living legend" in the field of statistics, and known for Cramer-Rao Bound, Rao-Blackwell Theorem, Rao Score Test, Fisher-Rao distance, Generalized Inverse, Quadratic Entropy, and Orthogonal Arrays, among his numerous path breaking contributions. Along with P.C. Mahalanobis, he played a major role in developing the Indian Statistical Institute (ISI) into the major center of statistical research and education by the 1940s. C.R. Rao's long list of awards includes the Padma Vibhushan (2001) and Padma Bhushan (1968), India Science Award (2009), the United States (US) National Medal of Science (2002), Wilks Memorial Award (1989), the Guy Medal in both Silver (1965) and Gold (2011).

Compared to some of Rao's other breakthroughs in statistics as stated above, the topic of my lecture, weighted distributions, is relatively less well known but has, nonetheless, led to great advances of research in the subsequent decades. It provides a curious counterpoint to the popular refrain that every student of statistics gets used to in her daily practise: "randomly

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

drawn samples are assumed to be independently and identically distributed." The topic has been found relevant to many theoretical and applied areas of statistics, and could, in fact, help modern data science to deal with analysis of samples that may not be collected through well-designed experiments. Interestingly, in the arc of Rao's remarkable career as it spanned over the better part of the last century, this topic happens to make a uniquely historic contribution.

In 1911-1914, (later Sir) Henry Wellcome, the British pharmaceutical entrepreneur had led, over four seasons, an archeological expedition at the Jebel Moya site in the southern Gezira plain of Sudan (Addison, 1949). It is the site of the largest pastoralist cemetery in Northeast Africa, currently dated to 5000–500 BC (Brass, 2016). The excavated skeletal remains were shipped to London where these were warehoused poorlyand even suffered from flooding. After the death of Wellcome (in 1936), and by the end of two World Wars, the remains and the excavation records were transferred to the Duckworth Laboratory in the University Museum of Archaeology and Ethnology at Cambridge University. The Wellcome Trust wanted the study to be completed under the curatorship of J.C. Trevor.

In March 1946, Trevor sent a telegram to Mahalanobis asking him to send someone to help with the anthropomorphic analysis of the collection. Incidentally, just prior to that, Mahalanobis had assigned a project on analysis of anthropometric data to Rao, who used the D^2 distance for grouping of Indian populations (Mahalanobis, Majumdar and Rao, 1949). This experience led to his selection to go, along with Ramkrishna Mukherjee, an anthropologist, to England in August 1946 to work as visiting scholars at the Cambridge University Museum of Archaeology and Anthropology. The aim of their study was "to undertake laboratory examination of identifiable and usable adult specimens, to analyse the measurements and observations of the field physical anthropologists, and to determine the relationship between the Jebel Moya inhabitants and other African peoples" (Brass, 2016).

Upon completion of their analysis, the long overdue report of this Wellcome Trust project was finally published in 1955 by Mukherjee, Rao and Trevor as a book titled, '*The Ancient Inhabitants of Jebel Moya* (*Sudan*)'. (Mukherjee, Rao and Trevor, 1955) Unfortunately, the 40-year hiatus between the excavation and the anthropometric data analysis had proved to be catastrophic for the remains, which had "disintegrated beyond hope of repair", according to Trevor. Out of more than 3000 skeletal parts originally excavated, only 98 crania, 139 mandibles and a few post-cranial elements had survived for conducting the anthropomorphic studies by Rao and his co-workers at Duckworth Laboratory.

Rao's task was to estimate the unknown mean cranial capacity and other features of the *original* Jebel Moya population from the damaged remains, many of which had measurements missing due to damage. Towards maximum likelihood estimation, one could write the likelihood function using a multivariate (normal) distribution based on the samples with complete measurements, and the derived marginal distribution for those with incomplete set of measurements. However, such estimation assumes that each skull – with all or some of the measurements – can be considered as part of a *random* sample from the original population of skulls. The key question was, however, was such an assumption valid?

While looking at the samples that survived, one could make a curious observation, "are only small skulls preserved?" (Figure 1). If $w(c)$ is the probability that a skull of capacity c is unbroken, then in archeological recovery, *i.e.*, during the data gathering process, it is known that $w(c)$ is a decreasing function of c. The larger a skull, the greater is its chance of being damaged upon burial and recovery, as Rao noted in his study of another collection (Rao and

Shaw, 1948). This will lead to a larger representation of small skulls among the unbroken cranial remains, and, therefore, the mean of the available measurements of the corresponding random variable C will be an underestimate of the mean cranial capacity of the original population.

While analysis of Jebel Moya remains was his full-time job, Rao, who was also a student in King's College, Cambridge, was "suggested" by his Ph.D. advisor R.A. Fisher to work simultaneously at the latter's genetics laboratory. Fisher had long been interested in the concept of "ascertainment" – a mode of sampling that depends on the outcome that one wanted to analyze as a dependent variable. In a classical paper, Fisher studied how the methods of ascertainment can influence the form of the distribution of recorded observations (Fisher, 1934).One can assume a model that has been adjusted for ascertainment will estimate parameters in the *general* population from which the sample was drawn. In usual statistical practice, it is generally assumed that a random sample from a population to be studied can be observed in data. Not surprisingly, therefore, the key specification of 'what population does a sample represent?' might be taken for granted.

However, in many situations, obtaining a random sample may be practically too difficult, or too costly, or indeed, even less preferable to the non-random data that are actually available, *e.g.*, from field observations or non-experimental data or a survey lacking a suitable sampling frame, as in many 'big data' problems. Sometimes the events may be observed only in modified form, *e.g.*, in damage models. Indeed, certain events may either be unobservable due to adoption of the very method that is used for making observations, and therefore, missed in the record. Or, they might be observable but only with certain probabilities or weights that may depend on the samples' specific characteristics, such as conspicuousness, as well as other unknown parameters.

2. Weighted Distributions

On 15-20 August of 1963, G.P. Patil organized the 'First International Symposium on Classical and Contagious Distributions' at McGill University in Montreal. Patil had visited ISI a decade earlier, where he recollected being advised by Rao to study discrete distributions. In the Montreal symposium, at a 2nd day session chaired by Jerzy Neyman, Rao presented the first paper that formulated and unified weighted distributions in general terms. It was titled, '*On discrete distributions arising out of methods of ascertainment*'. Later, the paper was published by the Statistical Publishing Society, Calcutta, in the Proceedings, 'Classical and Contagious Discrete Distributions', edited by Patil (Rao,1965).

Let X be a random variable (rv) with probability density function (pdf) $g(x;\theta)$ with parameters θ. Traditional statistical analysis assumes that an identically (as well as independently) distributed random sample $X_1, \ldots, X_n$ can always be observed. Weighted distributions arise when $X = x$ enters the sample with a non-zero weight $w(x,\alpha)$ that depends on the observed value x and possibly also on some unknown parameter α. Then x is not an observation on X but on the resulting rvX^w which has the weighted pdf:

$$f(x;\theta,\alpha) = \frac{w(x,\alpha)\cdot g(x;\theta)}{E[w(X,\alpha)]}$$

The denominator $E[w(X;\alpha)]$ is a normalizing constant so that $f(x;\theta,\alpha)$ integrates to 1. The weight $w(x,\alpha)$ could be any non-negative function for which $E[w(X;\alpha)]$ exists.

When x is univariate and non-negative, then the weighted distribution $f(x;\theta) = x \cdot g(x;\theta)/E(X)$ for $w(x,\alpha) = x$ is called *size-biased*; and *length-biased* for $w(x,\alpha) = |x|$ where $|x|$ is some measure of "length" of x. Many length-biased distributions could be shown to belong to the same family as their unweighted versions (Rao, 1965).

Weighted distributions are utilized to modulate the probabilities of the events as they are observed by means of collecting data possibly under less than perfect conditions. In this context, prominent application areas of weighted distributions include truncation and censoring, damage models, size-biased sampling, quadrat sampling, nonresponse in data, "file-drawer" problem in meta-analysis, *etc*. Even mixtures of distributions can be shown to belong to this general formulation (Larose and Dey, 1996). Interestingly, sometimes biased samples available from observational studies may contain more (Fisher) information than their randomly drawn counterparts (Bayarri and DeGroot, 1992). Patil and Rao (1978) studied size biased sampling with applications to wildlife populations and human families. Different applications of weighted distributions were reviewed by Patil and Rao (1977), and in reports and articles written byRao during his Pittsburgh years (Figure 2), *e.g*., Rao (1985), Rao (1988).

As an example of such special cases of weighted distributions, we consider skewed data. Let the p-variate Generalized Skew Elliptical (GSE) distribution of rv $z \in R^p$have pdf of the form $2g(z;\xi,\Omega)\pi(z-\xi)$ where g is an elliptically contoured pdf with location parameter ξ, scale matrix Ω, and skewing function π. An example of a GSE pdf due to Branco and Dey (2001) is $\frac{2}{\sqrt{\Omega}} g(\Omega^{-\frac{1}{2}}(z-\xi)) \cdot \pi(\Omega^{-\frac{1}{2}}(z-\xi))$, which could be formulated as a weighted distribution as follows: $f(z;\theta) = g(\Omega^{-1/2}z)/|\Omega|^{1/2}$, $w(z) = \pi(\Omega^{-1/2}z)$, $E[w(Z)] = 1/2$ (assuming ξ=0). Here, the weight function w distorts the elliptical contours of f*via* generation of asymmetric outliers in the observed sample due toGSE (Genton 2005).

Such formulations demonstrate the capacity of introducing weighting mechanisms to "distort" distributions as required. This allows important applications such as modeling of dynamic patterns as they emerge in data over space and/or time, as in environmental monitoring, statistical ecology, public health, etc. For describing the changes over time in the distribution of an environmentally important variable X, a *propagation function* (PF) models how the frequency of sampling units with the value $X = x$ at one point in time must change in order to produce the distribution that occurs at a later point in time. Thus, PF is a useful tool in long-term monitoring studies since all changes in a distribution can be examined together rather than just changes in single parameters such as the mean.

Let X and Y be the values of an environmental variable on a population unit at two consecutive time-points, with the marginal densities $f_X(\cdot)$ and $g_Y(\cdot)$ (Kaur, *et al*., 1995). The PF is defined as: $w(x) = g_Y(x)/f_X(x)$. This gives a weighted distribution: $g_Y(x) = w(x) \cdot f_X(x)$ such that $E_f[w] \equiv E[w(X)] = 1$. A similar concept of *resource selection function* (RSF) in wildlife habitat modeling was defined as a logistic discriminant function in terms of a ratio of pdf-s for used (f_u) and available (f_a) resources in k habitats: $f_u(x)/f_a(x) = w(x) = \exp(\sum_{i=1}^{k} \beta_i x_i)$, where β_i is the model coefficient for the i^{th} habitat's covariate x_i(McDonald, Gonzalez and Manly,1995).A sampling strategy could be biased in its observation of the used sites, given their notably used features, over the unused sites. Later, the idea of a general density ratio model (DRM) was further developed and applied to case-control studies (Qin, 1998).

3. Recent Extensions to Weighted Systems

Recent studies have extended the concept of weighted distributions to weighted *system* of distributions. This general approach allows us to develop a powerful computational framework for modeling the combined dynamics of environmental samples as collected from multiple sources, *e.g*., spatially distributed multivariate data streams, automated sensors and surveillance networks (Zhang, Pyne and Kedem, 2020a; Zhang, Pyne and Kedem, 2020b). Often, environmental monitoring stations are not located randomly, may contain built-in redundancies, and are difficult to maintain. Therefore, by systematic fusion of a moderate number of such nearby data sources, we can increase the predictive capability with a combined model.

We assume a reference event distribution g_0, and its possible distortions or "tilt"-ed forms $g_1, \dots, g_m$ due to m different sources. This gives us a weighted system of distributions:

$$\begin{aligned} g_1(x) &= w_1(x) \cdot g_0(x) \\ &\vdots \\ g_m(x) &= w_m(x) \cdot g_0(x) \end{aligned}$$

Assume that we have data from each of $g_0, g_1, \dots, g_m$. Then, the relationship between a baseline distribution and its distortions or tilts allows us to do inference on the system's behavior based on fusion of data observed from multiple, possibly dependent, sources. In a weighted system using density ratio models, various distributions are "regressed" on a common reference distribution, and our data fusion approach to estimate the parameters, including the densities, uses the entire combined data and not just the reference sample. Tail probabilities of exceeding a pre-specified event threshold can be estimated by using the DRM with variable tilt functions (Zhang, Pyne and Kedem, 2020a).

As an application of our data fusion approach, we combined data from multiple sites on environmental exposures of Radon, a known radioactive carcinogenic gas, using a multi-sample DRM defined as follows:

$$\frac{g_k(x)}{g_0(x)} = \exp(\alpha_k + \beta_k^T \boldsymbol{h}_k(x))\, k = 1, \dots, m$$

where g_0 represents the density of residential radon levels of the county of interest and $g_1, \dots, g_m$ represent the densities of its m nearby sites. Instead of making parametric assumptions on each of these densities, DRM captures the parametric structure of their ratios with a common model. To prevent bias, large standard errors, and loss of power in inference, it is important to properly select the variable tilt functions $\boldsymbol{h}_k$'s.

Let $\boldsymbol{X}_0, \dots, \boldsymbol{X}_m$ be the samples from the area of interest and its m nearby sites with sample sizes $n_0, \dots, n_m$ respectively. The sample $\boldsymbol{X}_0$ is referred to as the reference sample and let G denote the corresponding reference cumulative distribution function (CDF). The fused sample is defined as $\boldsymbol{t} = (\boldsymbol{X}_0^T, \dots \boldsymbol{X}_m^T)^T$, with size $n = \sum_{k=0}^{m} n_k$. Inference can be based on the following empirical likelihood obtained from the fused sample $\boldsymbol{t}$:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{k=1}^{m} \prod_{j=1}^{n_k} \exp(\alpha_k + \beta_k^T \boldsymbol{h}_k(X_{kj}))$$

where $p_i = d(G(t_i))$, and the estimates $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and hence $\tilde{p}_i$'s are obtained by maximizing the above likelihood with constraints:

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(t_i)\right) = 1 \qquad k = 1, \ldots, m.$$

This gives the estimated reference CDF $\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}_i I[t_i \le t]$ and the asymptotic result

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{d} N(0, \sigma(t)),$$

as $n \to \infty$. The expression of $\sigma(t)$ and other details regarding estimation and asymptotic result can be found in our recent paper (Zhang, Pyne and Kedem, 2020a). Based on the above result, a 95% confidence interval of the tail probability $1 - G(T)$ for a given threshold $T$is given by

$$(1 - \tilde{G}(T) - z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}, 1 - \tilde{G}(T) + z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}).$$

Finally, an optimal choice of tilt function may be made based on a criterion that ensures better specification of the density ratio structure. For instance, such selection can be made using Akaike Information Criterion (AIC) given by $-2\log L(\tilde{\alpha}, \tilde{\beta}, \tilde{G}) + 2q$ where q is the number of free parameters in the model (Zhang, Pyne and Kedem, 2020b).

4. Conclusion

In statistical inference, Rao noted, "wrong specification may lead to wrong inference, which is sometimes called the third kind of error in statistical parlance. The problem of specification is not a simple one." (Rao, 1988) The aim of my lecture, therefore, was to draw the attention of students and researchers to the fascinating area of weighted distributions that is all the more relevant in the current age of data science. Patil and Rao (1977) remarked, "although the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized." As more datasets of value seemingly lacking any rigorous design are collected, and possibly shared through unconventional and occasionally biased sources, the need for working with such less than ideal yet practically useful observations will have to be addressed by statistical pedagogy.

Personally, I had the good fortune to have had Professor Rao as my colleague, mentor and collaborator while serving as the P.C. Mahalanobis Chair Professor (sponsored by the Ministry of Statistics and Program Implementation, Government of India) and Head of Bioinformatics at the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS) in Hyderabad during 2012-2015. Starting in 2009, we had interactions at the early stages of planning the creation of the institute. We worked together on multiple projects that remain close to my heart, including the organization of the 'International Year of Statistics 2013' (STAT 2013) Conference held at CR Rao AIMSCS on December 28-31, 2013, under my convenorship; inauguration of the 'Data Science Laboratory for Environmental and Health Sciences' at my initiative at CR Rao AIMSCS by Professor Amartya Sen on 19 December, 2013 (Figure 3); co-editing a 2-volume 'Handbook of Statistics' on Disease Modeling and Public Health' (Rao, Pyne and Rao, 2017); founding in 2014 of a Special Interest Group in Computer Society of India on 'Big Data Analytics' with

myself as the founding chair. Notably, in STAT 2013, we hosted an early meeting of international experts on 'big data' in India (Pyne, Rao and Rao, 2016).

I, therefore, felt immensely proud and privileged to deliver the first lecture in the Professor C.R. Rao Centenary Lecture Session on January 2, 2020, of the International Conference ISGES 2020 organized by the Society of Statistics, Computer and Applications at the Department of Statistics, Savitribai Phule Pune University. Incidentally, the University is also my alma mater, which doubled my pleasure to speak both at this occasion as well as the venue. For this, I thank the organizers of the conference, and, in particular, my dear friend and former ICAR-IASRI National Professor, Professor Vinod K. Gupta, for kindly inviting me to deliver this lecture.

The citation with the National Medal of Science, the highest award in the US in a scientific field, honored Rao "as a prophet of new age for his pioneering contributions to the foundations of statistical theory and multivariate statistical methodology and their applications, enriching the physical, biological, mathematical, economic and engineering sciences."Indeed, to truly appreciate Rao's "putting chance to work", I encourage students to read his popular writings (Rao 1997). It has been a high honour for me to have worked alongside this living legend and received his guidance and blessings. In August 2019, I visited him (and his daughter Teja Rao) at his family home in Buffalo, New York, and expressed my deepest gratitude. I offer my heartiest congratulations to Professor C.R. Rao on his birth centenary year, and wish him a longer, healthy life.

References

Addison, F. (1949). *The Wellcome Excavations in the Sudan: Vol. I, Jebel Moya, 1910-1914*. Oxford University Press, Oxford.

Bayarri, M. J. and DeGroot, M. H. (1992) A "BAD" view of weighted distributions and selection models. In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics*,**4**, Oxford: University Press, 17-33 (with discussion).

Branco, M. D. andDey, D.K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99-113.

Brass, M. J. (2016). *Reinterpreting Chronology and Society at the Mortuary Complex of Jebel Moya* (*Sudan*). Archaeopress Publishing Ltd., Oxford.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, **6**, 13–25.

Genton, M. G. (2005). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman and Hall/CRC, Boca Raton.

Kaur, A., Di Consiglio, L., Patil, G. P. et al. (1995). Propagation functions for monitoring distributional changes. *Environmental and Ecological Statistics*, **2**, 239–269.

Larose, D.T. andDey, D. K. (1996). Weighted distributions viewed in the context of model selection: A Bayesian perspective. *Test*, **5**, 227–246.

Mahalanobis P.C., Majumdar D.N. and Rao C.R. (1949). Anthropometric survey of the United Provinces: a statistical study. *Sankhya*, **9**, 89–324.

McDonald, L. L., Gonzalez, L. and Manly, B. F. J. (1995). Using selection functions to describe changes in environmental variables. *Environmental and Ecological Statistics*, **2**, 225-237.

Mukherjee, R., Rao, C. R. and Trevor, J. C. (1955). *The Ancient Inhabitants of Jebel Moya* (*Sudan*). Cambridge University Press, Cambridge.

Patil, G. P. and Rao, C. R. (1977). The weighted distributions: A survey of their applications. In P.R. Krishnaiah (Ed.), *Applications of Statistics*, North Holland, Amsterdam, 383–405.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179–189.

Pyne, S., Prakasa Rao, B. L. S., Rao, S. B. (Eds.) (2016). *Big Data Analytics*: *Methods and Applications*. Springer, New Delhi.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**(**3**), 619-630.

Rao, A.S., Pyne, S. and Rao, C.R. (Eds.) (2017). *Handbook of Statistics*: *Disease Modelling and Public Heath*, *Part A and Part B*. Elsevier, Amsterdam.

Rao, C. R. and Shaw, D. C. (1948). On a formula for the prediction of cranial capacity. *Biometrics*, **4**, 247–253.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In G.P. Patil (Ed.), *Classical and Contagious Discrete Distributions*, Calcutta: Statistical Publishing Society, 320–333. Reprinted in *Sankhya*, **A27**, 311–324.

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In Atkinson A.C., Fienberg S.E. (Eds.), *A Celebration of Statistics*: *The ISI Centenary Volume*. Springer, New York. An earlier version appeared as the Technical Report No. 84-38, July 1984, Center for Multivariate Analysis, University of Pittsburgh, Pittsburgh, PA 15260.

Rao, C. R. (1988). *Weighted and Clouded Distributions*. Technical Report No. 88-01, Center for Multivariate Analysis, University of Pittsburgh, Pittsburgh, PA 15260.

Zhang, X, Pyne, S. and Kedem, B. (2020a). Estimation of residential radon concentration in Pennsylvania counties by data fusion. *Applied Stochastic Models in Business and Industry*, 1-17.

Zhang, X., Pyne, S., Kedem, B. (2020b). Model selection in radon data fusion. *Statistics in Transition*, *New Series*, Special Issue, August, 167-174.

ANNEXURE

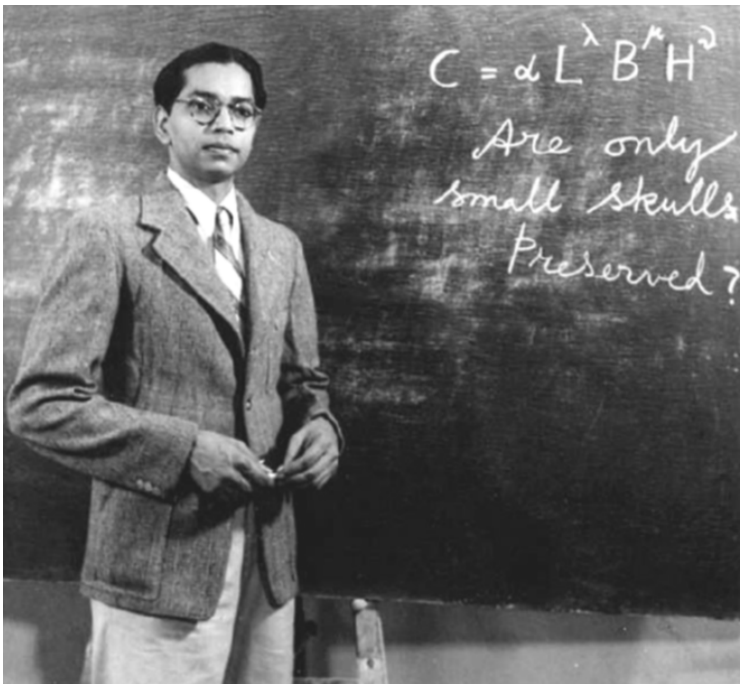


Figure 1: C.R. Rao in Cambridge during the 1940s. For discussion on the cranial capacity formula shown in the picture, see Rao and Shaw (1948).

Figure 2: C.R. Rao (standing 4th from the left) with family, friends, students in Pittsburgh, USA, during the 1980s. With P.R. Krishnaiah (standing 2nd from the left), Rao established in 1982 a unique Center for Multivariate Analysis at the University of Pittsburgh.

Figure 3:Inauguration of Data Science Laboratory, CR Rao AIMSCS, Hyderabad (December 19, 2013). From left to right: (late) Bhargavi Rao, C.R. Rao, Amartya Sen, Saumyadipta Pyne, S.B. Rao, Teja Rao (Inset: foundation plaque).

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Ganeshkhind, Pune, during January 02-04, 2000; pp 163-174

Digital Transactions Through Debit Cards – Costs and Prices for the Payment Services

Ashish Das
Department of Mathematics, Indian Institute of Technology Bombay, Mumbai 400076, India

Received: 07 August 2020; Revised: 27 August 2020; Accepted: 30 August 2020

Abstract

To promote small ticket debit card transactions up to Rs. 2000, the government during the calendar years 2018 and 2019 made merchant discount rate (MDR) zero for the merchants, while providing monetary support to banks @ 0.4%, towards MDR. In contrast, effective January 2020, the government made MDR zero for any transaction amount for use of RuPay debit cards alone and neither merchants nor the government paid the banks for such merchant transactions. However, banks were allowed to impose MDR for any transaction amount onto the merchants for use of mastercard/VISA debit cards.

For the period January-June 2020, with about Rs. 3,69,609.00 crore worth of debit card merchant transactions, the government has done away with the merchant's zero MDR regime (on ticket sizes up to Rs. 2,000.00) for about Rs. 1,88,867.00 crore worth of transactions that were done through mastercard/VISA debit cards. A simple projection implies that merchants would be overburdened in the calendar year 2020 in the range of Rs. 1,500.00 crore and Rs. 3,400.00 crore depending upon the MDR ranging between 0.4% and 0.9% for mastercard/VISA's sub Rs. 2,000.00 ticket transactions, as against nil burden in calendar years 2018 and 2019.

We show that during the one year period August 2018 through July 2019, there had been issuance of at least 455 lakh RuPay debit cards corresponding to at least 422 lakh new accounts added under PMJDY. In contrast, for the same tenure during August 2019 through July 2020, we see a subdued issuance of around 77 lakh RuPay debit cards despite at least 366 lakh new PMJDY accounts added. This showcases that with a revenue differential between RuPay and mastercard/VISA, banks and system providers, in their commercial interest have taken steps to move away from RuPay and promote a card scheme which generates more revenue for them.

The government and RBI have effectively implemented a net increase of debit card MDR expenses at least for the small and medium merchants. Now, there are two contrary aspects to such an increase. (A) Such an increase in the "cost to merchant" is good for the development and increased acceptance of debit cards, and (B) The increased net MDR, onto small and medium merchants, is bad for the development and increased usage of debit cards.

If (A) holds (and merchants pay a controlled MDR), there was no need for the induced discrimination between RuPay and mastercard/VISA. The same could have been achieved by arriving at a lower controlled MDR, uniform across all card schemes. However, if (B) holds, by putting restrictions only for RuPay and discriminatorily allowing mastercard/VISA to impose MDR onto merchants, we have not quite achieved desired results, unlike the government's strategy of zero MDR for merchants for the two calendar years 2018-19.

Corresponding Author: Ashish Das
Email: ashish@math.iitb.ac.in

Using debit card and ATM data, this paper prepare grounds for policy guidance.

Keywords: ATM cash withdrawal; Merchant Discount Rate; RuPay Debit Card.

1. Introduction

Debit cards are issued by banks for facilitating bank account holders towards interoperable ATM cash withdrawal and carrying out merchant transactions. For merchant payments, such debit cards are issued under a card schemes, which are primarily *mastercard*, *VISA* and *RuPay*. Historically, card payments for merchant transactions had a well-defined revenue generating structure, where the revenue came from Merchant Discount Rate[1] (MDR).

In order to set catalysts for the digital payment systems, Government of India on February 29, 2016 came out with cabinet approved guidelines for the 'Promotion of Payments through Cards and Digital means'. The Finance Ministry's office memorandum provides broad guidelines on the way forward for promotion of digital payments. Among several measures for wider adoption of card/digital transactions, two specific measures therein was to take steps to "rationalize MDR on card transactions" and to ensure that the card holders are not imposed a charge for using such a digital means of payment.

1.1. The history of MDR regulations

In September 2012, Reserve Bank of India (RBI) mandated to cap debit card MDR at 0.75% for transactions up to Rs. 2,000.00 and 1% for transactions above Rs. 2,000.00. This continued till 08 November 2016.

Immediately after the demonetization of the specified bank notes on 08 November 2016, the government instructed banks to temporarily waive MDR imposed on merchants.

As an interim measure, RBI effective 01 January 2017 rationalized the MDR on debit cards by capping it at (i) 0.25% for transactions valued up to Rs. 1,000.00; (ii) 0.5% for transactions valued in excess of Rs. 1,000.00 but not exceeding Rs. 2,000.00; and (iii) 1% for transactions valued in excess of Rs. 2,000.00. RBI's new caps on debit card MDR were a substantial reduction to the RBI's pre-demonetization cap of 0.75% for transactions valued up to Rs. 2,000.00.

Subsequently, effective 01 January 2018, RBI tweaked MDR rules claiming that such tweaks would encourage some small businesses to accept debit card payments. For businesses with annual turnover below Rs. 20.00 lakhs, RBI capped the debit card MDR at 0.4% of transaction value or Rs. 200.00, whichever is lower. For others, the debit card MDR was capped at 0.9% of the transaction value or Rs. 1,000.00, whichever is lower. For QR-code based debit card acceptance, the MDR caps were set 10 basis points lower than the physical POS and online debit card acceptance infrastructure.

In parallel, effective 01 January 2018, the government decided to bear MDR for two years on all debit card transactions valued up to Rs. 2,000.00. However, the government fixed

[1] Merchant Discount Rate or Merchant Discount Fee is a service charge that banks take from merchants accepting card/ digital payments, which is usually a certain percentage of the transaction amount. The MDR paid by merchants is shared between acquirer banks, issuer banks and the card payment networks.

the MDR at 0.4% for debit card transactions up to Rs. 2,000.00. In effect, due to the government's intervention, RBI's decision to allow banks to charge up to 0.9% as MDR for businesses with annual turnover of Rs. 20.00 lakh or more (even for transaction amounts less than Rs. 2,000.00), got overruled and the banks got only 0.4% as MDR for such transactions.

Corresponding to this MDR of 0.4%, the interchange[2] fixed by card payment networks is 0.15%. Thus, RBI's MDR mandates could never get implemented since the government felt otherwise on small ticket sized transactions up to Rs. 2,000.00 and reduced the MDR to zero for all merchant categories and restricted the banks to receive no more than 0.4% as MDR.

In fact, National Payments Corporation of India (NPCI) was the only network to adopt lower-than-cap MDR. The MDR pricing structure arrived at (effective October 2019) for RuPay debit card had been 0.4% (0.3% when the transaction is QR-code based) for transactions up to Rs. 2,000.00 and 0.6% (0.5% when the transaction is QR-code based) for transactions exceeding Rs. 2,000.00, with a ceiling on MDR of Rs. 150.00 for any transaction.

1.2. The present *avatar* of MDR

Effective 01 January 2020, the government decided not to bear MDR any further on all debit card transactions valued up to Rs. 2,000.00.[3] In effect, due to this decision, RBI's mandate got re-invoked and banks got the leverage to charge MDR @ 0.9% or less from businesses with annual turnover of Rs. 20.00 lakh or more for transactions of any value. Furthermore, for businesses with annual turnover of less than Rs. 20.00 lakh, banks got the freedom to impose an MDR of 0.4% or less.

However, the government simultaneously brought in a new law where RuPay debit card had been identified as a prescribed payment mode such that banks and system providers could no longer charge any fee to the merchants for whom they setup the payment acceptance infrastructure. Consequently, any charge, including the MDR, was no longer applicable on payments made through RuPay debit card.

While taking such a step, the government envisage that among low-cost digital modes of payment, RuPay debit cards (and not mastercard/VISA debit cards) will promote less cash economy through their extensive use for P2M (person-to-merchant payment) transactions. The underlying philosophy is that neither merchants nor consumers should get any feel of extra cost while adopting such digital modes of payment. An impression given is that RBI and banks will be able to absorb the associated costs from the savings that will accrue to them on account of handling less cash as people move to these digital modes of payment.

1.3. The law

The government under Section 10A of the Payment and Settlement Systems (PSS) Act, 2007, indicate that no bank or system provider shall impose any charge upon a person making or receiving a payment by using the electronic modes of payment prescribed under section 269SU of the Income-tax Act, 1961. In the Income-tax Rules, 1962, a new Rule 119AA has been inserted that prescribed RuPay debit card and BHIM-UPI as the electronic modes of payment for the purpose of Section 269SU.

---

[2] Interchange or issuer interchange is the share of the MDR that the issuer bank keeps as their commission. Thus, MDR comprises of the interchange and the acquirer's commission.

[3] Earlier banks were getting reimbursement of 0.4% MDR for transactions up to Rs. 2,000.00 from MeitY.

Technically, this implies that banks shall not levy any charges to a person for payments made or received through RuPay debit card and BHIM-UPI.

2. Banks and System Providers

To harmonise a way forward for the banks, early January 2020, the Indian Banks' Association (IBA) made a move when 15 major banks came together to decide what should be the MDR for mastercard/VISA debit cards. Considering the government's agenda of promoting digital transactions and encouragement to merchants for promoting low value transactions, the IBA indicate that the banks reached a consensus that for transactions up to Rs. 2,000.00, the applicable MDR should be 0.4% irrespective of the merchant category.

Like the government, IBA too did not feel it appropriate to charge a high MDR of up to 0.9% for merchants having annual turnover of Rs. 20.00 lakh or more. This, possibly raises doubt on RBI's December 2017 regulation, where it had removed the concept of MDR based on ticket size and set a very low benchmark of Rs. 20.00 lakh to categorise small and medium merchants. Actually, that may have triggered the government to intervene, and now IBA too.

The norms recommended by the banks that mostly prevail are:

a) MDR cap should be 0.4% for businesses with annual turnover of less than Rs. 20.00 lakh.
b) For businesses with annual turnover of Rs. 20.00 lakh or more,
    i. MDR to be capped @ 0.4% for transactions up to Rs. 2,000.00.
    ii. MDR to be capped @ 0.9% for transactions above Rs. 2,000.00.

2.1. Moving towards a lower controlled MDR

IBA's new norm is not quite in sync with RBI's attempt in January 2018 to eliminate the concept of MDR based on ticket size (which had been in place since 2012 for transactions up to Rs. 2,000.00). It is felt that had RBI's merchant categorisation been more rationale, we would not have seen intervention by the government with such vigour.

In all this zero or low MDR mess, what had been at stake is the banks' and system providers' revenue losses from large merchants like Amazon, Big Bazaar, IRCTC and the like, where they have been kept at par with small and medium merchants, with respect to small ticket transactions up to Rs. 2,000.00. Moreover, it does not make sense to see only RuPay debit cards offering zero MDR to small and medium merchants while mastercard/VISA debit cards imposing MDR @ 0.4% for transactions up to Rs. 2,000.00.

What is possibly missing is the payment industry's will to see a modification of the merchant categorization, where a lower controlled MDR is set for large merchants after a more rational merchant categorised than RBI's present categorization. There is an urgent need for a reasonable definition of large merchants for the purpose of MDR.

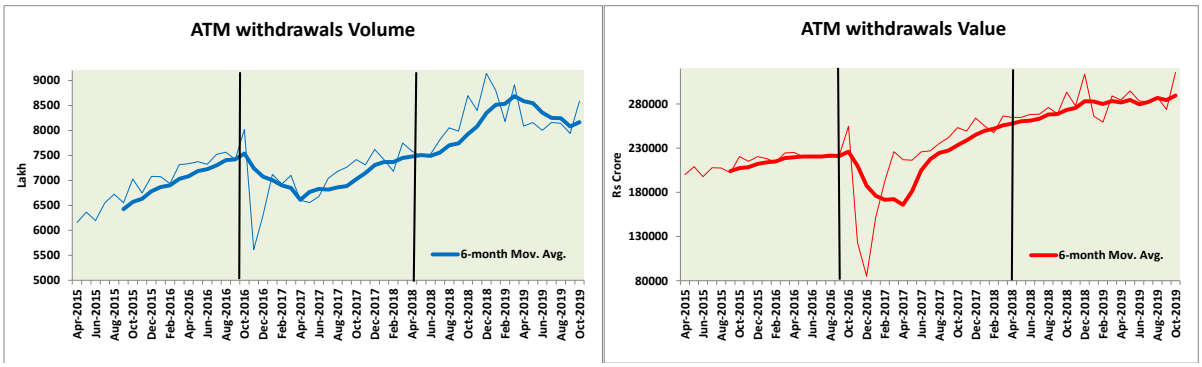
3. From Cash to Cards

Debit cards are extensively used by bank account holders towards cash withdrawal at ATM. Currently RBI and banks are absorbing significant costs while they provide cash as a prominent mode of payment. The promotion of excessive cash needs to be mitigated in such a

way that it not only reduces cash handling costs for the banks but also saves enough to support digital payments.

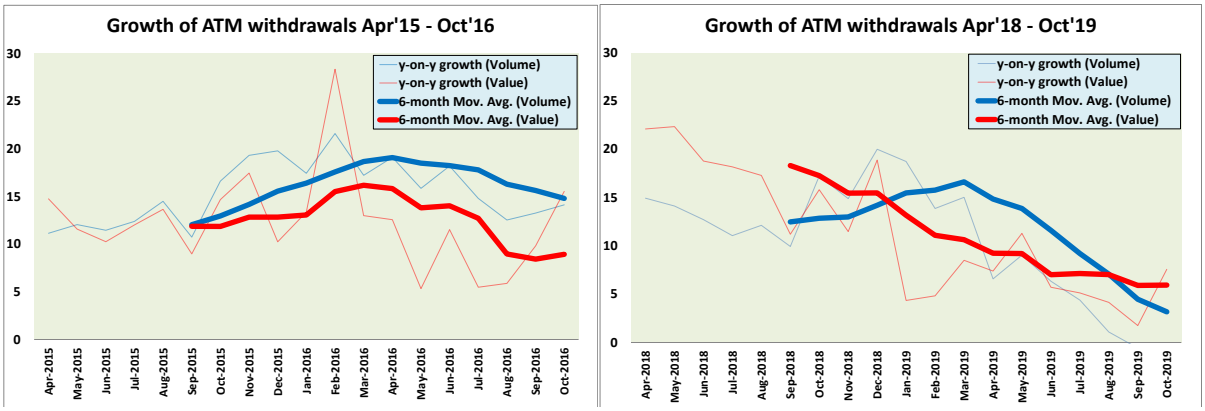
3.1. Cash from ATM

Cash is predominantly promoted in India with 8 to 10 free ATM withdrawals. This potentially amounts to bank's disbursement of up to Rs. One lakh of free cash per month to an individual holding a bank account. While keeping 17 months of wash-out period in between the pre- and post-demonetisation periods, Figures 1 and 2 show how the debit card usage at ATM behaved during the pre-demonetisation period April 2015 – October 2016 and the post-demonetisation period April 2018 – October 2019. The figures indicate that though in absolute terms there has not been any significant respite from predominant ATM usage in the country, there are some signs of reduced y-on-y growth in later months. The period November 2019 – July 2020 has been dealt separately since effective November 2019, RBI in its monthly ATM data dissemination has changed the definition of the ATM usage.



Source: RBI data

Figure 1: ATM withdrawals during the pre- and post-demonetisation (after wash-out period)

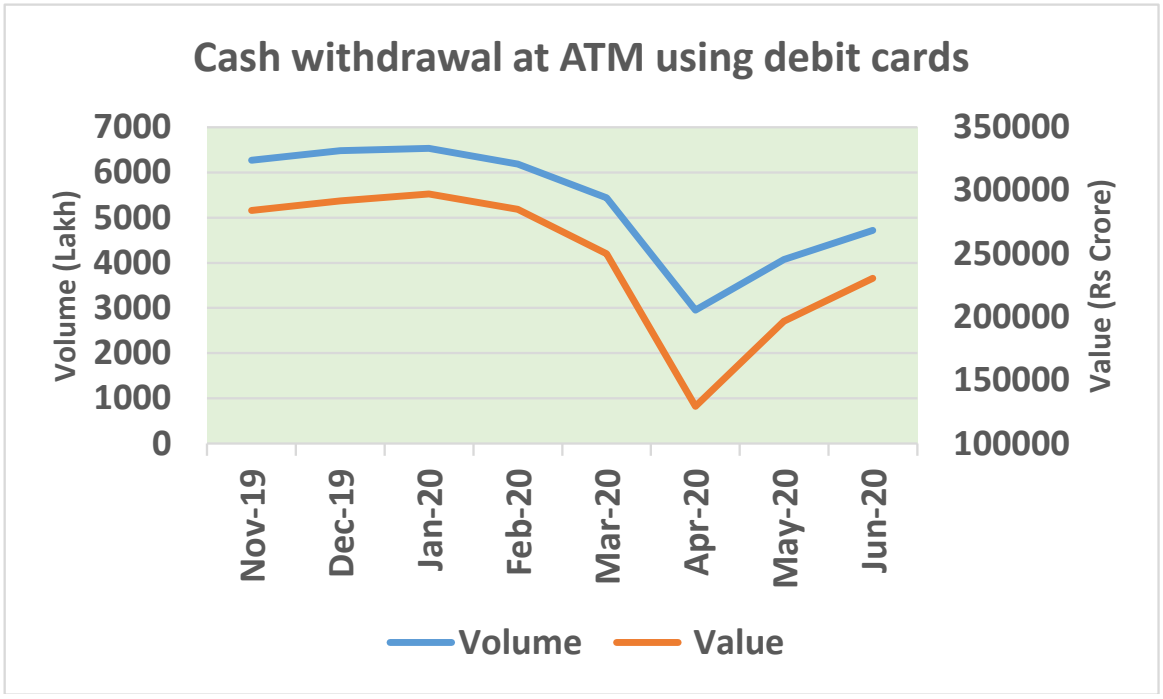


Source: RBI data and author's computation

Figure 2: y-on-y growth of ATM withdrawal during the pre- and post-demonetisation (after wash-out period)

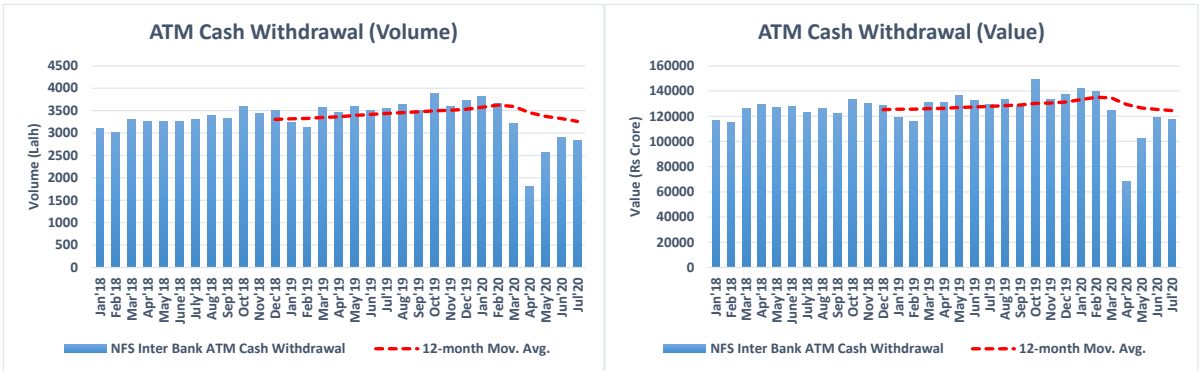
Figure 3 shows how the ATM cash withdrawal using debit cards behaved during the period November 2019 – June 2020. The period since March 2020 shows the effects of

COVID-19 lockdown and partial unlocks. As such, even if we supplement with more recent National Financial Switch (NFS) off-us ATM data till July 2020, as shown in Figure 4, we see no respite from predominant ATM cash withdrawals in the country.



Source: RBI data

Figure 3: Cash withdrawal at ATM



Source: NPCI and RBI data

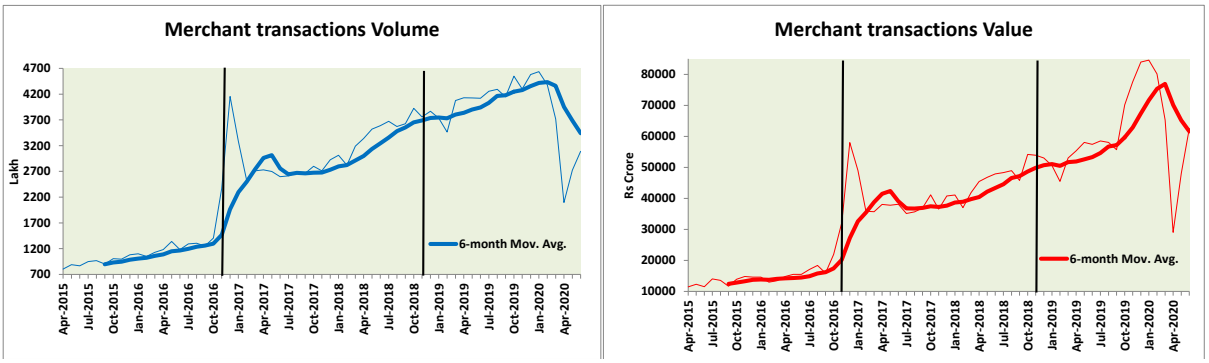
Figure 4: NFS inter bank ATM cash withdrawal

But for RBI's mandate allowing significant amount of cash withdrawal free for many bank customers, technically speaking, banks would not have incurred such avoidable and non-remunerating expense. There is nothing that RBI appears to have done as a deterrent, which strongly prompts a reduction of large amounts of cash withdrawal in a month. Digital payment modes are now amply available where large and frequent cash is still in use. May be RBI advocating banks of charging a fee in a tiered fashion for total cash withdrawals in excess of a reasonable amount, say Rs. 20,000.00 a month, could create enough deterrent. Such a move

would allow generating desirable revenue for the banks to meet cash handling costs and additionally support costs for digital payments infrastructure.

3.2. Merchant transactions using debit cards

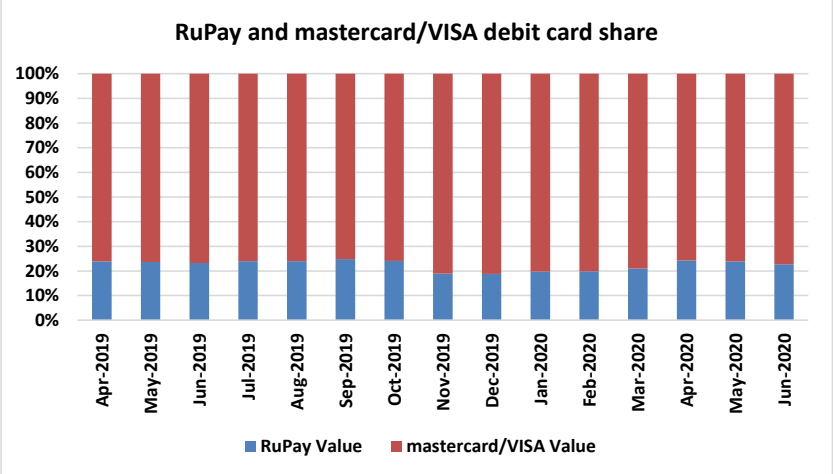
Figure 5 shows the trend for debit card merchant payments during the period April 2015 – June 2020 where we have indicated a wash-out period of 12 months to account for the disturbances due to impact of demonetisation. Clearly, under a controlled MDR-revenue model, debit card usage and acceptance for merchant payments has shown a consistent growth for some time now (exception being the COVID-19 lockdowns and partial unlocks). This has an associated cost for which revenue is collected either directly or indirectly from the users of the banking system.



Source: RBI data

Figure 5: Debit card merchant transactions during the pre- and post-demonetisation (after wash-out period)

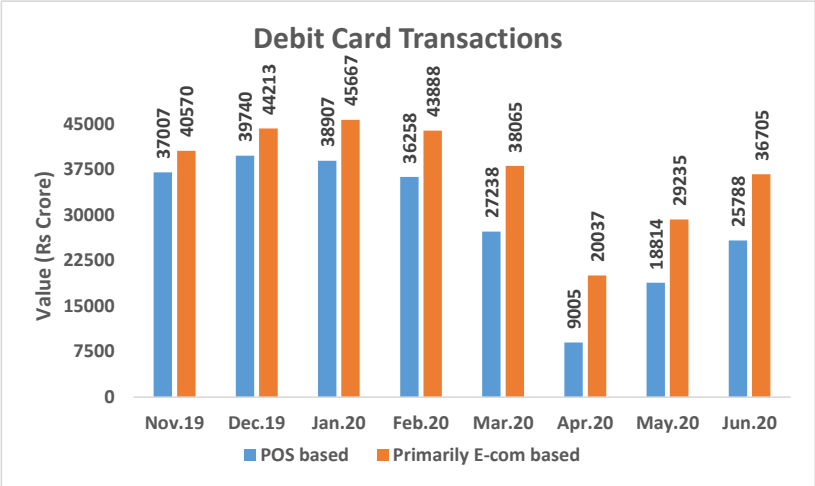
Figure 6 shows the percentage share of debit card transactions for RuPay and mastercard/VISA. During April 2019 – June 2020, in value terms, RuPay had an average share of only 22% whereas mastercard/VISA had 78%. In fact, even during January-June 2020, in value terms, RuPay had an average share of 22% whereas mastercard/VISA had 78%.



Source: RBI and NPCI data

Figure 6: Percentage share of RuPay and mastercard/VISA (in Value terms)

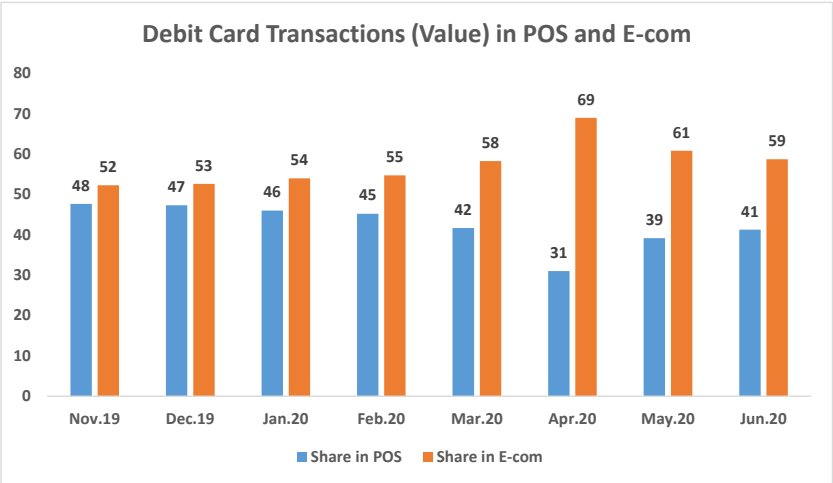
The use of debit cards, in value terms, is more prominent among E-com merchant transactions rather than physical POS transactions. Primarily, E-com constitute data on e-commerce transactions. However, though relatively meagre, RBI includes digital bill payments through ATMs and card to card transfers for debit cards under their E-com data. Primarily Figures 7 and 8 show the extent of E-com and POS transactions. Even before the COVID-19 pandemic, in value terms, the E-com transactions had been more than the POS transactions (though in volume terms it is only since April 2020 that we see the same trend). The effect of lockdown on physical retail shops and other services had its impact in increasing the gap between E-com and POS transactions.



Source: RBI

Figure 7: Share of POS and E-com merchant transactions using debit cards

In value terms, there had been a consistent increase in percentage share of E-com over POS transactions. During the first six months of 2020, percentage share of E-com had been 58%. For bank account based transactions, the trend in debit card acceptance for E-com is unavoidable unless BHIM-UPI becomes a better choice for all. Unlike E-com, cash is always an alternative for POS since POS is an expensive proposition for many small and medium merchants. However, such merchants need to be migrated and enable for the asset-lite BHIM-UPI acceptance.



Source: RBI

Figure 8: Percentage share of E-com and POS transactions (value)

4. Costs and Prices

We now need to address three pertinent questions.

(i) Given that there is zero MDR on RuPay alone, would there be an increase in the card acceptance desire at merchant locations?

(ii) If 'cost to merchant' is an important attribute for merchant's choice for card acceptance and the bank's desire to provide the card acceptance infrastructure, how would it impact continuation of the card acceptance trend?

(iii) What would be the consequence of the discriminatory approach adopted for RuPay?

The Watal Report[4] highlights the breakup of debit card transactions of less than Rs. 2,000.00 and more than Rs. 2,000.00 in value terms (see Table 1). Nearly 65% of the total values of debit card transactions fall in the sub Rs. 2,000.00 tickets category, and all these transactions would now attract MDR within the RBI set cap of 0.9% for about 78% of debit card usage amounts (*i.e.*, for mastercard/VISA but not RuPay).

As a result, we now see merchants paying MDR for sub Rs. 2,000.00 tickets, contributed by 78% of the debit card usage. The banks' re-imposition of debit card MDR @ 0.4-0.9% has affected the *small and medium* merchants when mastercard/VISA cards are used. This being the only source of MDR revenue in the debit card business for banks, it creates a strong potential for RuPay debit cards (constituting about 22% of the total values of debit card transactions) being marginalized in due course. As there is a revenue differential for banks between RuPay and mastercard/VISA, banks would always, in their commercial interest, tend to promote that card scheme which generates more revenue for them.

Table 1: Distribution of debit card transactions in value terms

Table 6.10.: *Breakup of acquirer gross margins**

Type of transaction	% of transactions value by value	MDR charged to merchants	Issuer interchange fees	Scheme fees	Acquirer gross margin
Debit Regular $\lesssim$ INR 2,000	20%	0.75%	0.50%	0.07%	01.8%
Debit Premium $\lesssim$ INR 2,000	15%	0.75%	0.65%	0.07%	0.03%
Debit Regular $>$ INR 2,000	10%	1.00%	0.75%	0.07%	0.18%
Debit Premium $>$ INR 2,000	10%	1.00%	0.90%	0.07%	0.03%
Prepaid	5%	1.00%	1.85%	0.07%	0.92%
Credit Regular	10%	1.30%	1.10%	0.07%	0.13%
Credit Premium	25%	1.90%	1.80%	0.70%	0.03%
Credit Corporate	5%	1.90%	2.00%	0.07%	0.17%

*Source: See, submission dated 28-10-2016 by Payment Council of India

Source: Shri Ratan P. Watal Report "Committee on Digital Payments – Medium Term Recommendations to Strengthen Digital Payments Ecosystem"

[4] Report of the Committee on Digital Payments headed by Shri. Ratan P Watal, December 2016. Ministry of Finance, Government of India.

4.1. The increased MDR burden for *small and medium* merchants

In Table 2, we present the debit card data for January-June 2020 and compute the merchant payoffs towards MDR for sub Rs. 2,000.00 ticket transactions. Prior to this, the government had made such merchant MDR zero for two calendar years 2018-19. Unlike pre-January 2020, for the first six month of 2020, when 78% of the sub Rs. 2,000.00 debit card transactions (in value terms) attracted MDR @ 0.4% from merchants (including small and medium merchants), it amounted to a total payoff of Rs. 755.00 crore. When projected (linear projection) for the full year 2020, the MDR payoff amounts to Rs. 1,511.00 crore. These figures would more than double when MDR is applied @ 0.9%.[5]

Table 2: Debit card transactions and computation of MDR payoffs

Value (Rs Crore)	Debit cards (1)	RuPay cards (2)	mastercard/ VISA cards (3)	Sub Rs 2000 transactions @65% of (3) (4)	MDR Revenue @0.4% of (4) (5)	MDR Revenue @0.9% of (4) (6)
Jan-2020	84575	16728	67846	44100	176	397
Feb-2020	80146	15902	64244	41758	167	376
Mar-2020	65303	13745	51558	33513	134	302
Apr-2020	29043	7051	21991	14294	57	129
May-2020	48049	11438	36611	23797	95	214
Jun-2020	62494	14180	48314	31404	126	283
Jan-Jun 2020	369609	79044	290565	188867	755	1700
Jan-Dec 2020					1511	3400

Source: RBI/NPCI data and author's computation

But for the COVID-19 lockdowns in the country, the level of debit card transactions would have been much higher and would have led to much larger MDR payoffs by the debit card accepting merchants. The government and RBI have effectively implemented a net increase of debit card MDR expenses at least for the small and medium merchants. Now, there are two contrary aspects to such an increase.

(A) Such an increase in the "cost to merchant" is good for the development and increased acceptance of debit cards, and
(B) The increased net MDR, onto small and medium merchants, is bad for the development and increased usage of debit cards.

If (A) holds (and merchants pay a controlled MDR), there was no need for the induced discrimination between RuPay and mastercard/VISA. The same could have been achieved by arriving at a lower controlled MDR, uniform across all card schemes.

However, if (B) holds, by putting restrictions only for RuPay and discriminatorily allowing mastercard/VISA to impose MDR onto merchants, we have not quite achieved desired results, unlike the government's strategy of zero MDR for merchants for the two calendar years 2018-19.

[5] The average ticket size on RuPay debit card transactions being less than Rs. 1,200.00, the benefit of zero MDR on RuPay debit cards on ticket sizes more than Rs. 2,000.00 is minimal.

We have to choose between (A) and (B).

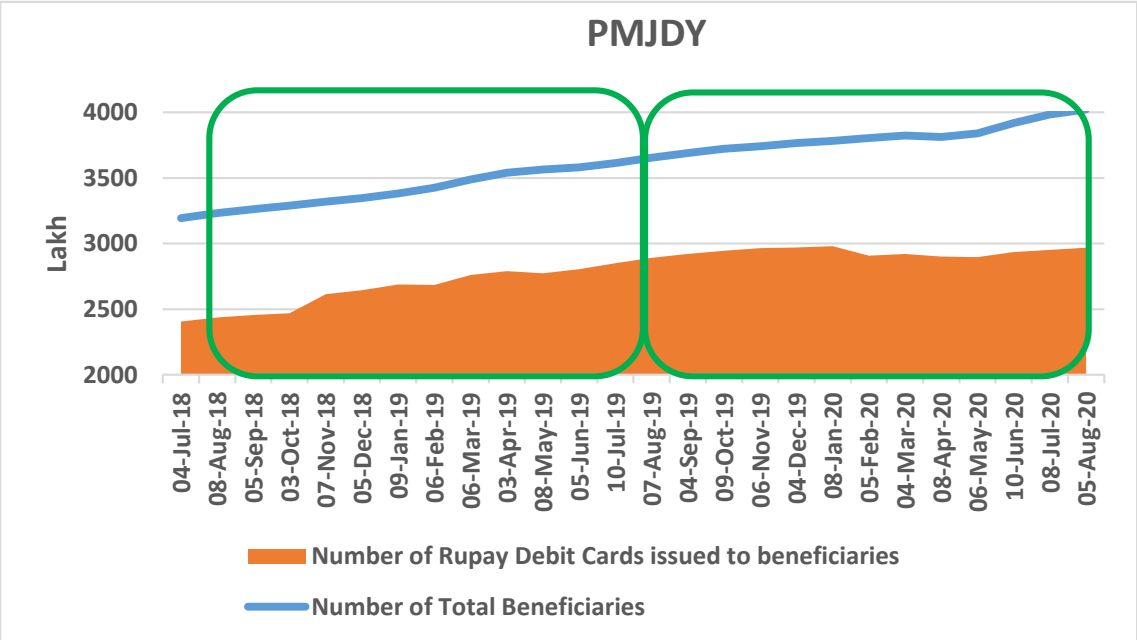
Even if we keep aside the issue of MDR, Section 10A of the Payment and Settlement Systems Act 2007 emphasises that no bank or system provider shall impose any charge upon anyone, either directly or indirectly, for using RuPay debit cards. If we can have a bundled pricing for RuPay card acceptance, the monthly/ yearly rentals for POS terminals and PGs would increase with rentals being attributed to mastercard/VISA (and not RuPay) debit cards. This would actually lead to an indirect charge being imposed on RuPay debit cards.

An important question remains as to whether zero MDR for RuPay debit cards would serve the purpose of promoting card payments in a situation where merchants are still overburdened from the fee for accepting other cards (cards other than RuPay debit cards). Could the answer lie in allowing merchants, at their discretion, not to accept cards other than RuPay debit cards? Surely not, since that may be a hindrance in their sales.

4.2. Discriminatory approach for RuPay debit cards

So how would the system work now without any revenue stream for RuPay debit card (a prescribed modes of payment)? Also, how would the system work in presence of the induced discrimination between RuPay on the one hand and mastercard/VISA on the other hand?

Though the zero MDR for RuPay debit card has led to savings for some merchants (of about Rs. 1,000.00 crore for calendar year 2020), an important question remains as to whether it would serve the purpose of promoting card payments in the presence of merchants being still overburdened on the fee for accepting other cards (cards other than RuPay debit cards). Note that for mastercard/VISA, effective January 2020, the merchants no longer enjoy zero MDR on transactions up to Rs. 2,000.00. Could the answer lie in allowing merchants, at their discretion, not to accept cards other than RuPay debit cards?



Source: Data submitted to DFS by Public Sector Banks, Regional Rural Banks and Major Private Sector Banks

Figure 8: RuPay debit card issued against PMJDY accounts

How should the government and the RBI solve this complex problem? If there is a revenue differential for banks between RuPay and mastercard/VISA, banks would always in their commercial interest have a tendency to promote that card scheme which generates more revenue for them. This is clearly reflected in Figure 8 where, during the one year period August 2018 through July 2019, there had been issuance of at least 455 lakh RuPay debit cards corresponding to at least 422 lakh new accounts added under PMJDY. In contrast, for the same tenure during August 2019 through July 2020, we see a subdued issuance of around 77 lakh RuPay debit cards despite at least 366 lakh new PMJDY accounts added. This showcases that banks have taken steps to move away from RuPay and promote a card scheme which generates more revenue for them.

5. Concluding Remark

Using debit card and ATM data, this paper prepares grounds for policy guidance. Pricing policy can be ideally based on economic and accounting principles. For a way forward for debit cards, based on cost to the banking industry and prices implicitly paid by bank depositors, we refer to Das A. (2020), "*Merchant transactions through debit cards – costs and prices*", IIT Bombay Technical Report (forthcoming).

References

Basic Statistical Returns of Scheduled Commercial Banks (SCBs) in India. Volume **47**, March 2018. RBI. 03 January 2019.

Basic Statistical Return BSR 2 - Deposits with Scheduled Commercial Banks (SCBs). March 2019. RBI, 15 November 2019.

Concept Paper on Card Acceptance Infrastructure. RBI, 08 March 2016.

Das, Ashish (2016a). *Incentivising ATM-Cash and Cheques Over Electronic Transactions - A Policy Gap*. IIT Bombay Technical Report. 26 January 2016.

Das, Ashish (2016b). *Promotion of Payments Through Cards and Digital Means*. IIT Bombay Technical Report. 24 April 2016.

Das, Ashish (2017). *Discovering the Right MDR for India*. IIT Bombay Technical Report. 28 February 2017.

Das, Ashish (2020). *Discriminatory Approach for RuPay Debit Cards: Some Suggestions for Corrective Measures*. IIT Bombay Technical Report. 07 January 2020.

Das, Ashish (2020). *Deviating from the BHIM-UPI Law*. IIT Bombay Technical Report. 24 August 2020.

Das, Ashish and Das, Praggya (2016). *Sanitising Distortions in Digital Payments*. IIT Bombay Technical Report. November 28, 2016.

Das, Ashish and Agarwal, Rakhi (2010). *Cashless Payment System in India - A Roadmap*. IIT Bombay Technical Report. May 31, 2010.

Promotion of payments through cards and digital means. Government of India (Ministry of Finance), 29 February 2016.

Report of the committee to review the ATM interchange fee structure. Report of the RBI constituted committee. October 2019.

Report of the committee on the analysis of QR (Quick Response) code. RBI. 22 July 2020.

Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Pune, during January 02-04, 2020; pp 175-180

Role and Importance of Statistics in Business Management

Rashmy Moray
Symbiosis Institute of Management Studies
Symbiosis International University (SIU), Pune

Received: 19 August 2020; Revised: 31 August 2020; Accepted: 01 September 2020

Abstract

Things that are managed are always measurable and termed as statistics. Use of statistical techniques in business management is vital in the contemporary scenario of competitive environment of globalisation. The various functional areas of business like marketing, finance, production, operations and HR encounters ample statistical data that has to be collected, processed, recorded and transmitted to the various stakeholders. The spectrum of statistics is widely used by the business administrators to run the business. Applying and using various statistical methods, techniques help the mangers to combat the business uncertainties. Considering this backdrop an attempt is made to highlight the role of statistics in business management.

Key words: Statistics; Business; Organisations; Finance; Marketing; HR.

1. Introduction

Envisioning the trade and industrial environment of the future is a key management function and statistics play a dynamic role in all the types of business dealings. Every commercial proposal starts with an extensive research and all the data collected is compiled into statistics for decision(Balaji, 2013). Facts, figures and statistics are connoted to exhaustively explain and measure uncertainty and allows the managers to predict the future. Organisations encounter abundant information across all the media platforms and produce database for strategic operations in the day-to-day functioning of their business. From the perspective of businessoperations, a financial analyst useswidespread financial information to guide their speculative avenues and draw conclusions whether the stock is under-pricedor overpriced. Marketing managers using the electronic scanners at retail checkout counters collect data for a variety of marketing research applications. Statistics helps the business in producing goods with limited variations and minimum wastage and increase in the workers' productivity(Scott, 2017). Considering the operationalvalue of the business, now-a-days all the trading concerns highly reply on information technology (IT) systems to manage data, facilitate remittances and run day to day operations. To overcome the obstacles, statistical algorithms are used. In process of steering the business direction statistics is used as device to forecast the future for strategic planning. Application of statistical models provides a base to predict the business future expenses and revenues which would help the business to adjust with new market developments and track the competitors' activities(Salleh, 2018). Against this backdrop let us look at a few details regarding the role and significance of statistics in the work of managing business. The objective is to portray the importance of statistics in business management.

Corresponding Author: Rashmy Moray
Email: rashmy.moray@sims.edu

2. Spectrum of Literature

According to Horace Secrist, Statistics is the science of collection, organisation, presentation, analysis and interpretation of numerical data. Various definitions are found to define statistics but it is universal in character when it comes to its application to the world of business and industry. What is measurable is always in statistics and the things that are measurable are always manageable. Fundamentally statistics is aggregate of facts collected in a systematic manner and that are affected by number of factors, numerically expressed, which can be estimated according to reasonable standard of accuracy and presented/placed in relation to each other for a pre-determined purpose.

In the olden days' statistics was considered as science meant purely for federal governance and used to collect data for law and order, defence strength, population wealth and welfare for formulating policies and practices. But statistics is a study which is pervasive in nature. It is the division of science dealing ingathering, assembling, analysing the data collated and depicting of inferences drawn from the samples to the whole population(Ryan Winters, 2010). It includes forecasting, planning, organising and decision making which are the primary activities under taken by the business manager.

Contemporary world of trade and commerce is more of art than science. The growing worldwide competitive race amongst the corporatescompels business managers to address uncertainties by applying technicalmeans and be objective decision makers. Elimination of uncertainties is inevitable but using statistics, business executives make informed decisions about their products, customers, operations by applying statistical thinking and methods. This demandsthe unravellingmuscle of statistics for business administrators in the field of marketing, finance and operations. Largely, knowledge of statistics enables the business operational leaders to identify the problem, describe its nature of existence, estimatealternate course of actions, approximate errors, monitor methodsand practices and allow them to compare actual results with standards and upon any deviations enables the managers to take appropriate corrective measures(Veena, 2014). Let us consider each of the functional areas of the business and deliberate how statistics discourses the need of business management.

Marketing research heavily relies on statistical techniques to bring in insights to the usual deliverables and volume of output. According to American Marketing Association (AMA) Marketing research is examining the company's marketing processes exhaustively and a marketing executive needs to collect and analyse ample volume of data related to market crescendos and target clienteles. Marketing strategy developed by the marketing executives depends upon the significant results of bazaarinvestigations which involves numerical methods for collection, using sampling techniques, analysing dataand assessing the effect of various marketing strategies. Sample customers from consumers' population are selected to understand the perception with respect to a particular product or service and their buying and spending behaviour. It also helps in predicting the future preferences and purchasing habits of the consumers.(Cremonezi, 2018). The inferences derived from the gathered information from these surveys tend to widen the aspect of the population. This in turn depends on the method and source of data collection which encompasses the basic areas of survey design and sampling. A poorly designed survey and insufficient sample may result in biased and misrepresenting outcomes affecting the business advertising and sales promotional strategies. In Marketing management statistics is used extensively in determining advertising expenditures influencing sales and increasing the bottom lines i.e. net operating profits of the business(Saxena Parikshit, 2011). Statistics in marketing help the

marketing executives to identify which consumer buys what product at what frequency, which consumer is an active complaint launcher, recognises the sales personnel level of productivity and which mode of media would help in reach out the consumers in the digital era.

From the perspective of production and quality control point of view statistics is very vital since the volume of output produced directly impacts the top line of the business that is gross revenues made from sales. A production manager uses statistical process control techniques to enhanceefficiency and quality(Veena, 2014). To enhance the production efficiency, the production executives, depend on application of control charts, sampling techniques and probability distribution ensuringimprovedmethods and produces leading to fall in manufacturingexpenses(Cremonezi, 2018). Organisations who have always thrived onto continuous improvement or heavily relied on quality assurance programs such as, Six Sigma or Lean Manufacturing, understands the significance of statistics. Empirical evidence submits that diverseintensities of statistical quality control methods application exist in the organisations with ISO-certified quality management systems Statistics provides a platform to quantify and regulate manufacturing processes to reduce variations in the occurrence of errors leading to waste ensuring consistency in the process of production. This ultimately helps in reduction in direct costs of material and labour(Williams, 2019).

With the help of statisticians'promotion of statistical quality control and its integration into quality management systems isexecutedtogether at the micro and the macro level of the production management. Application of statistical controlling measures in productions management provides a means of detecting error at inspection, creates a basis for attainable specifications, plugsblockages and concerned areas, reduces inspection costs leading to uniform quality of production. The use of statistics in quality improvements includes hypothesis testing regression analysis, statistical process control(SPC) that helps the production engineers and managers identify when methods and practices are beyond resistance, as a result of deviations caused by uncertain situations and that are not integralpart of the procedures. Finally helps in designing and analysis of experiments. Further if considering the use statistics in finance, chompingfiscal data and using financial techniques is avitalmeasureundertaken by financial analyst who are none other than are the financial actuaries using financial data.

Deliberating the role of statistics in finance, statistical investigation has become aninfluentialdevice for economic, industry and market valuation in the hands of financial managers. The explosive innovations in the complex financial instruments makes markets complex making more difficult for the investors, lenders and practitioners to value financial assets accurately. Statistical analysis has come to their rescue by offering the options to hedge the amount of risk associated with investment avenues. The best example of how statistics exponentially progressed is witnessed from the derivation and application of Black-Scholes formula (Black and Scholes, 1972 and Merton, 1973) for option pricing. These statistical tools have been used across all major funds analysis to forecast the performance and meeting the investment goals.

The behavioural aspects of investors are used in predicting the stock prices based on the current price statistics of the index and the individual stock movements. Technical analysis is a classic example of statistical data presented in the form of charts and bars used by investors and financial houses for decision making analysis. Finance main concern is related to the valuation of assets. It is related to future receipts and payments called as cash flows. Statistical data helps the organisations in predicting the futures with respect to these cash

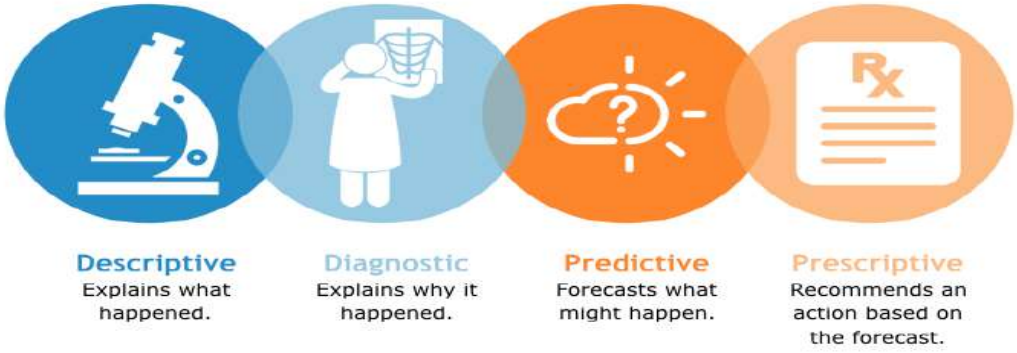
flows that impacts the earnings of the stakeholders. Practitioners use financial data for their own judgements regarding the investment positons in the stock markets(Gemeno, 2006).

The functional areas in accounting, like, financial accounting, managerial and cost accounting, auditing and taxation etc., includes the use of econometrics like regression analysis, time series analysis, discriminant analysis, probit, logit and Altman`s Zeta analysis are of statistical nature. These offer various means and mode to gather, analyse and evaluate data for decision making on the part of financial managers to anticipate, acquire and allocate financial resources for the business operations. Statistics is pervasive in nature when it comes to accounting and finance as usually surveys research on investors use of bookkeeping data, capital market effects (stocks, bonds, options *etc*.) pension data, commodity markets and adjusting inflation rates to asset values. Using statistics knowledge of how firms'records, process and reports and constructed and transmitted to various stakeholders of the business. In finance the statistical procedures and estimation techniques attempt to minimise the occurrence of loss by quantifying the risk in returns over investments(Robert Hamada, 1998). At the economy level financial statistics is considered asaninclusive repository of stock and data on the financial assets and liabilities of all sectors of an economy. It highlights the statistical data that condenses past actions or projects future behaviour with respect to individual financial security, group of securities and markets in a widely spread geographical territory. The crux of presenting the balance sheet is merely the statistical record of presented in a systematic manners following an accounting process. The collective financial data of all the companies, industries and economic sectors are the resultant outcome of statistical process that is presented in the form of economic parameters to analyse the nation's economic performance.

Finally let us have a bird's view over the Human resource management (HR), another important functional area of business management. The discipline HR and relevance of statistics is age old since personnel at work that are men in concept of management play a vital role among the five Ms that is, men, material, machine, money and markets. The focus on HR in management is important as men is the first factor that assembles all the other remaining factors. HR leaders in the organisation always on the edge to identify the changing perception of the employees with respect to upgradation in pay package based on their contributed productivity to the profitability of the organisation. HRM is a phenomenon to manage people in an organisation. The focus of the discipline HR is to develop strategies to improve the workforce experience with the rising industry trends. The statistics of employees is all about the data of workforce employed in an organisation with respect to their demographics, educational qualifications, strengths, weakness, skills, experience role and responsibilities. The purpose of HR management in an organisation is to identify the training needs to upgrade the employee skills to enhance efficiency and contribute to the organisational development.

The use of statistics in HR is primarily to measure the employees level of satisfaction with respect to job performance. Ever since the concept of globalisation has emerged corporates are facing employee retention and engagement problems. Statistics as discipline helps these organisation in creating, maintaining and upgrading the data related to recruitment, contracts, payroll, performance, insurance and benefits and training needs of their staff and division/departmental heads. Statistical data in hand enables organisations to optimise engagement practices for the employee retention and motivation. The concept of HR analytics is receiving attention of business managers in large firms. It is a combinations of two terms analysis and statistics which means any analysis driven through the application of

statistics(Sengupta, 2020). Companies used several arrays of metrics to assess the effect of HR initiatives with the emergence of HR metrics and HR scorecard in relation to efficiency effectiveness and the impact on business performance. Over a period, the use of statistical tool gained significance over metrics, leading to the use of analytics. In the recent times application of quantitative techniques in HR has been prompted by the datafication of HR (Bersin, 2013).This is how the role of statistics is profound in the world of business management. Further the picture below helps us to understand how it enables the managers to apply statistics in business decision making through various approaches to analytics.



Source: https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/

From the above depiction it can be comprehended that business enables managers to analyse past performance, predict future business practices and lead organizations effectively. And in the common parlance of statistics it can be in the form of descriptive; that summarises existing data showing cause and effect analysis which is usually used in functional areas of management, diagnostic that focuses on examining the past historical performance to identify and evaluate the causes behind the routines, predictive; using machine learning techniques, modelling an data mining that enabling the managers to predict future probabilities and trends based on historical data,*e.g.*, fraud detection and security, risk assessment*etc*.

3. Conclusion

From the above deliberation it is clear that in the age of technology statistics has wide range of applications. The discussion brings out clear deductions about the applications of statistics and its expediency in business managerial decision-making. Statistical methods used in day to day managerial aspects is a conspicuous reference for researchers, managers, consultants and academicians.In the fields of business, management science, operations research, supply chain management, financial econometrics and economics understanding statistical literature and applying quantitative practices is the need of the hour to combat business uncertainties and take smart decisions on a day to day front. To conclude researchers always are of the opinion that statistics enables the business managers to analyse past performance, predict future business practices and lead organizations effectively. In this entire process, descriptive analysis, predictive analysis and prescriptive analysis is key to success in business management.

References

Balaji, R. (2013). Role of Statistics on Business Research. *International Journal of Engineering Research and Technology* (*IJERT*), **2**(**10**).

Bersin, J. (2013). *https://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/#1a4ca3464cd0.* Retrieved from https://www.forbes.com: https://www.forbes.com

Cremonezi, L. (2018). Introducing statistics in marketing research. *Ipsos Connect*.

Gemeno, R. (2006). *https://iase-web.org/documents/papers/icots7/5G2_GIME.pdf.* Retrieved from https://iase-web.org.

Radha, K. and Scott, E. (2017). *https://blog.essaycorp.com/importance-of-statistics-in-business/.* Retrieved from https://blog.essaycorp.com/importance-of-statistics-in-business/

Robert Hamada, J. M. (1998). The role of statistics in accounting, marketing, finance and production. *Journal of Business Economics and Statistics,* **6**(**2**), 261-272. doi:10.2307/1391563

Ryan Winters, A. W. (2010). Statistics a brief overview. *Ochsner Journal,* **10**(**3**), 213-216.

Salleh, S. (2018). *https://yourbusiness.azcentral.com/importance-statistics-management-decision-making*-25518*.html.* Retrieved from https://yourbusiness.azcentral.com: https://yourbusiness.azcentral.com/importance-statistics-management-decision-making-25518.html

Saxena, Parikshit (2011). Application of statistical techniques in market research: A sample survey. *International Journal of Applied Engineering Research, Dindigul*, 163-171.

Sengupta, R. (2020). *https://www.analyticsinhr.com/blog/hr-analytics-statistics-introduction/.* Retrieved from https://www.analyticsinhr.com.

Veena (2014). *https://iba.ac.in/statistics-for-managersb/.* Retrieved from https://iba.ac.in: https://iba.ac.in/statistics-for-managers/

Williams, J. T. (2019). *The Importance of Statistics in Management Decision Making.* Retrieved from https://smallbusiness.chron.com.

Special Proceeding of the 22 nd Annual Conference of SSCA held at Savitribai Phule Pune University, Ganeshkhind, Pune, MH, during January 02–04, 2020; pp 181-192

High Frequency Financial Data and Associated Financial Point Process: A Nonparametric Bayesian Perspective

Anuj Mishra and T. V. Ramanathan
Department of Statistics and Centre for Advanced Studies in Statistics
Savitribai Phule Pune University, Pune 411 007, India

Received: 22 August 2020; Revised: 03 September 2020; Accepted: 06 September 2020

Abstract

This paper review the theoretical framework of modeling high frequency financial data (HFD) using a point process approach. We represent the financial event arrival times as the realization of non-homogeneous Poisson process with an intensity function $\lambda(t)$, which is assumed to be periodic. In the case of HFD, this periodic pattern is quite well known, as the intensity of trades is higher in the morning and just before closing the market and lower during the afternoon. We make an attempt to study this intra-day cyclic behaviour with an intensity modelling approach using Bayesian nonparametric method. The posterior consistency of the proposed nonparametric Bayesian procedure is established. The Bayesian estimation of the intensity function is described for a specific case where the conditions of the prior are satisfied. This paper is just a first step towards modelling the HFD using point process approach and the corresponding Bayesian nonparametric analysis. It may be mentioned that lot more computations need to be done to complete this ongoing work.

Key words: Bayesian nonparametrics; Financial point processes; High frequency financial data; Intensity function; Intensity process.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The empirical studies in finance literature usually concentrate on opening, closing or average prices of stocks from financial markets. However, due to technological advancements, researchers can now work with the high frequency data (HFD), which contains details of all the transactions along with the marks such as price, volume, time of transaction etc. Such data has attracted lot of researchers and this has become a new area of research these days. Known as 'high-frequency finance', it helps to understand the financial markets at a micro level, see Viens *et al.* (2011), Gregoriou (2015) and Florescu *et al.* (2016) for a broad overview. Here, there can be details of hundreds of transactions happening in a micro time interval corresponding to a particular stock from an electronic stock exchange. An important feature of HFD is that the transactions are recorded as and when they occur, hence the observations are irregularly time spaced. This prevents the use of standard time series methods in high frequency finance. The timing of transactions carry substantial amount of information, which can be used in studying the micro structure of a financial market. Therefore, it is very important to model the time interval between transactions (durations) appropriately.

Corresponding Author: T. V. Ramanathan
Email: ram@unipune.ac.in

One way of modeling the durations is to make use of the autoregressive-type conditional duration (ACD) models introduced by Engle and Russell (1998), which attempts to model the time between the events occurring at t_{i-1} and t_i, by defining $x_i = t_i - t_{i-1}$, where $\{t_0,t_1,...,t_n,...\}$ is a sequence of arrival times with $0 = t_0 \leq t_1 \leq t_2 \leq ... \leq t_n \leq ....$ The sequence $\{x_1,x_2,...\}$ of non negative random variables form the durations. Here, the arrival times t_i may not necessarily mean the time corresponding to consecutive trades. There can be other type of events of interest as well. For example, t_i can be time of occurrence of a volume event, which is said to have occurred if the cumulative trade volume since the last volume event is at least a preset amount v. Similarly, price event is said to have occurred if the cumulative price change since last price event is at least of a preset amount p. Thus, x_i is the interval between consecutive events of interest leading to trade, volume or price durations.

Let the conditional expected duration be

$$\psi_i = E(x_i|\mathcal{F}_{i-1}), \tag{1}$$

where, $\mathcal{F}_i$ is the information set at transaction i, that is, $\mathcal{F}_i = \sigma(x_i,x_{i-1},\cdots,x_1)$. The main assumption of an ACD model is that the durations are of the form

$$x_i = \psi_i \, \epsilon_i, \tag{2}$$

where ϵ_i are independent and identically distributed (i.i.d) random variables with $E(\epsilon_i) = 1$ (In fact, without loss of generality, it is possible to assume that this is true). The above set up is very general and it allows a variety of models which can be obtained by choosing different specifications for the expected duration ψ and different distributions for ϵ. We refer to two interesting review papers Pacurar (2008) and Bhogal and Ramanathan (2019) for detailed discussions on this approach.

Another approach of modeling HFD is using a point process. In this method, one represents the event of arrival times as a realization of non-homogeneous Poisson process with an intensity function $\lambda(t)$ having a specific structure. This method is usually known as *financial point process* method. A further extension of this is modelling based on the intensity function of the process, which leads to more flexible and powerful models. Such an approach is recommended when we deal with multivariate processes, in which case, the conditional duration approach is not very successful (see Russell (1999), Hautsch (2004) and Bauwens and Hautsch (2006)).

Researchers have been using the periodicity adjustment procedure of Engle and Russell (1998), which we have also used to remove intra-day effect in a paper published recently Mishra and Ramanathan (2017). However, WSu (2012) had claimed that this procedure is not very satisfactory. Hence, using a non-homogeneous Poisson process may resolve the issue, as suggested by various researchers in related problems. Belitser *et al.* (2013) proposed an M-estimator to estimate the period of a cyclic non-homogeneous Poisson process, established its consistency and demonstrated the effectiveness by applying it to a call center data. In our case, we already know the period, which is 'daily' and hence we are not interested in estimating the period. Weinberg *et al.* (2007) modelled the day-to-day as well as intraday variations in the same call center data using a normal approximation to Poisson. Specifically, in finance, Andersen *et al.* (2019) develop a procedure to test intra-day periodicity in return volatility. We propose a procedure to adjust the periodicity using a Bayesian approach and prove its consistency.

Bayesian approaches for Hawkes models have received much less attention. The only contributions for the Bayesian inference are due to Gulddahl (2013) and Blundell *et al.* (2012) who explored parametric approaches and used MCMC to approximate the posterior distribution of the parameters. Donnet *et al.* (2018) study the properties of Bayesian nonparametric procedures in the context of multivariate Hawkes processes.

The paper is organized a follows. In Section 2, we discuss the financial point process. Bayesian nonparametric approach for financial point process is described in Section 3. Extensions to Hawke-type processes is described in Section 4. Section 5 concludes with some future directions.

2. Financial Point Processes

Let $\{t_i, i = 1, \cdots, n\}$ denote a random sequence of increasing event times $0 < t_1 < \cdots < t_n$ associated with an orderly (simple) point process. Then,

$$N(t) = \sum_{i \geq 1} I_{\{t_i \leq t\}}$$

define a right-continuous counting function which gives the number of events of some type in the time interval $(0,t)$. The $\mathcal{F}_t$-intensity process $\lambda(t)$ of the counting process $N(t)$ is defined as

$$\lambda(t;\mathcal{F}_t) = lim_{\Delta \to 0} \frac{1}{\Delta} E[N(t+\Delta) - N(t) | \mathcal{F}_t], \tag{3}$$

where $\mathcal{F}_t = \sigma\{N(s); 0 \leq s \leq t\}$. Therefore, the sequence of event arrival times $\{t_i\}$ can be modeled as a point process by modelling the intensity $\lambda(t)$.

The simplest type of point process is the homogeneous Poisson process defined by

$$\begin{aligned} Pr((N(t+\Delta) - N(t)) = 1 | \mathcal{F}_t) &= \lambda\Delta + o(\Delta), \\ Pr((N(t+\Delta) - N(t)) > 1 | \mathcal{F}_t) &= o(\Delta), \end{aligned} \tag{4}$$

with $o(\Delta)/\Delta \to 0$, as $\Delta \to 0$. Note that in this case, the intensity is constant and it leads to

$$P(t_i > x) = P(N(x) < i) = \sum_{j=0}^{i-1} \frac{e^{-\lambda x}(\lambda x)^j}{j!}.$$

A straight generalisation from here is the case when the intensity function is a deterministic function of time or a non-homogeneous Poisson process with intensity function $\lambda(t)$. This can be particularly useful in modelling the intra-day cyclic behaviour of durations with an appropriate choice of $\lambda(t)$. One another possibility here is to use a marked point process defined with marks such as arrival of buys, sells and certain limit orders, see Bauwens and Hautsch (2006).

3. Bayesian Point Processes

Let $(N_t)_{t \geq 0}$ be a non-homogeneous process on $[0,T]$, that is, the sample paths of $(N_t)_{t \geq 0}$ are right-continuous step functions with $N_0 = 0$ and with jumps of size 1. Let N_t be the number of jumps in $[0,t]$ and $N_t < \infty$ almost surely. We assume the following about the process $N(t)$ and the intensity function $\lambda(t)$.

A1 For any disjoint subsets $B_1, B_2, \cdots, B_m \in \mathcal{B}([0,T])$, $\mathcal{B}([0,T])$ the random variables $N(B_1), N(B_2), \cdots, N(B_m)$ are independent random variables denoting the number of jumps in $B_1, B_2, \cdots, B_m$ respectively.

A2 For any $B \in \mathcal{B}([0,T])$ the random variable $N(B)$ is distributed as Poisson with parameter $\Lambda(B)$; where Λ is a finite measure on $([0,T], \mathcal{B}([0,T]))$, called as the compensator of the process.

A3 Λ admits a density λ with respect to the Lebesgue measure on $\mathcal{B}([0,T])$. That is,

$$\Lambda_t = \int_0^t \lambda(s) ds,$$

where, $\lambda(t)$ is called the intensity of N_t.

We estimate $\lambda(t)$ using Bayesian procedure and establish the posterior consistency. Consistency results for intensities of Poisson processes can be established by connecting and extending the two main approaches regarding consistency for i.i.d. observations. The first approach, due to Schwartz (1965), Barron *et al.* (1999), and Ghosal *et al.* (1999), requires construction of an increasing sequence of sets, a sieve, and a sequence of uniformly consistent tests. This is the approach followed by Belitser *et al.* (2015) and Donnet *et al.* (2018). An alternative approach, which we follow in this paper, provided by Walker (2004) relies on a martingale sequence to obtain sufficient conditions for posterior consistency in the i.i.d. case. This alternative approach, though equivalent to the use of a suitable sieve, simplifies the verification of necessary conditions for consistency.

Below we state Theorem 1.3 of Kutoyants (1998) as a lemma.

Lemma 1: For any λ, the law P_λ of N under the parameter value λ admits a density p_λ with respect to the measure induced by a standard Poisson point process with intensity 1. This density is given by

$$p(\lambda)=exp(\int_0^T log\lambda(t)dN_t-\int_0^T(\lambda(t)-1)dt).$$

Suppose we observe n independent non-homogeneous Poisson processes $N^{(1)},N^{(2)},\cdots,N^{(n)}$ on $[0,T]$ with a common intensity λ, which is a positive integrable function on $[0,T]$. Then, by Lemma 1, the likelihood is given by

$$L(\lambda)=\prod_{i=1}^{n}exp(\int_0^T log\lambda(t)dN_t^{(i)}-\int_0^T(\lambda(t)-1)dt).$$

We define the parameter space as

$$\mathcal{F}=\left\{\lambda:[0,T]\to R_+|\int_0^T\lambda(t)dt<\infty\right\}.$$

Here, we are estimating the intensity function $\lambda(t)$, given $N^{(1)},N^{(2)}\cdots N^{(n)}$, using a Bayesian nonparametric approach. Let λ belong to the class $\mathcal{F}$ of intensities which need not be indexed by a finite dimensional parameter. Let Π be a prior on $\mathcal{F}$, $\Pi:(\mathcal{F},\sigma(\mathcal{F}))\to[0,1]$. Let $\Pi(.|N^{(1)},N^{(2)},\cdots,N^{(n)})$ stand for the posterior distribution of λ given the data. So, if B is a set of intensities, the posterior mass assigned to it is given by

$$\Pi(B|N^{(1)},N^{(2)},\cdots,N^{(n)})=\frac{\int_B R_n(\lambda)d\Pi(\lambda)}{\int_{\mathcal{F}}R_n(\lambda)d\Pi(\lambda)},$$

where

$$R_n(\lambda)=\prod_{i=1}^{n}\frac{p(N^{(i)},\lambda)}{p(N^{(i)},\lambda_0)}$$

is the likelihood ratio with $\lambda_0\in\mathcal{F}$ being the true fixed but unknown transition density.

The Bayesian model is consistent if the posterior mass increases around λ_0 as n increases. Suppose that a topology on $\mathcal{F}$ has been specified. Then posterior distribution is said to be consistent at λ_0 if for every neighborhood U of λ_0, we have that,

$$\Pi(U^c|N^{(1)},N^{(2)},\cdots,N^{(n)})\to 0 \quad a.s.$$

3.1. Posterior consistency

For a continuous function f on $[0,T]$ we define the norms $\|f\|_2$ and $\|f\|_\infty$ as usual by defining

$$\|f\|_2=\int_0^T f^2(t)dt, \text{ and } \|f\|_\infty=sup_{t\in[0,T]}|f(t)|.$$

In the following theorem, we establish the posterior consistency of Bayesian procedure under a couple of conditions on the prior assumed and on the space of intensity functions.

Theorem 1: For any given $\epsilon>0$, let $A_\epsilon^{'}$ be a set of intensities around the true intensity $\lambda_0\in\mathcal{F}$ defined as,

$$A_\epsilon^{'}=(\lambda\in\mathcal{F}:\|\sqrt{\lambda}-\sqrt{\lambda_0}\|_2>\sqrt{2}\epsilon). \tag{5}$$

Let the prior Π be such that

a) $\Pi(\lambda:\|\lambda-\lambda_0)\|_\infty<\epsilon)>0$ and

b) $\sum_{j=1}^{\infty}\sqrt{\Pi(A_j^{'})}<\infty$, where $\{A_j^{'}\}$ is a countable cover of size $\frac{\delta}{\sqrt{2}}(\delta<\epsilon)$ for $A_\epsilon^{'}$.

Then,

$$\Pi(A_\epsilon^{'}|N^{(1)},N^{(2)},\cdots,N^{(n)})\to 0\ a.s. \tag{6}$$

Condition (a) is similar to Belitser *et al.* (2015) and Donnet *et al.* (2018). But we have an easily verifiable condition (b), as compared to other conditions provided by the same authors.

Let the square of the Hellinger distance $h(p_\lambda,p_{\lambda'})$ be defined as

$$h^2(p_\lambda,p_{\lambda'})=2\left(1-E_{\lambda'}\left(\sqrt{\frac{p_\lambda(N)}{p_{\lambda'}(N)}}\right)\right),$$

where E_λ is the expectation corresponding to the probability measure under which the process N is a Poisson process with intensity function λ. Let the Kullback-Leibler divergence $K(p_\lambda,p_{\lambda'})$ be defined by

$$K(p_\lambda,p_{\lambda'})=-E_{\lambda'}\left(log\left(\frac{p_\lambda(N)}{p_{\lambda'}(N)}\right)\right).$$

We state two lemmas before proving Theorem 1. Lemma 2 constitutes a part of Lemma 1 of Belitser *et al.* (2015) and Lemma 3 is nothing but Theorem 4 of Walker (2004), which gives posterior consistency result for density estimation. Proofs of these lemmas are omitted as they are available in the references mentioned.

Lemma 2: For the Hellinger distance $h(p_\lambda,p_{\lambda'})$ and the Kullback-Leibler divergence $K(p_\lambda,p_{\lambda'})$, the following hold good.

(i) $\frac{1}{\sqrt{2}}\|\sqrt{\lambda}-\sqrt{\lambda'}\|_2\le h(p_\lambda,p_{\lambda'})\le\sqrt{2}(\|\sqrt{\lambda}-\sqrt{\lambda'}\|_2)$

(ii) $\|\lambda-\lambda_0)\|_\infty\le K(p_\lambda,p_{\lambda'})$

Lemma 3: Let A_ϵ be a set of intensities defined in terms of densities with respect to Poisson measure p_λ h-bounded away from p_{λ_0},

$$A_\epsilon=(\lambda\in\mathcal{F}:h(p_\lambda,p_{\lambda_0})>\epsilon). \tag{7}$$

Assume that the prior Π has the following properties:

C1. $\Pi(K(p_\lambda,p_{\lambda_0})<\epsilon)>0$

C2. $\sum\limits_{j=1}^{\infty}\sqrt{\Pi(A_j)}<\infty$, where $\{A_j\}$ is a countable h-cover of size $\delta(<\epsilon)$ for A_ϵ.

Then,

$$\Pi(A_\epsilon|N^{(1)},N^{(2)},\cdots,N^{(n)})\to 0 \ a.s. \tag{8}$$

Proof: *of Theorem 1:* We can view our problem as a density estimation problem with respect to the Poisson measure, and as a consequence, we have the posterior consistency of Lemma 3. By Lemma 2(i), we have from (7) and (5), $A_\epsilon \subset A_\epsilon^{'}$, which gives

$$\Pi(A_\epsilon)\to 0 \Rightarrow \Pi(A_\epsilon^{'})\to 0. \tag{9}$$

By Lemma 2(ii), we have

$$\Pi(\lambda:\|\lambda-\lambda_0\|_\infty<\epsilon)>0 \Rightarrow \Pi(\lambda:K(p_\lambda,p_{\lambda_0})<\epsilon)>0. \tag{10}$$

Again, by Lemma 2(i), if $\{A_j\}$ is a countable h-cover of size $\delta(<\epsilon)$ for A_ϵ, then $\{A_j^{'}\}$ is a countable cover of size $\frac{\delta}{\sqrt{2}}(\delta<\epsilon)$ for $A_\epsilon^{'}$. Moreover, we also have, again by Lemma 3,

$$\sum_{j=1}^{\infty}\sqrt{\Pi(A_j^{'})}<\infty \Rightarrow \sum_{j=1}^{\infty}\sqrt{\Pi(A_j)}<\infty \tag{11}$$

We have shown in (10) and (11) that condition (a) and (b) of Theorem 1 implies Lemma 3(C1) and 3(C2). Under these conditions, we have the result (8) of Lemma 3, which implies the result (6) of Theorem 1 by (9). □

Thus we have proved a general result. We can use Theorem 1, to obtain the suitability of specific priors. For clarification, we illustrate an example in the subsection below, but would like to emphasize that these results may also be used to verify the posterior consistency of other priors.

3.2. Illustration

There is a huge literature on prior construction for nonparametric models, where space of functions serve as the parameter space, see Chapter 2 of Ghosal and van der Vaart (2017) for details. The simplest and most popular is the random basis expansion, which we discuss briefly.

Given a set of basis functions $\phi_j:[0,T]\to R$, one way of constructing prior on intensity functions $\lambda:[0,T]\to R_+$ is by writing $\lambda=exp(\sum_{j=0}^{\infty}\beta_j\phi_j)$ and putting priors on the coefficients β_j in this representation. There can be many choices of bases, such as, polynomials, trigonometric functions, wavelets, splines, spherical harmonics, etc. See De Boor (1978) for details on splines, Härdle *et al.* (2012) and Donald and Percival (2000) for details on wavelets and Fourier bases. Also see Appendix E of Ghosal and van der Vaart (2017) for their approximation properties. For a given application, the suitability of the prior is determined by the approximation properties of the basis together with the prior on the coefficients. Rivoirard and Rousseau (2012) discussed a very general adaptive priors based on wavelets and Fourier bases. Shen and Ghosal (2015) and Belister *et al.* (2014) carry out a similar study for priors based on spline basis. Shen and Ghosal (2015) assumed the knots as fixed, whereas, Belister *et al.* (2014) considered them as random.

We outline the procedure of estimation without getting into the details of computations.

Given infinite basis functions, the convergence of λ is not guaranteed always. However, it is true if and only if $\sum\beta_j^2<\infty$ *a.s.*. Let $\beta_j\sim N(0,\sigma_j^2)$. To ensure that λ defines a valid intensity function with probability 1, it is sufficient that $\sum\sigma_j<\infty$. Conditions for posterior consistency can be obtained by applying Theorem 1. Walker (2004) studied the same prior with adjustments for density estimation case and obtained a sufficient condition as $\sum(\frac{\sigma_j}{\omega_j})^{2m-\frac{1}{2}}<\infty$ for some sequence ω_j satisfying $\sum\omega_j<\infty$. Basically, if we want to use this basis representation as a prior on intensity

functions, then we will put priors on β_j, which will induce a prior on intensity functions. Also we want the prior on β_j to shrink to zero at an appropriate speed. For example, by taking $\sigma_j \propto j^{-1-q}$ for any $q>0$, this will be satisfied (put $\omega_j \propto j^{-1-r}$ for any $r>0$).

What we have provided till now is not a method of estimating the intensity, but, mentioned the conditions that are to be verified, in order to justify the Bayesian procedure.

3.3. Bayesian estimation procedure using HFD

In order to estimate the intensity using the Bayesian procedure, appropriate choice of prior needs to be made along with different bases, and then the posterior has to be computed. The purpose here is not to estimate the intensity in detail, but, to throw some light on the procedure using a particular case. The result will be true for any Bayesian procedure where the conditions on the priors are satisfied.

Let n denote the observed number of days of trading activity and Tn the total time in seconds during which market operates everyday. For example, if market timing is 10 AM to 4 PM, then $T=0$ *sec* corresponds to the time 10:00:00 and $T=21{,}600$ *sec* corresponds to the time 16:00:00. Then the complete event counting process is given by $N=\{N_t : t\in[0,nT]\}$, where N_t denotes the number of events in $[0,t]$. The assumption of periodicity implies that $\lambda(t+T)=\lambda(t)$, $\forall\ t\geq 0$. For $i=1,\cdots,n$, the event arrival counting process during day i may be defined as

$$N_t^{(i)} = N_{(i-1)T+t} - N_{(i-1)T}, \quad t\in[0,T].$$

Since the increments of the process N_t are independent, the processes $N_t^{(i)}$ are independent non-homogeneous Poisson processes with λ as the intensity function, restricted to $[0,T]$.

Our objective is to estimate the intensity function. Let Δ be a small grid of time, say 15 seconds, and $m=T/\Delta$ be the number of grids per day in the data set. Then the number of events in the j^{th} time grid on day i is given by

$$A_{ij} = N_{j\Delta}^{(i)} - N_{(j-1)\Delta}^{(i)},$$

which is assumed to follow a Poisson distribution with mean $\lambda_j = \int_{(j-1)\Delta}^{j\Delta}\lambda(t)$, for every $i=1,\cdots n$ and $j=1,\cdots m$. We denote the available data over grids as $A^n=(A_{ij} : i=1,\cdots,n, j=1,\cdots,m)$. Hence, the likelihood is given by

$$L(\lambda|A^n) = \prod_{i=1}^{n}\prod_{j=1}^{m}\frac{\lambda_j^{A_{ij}}\exp\lambda_j}{A_{ij}!}. \tag{12}$$

Putting a prior of the random basis type on λ, we have

$$\lambda = \sum_{k=1}^{J}\beta_k\phi_k. \tag{13}$$

This finite basis representation may not be truly non-parametric in nature. However, this problem can be addressed by putting another prior on J (see Chapter 2, Ghosal and van der Vaart (2017)). Therefore, a draw from prior Π can be constructed as follows:

1. Draw J from a Poisson distribution with mean μ (around 12).
2. Given $J=j$, draw $\underline{\beta}$ vector of j-independent $N(0,\sigma_j^2)$

Given the data, likelihood (3.8) and the prior (3.9), we can use MCMC to sample from posterior distribution, as it will be difficult to obtain the analytical form of the posterior. As long as the bases

considered are fairly easy to evaluate and integrate, we can compute the likelihood and posterior up to a normalising constant, without making any approximation. After that the computations will be a straightforward adoption of existing MCMC methods Polasek (2012). Also, having the data, form of the prior and the likelihood, we can use *Stan* programming language Carpenter *et al.* (2017), which does the posterior computation automatically using the Hamiltonian Monte Carlo (HMC) Betancourt (2017). While using HMC, proper care needs to be taken as a discrete parameter J is introduced into the model, and HMC does not work in discrete parameter case. However, properly augmenting HMC with Gibbs-type update for discrete parameter may circumvent this problem. This demand further investigations in this front.

We need to do a detailed computation considering all the bases and their variations in order to recommend an optimal choice of the prior and the basis for an intra-day periodicity adjustment. To carry out this task with the data, there has to be some additional changes in the structure of the intensity function. For example, here, we have considered the case of a Poisson process where the intensity does not depend on the observations. But in real cases we can incorporate a dynamic behaviour into the structure of the intensity function, so that, it actually affects the intensity of events in the market. This can be achieved by using the Hawkes- type models for the counts. Next we briefly discuss the Hawkes process and some of its generalizations.

4. Extensions to Hakwes-type Processes

A different generalization of the Poisson process is obtained by specifying $\lambda(t)$ as a (linear) self-exciting process given by

$$\lambda(t)=\mu+\int_0^t w(t-u)dN(u)=\mu+\sum_{t_i<t} w(t-t_i) \tag{14}$$

where $\mu(t)$ provides a Poisson base for the process and $w(u)$ is a kernel (exciting). This process is known as Hawkes process and was first proposed by Hawkes (1971) and was applied in seismology. The generalisation capability in Hawkes process over Poisson comes from kernel, which allows contribution by an event that occurs at a previous time $t-k$ to intensity at time t. This kind of a dynamic behaviour is not supported by the Poisson process.

following are some of the kernels that are frequently used in the case of Hawkes process.

1. **Exponential kernel**: The exponential kernel is given by
$$w(u)=\alpha\beta e^{-\beta u},\ u>0. \tag{15}$$
This kernel implies an exponential decay in the effect of an event on future events and β drives the strength of the time decay and α, the overall strength of excitation.

2. **Power-law kernel**: The power-law kernel is given by
$$w(u)=\frac{\alpha\beta}{(1+\beta u)^{1+p}},\ u>0 \tag{16}$$
This kernel implies a hyperbolic decay and capture long range dependence.

Another approach lies in using kernels that take the form of linear combination of exponential and/or power functions with different rate constants, which might help in capturing different short and long range dependence.

The intensity given in (4.1) can also be generalised to accommodate marks. There may be a mark, χ_i associated with the the event t_i, that can affect the intensity. For example, a trade with large volume may excite future trades more than a trade with small volume. Marks may be contained within the kernel function as $w(u,\chi)$. The Hawkes process with such a modification is known as Marked Hawkes process.

Multivariate Hawkes models can be also obtained by a generalization of (4.1). In such a case, $\lambda(t)$ is a $K\times 1$ vector defined by $\lambda(t)=(\lambda^1(t),\cdots,\lambda^K(t))$ with

$$\lambda^k(t)=\mu^k(t)+\sum_{m=1}^{K}\sum_{t_i^k<t} w_{mk}(t-t_i^k). \tag{17}$$

The function w_{mk} is a cross-exciting term with $w_{mk}(t-t_i^k)$ being the contribution to the intensity of type-m events made by a type-k event at t_i^k .

The probabilistic properties of Hawkes processes are discussed in Hawkes (1971), and Brémaud and Massoulié (1996). Hawkes and Oakes (1974) show that every self-exciting Hawkes processes can be represented as a Poisson cluster process. Thinking of each event as a parent, an event occurring at time t_i gives birth to offspring according to a Poisson process with intensity $w(t-t_i)$: these offspring generate their own offspring, and so on. Ogata (1978) discusses the maximum likelihood estimation of Hawkes process, whereas, asymptotic behaviour of such an estimate is investigated by Ozaki (1979).

Despite their usefulness, Hawkes-type models did not find their place in financial econometrics for a long time. Bowsher (2007) applied Hawkes type model in financial econometrics for the first time. He presented a continuous time, bivariate point process model ((4.4) with K=2) of the timing of trades and mid-quote changes for a New York Stock Exchange stock. Estimation was performed using maximum likelihood method as analytic likelihoods were available. Since then, there has been various developments in finance related to applications of Hawkes process. Bacry *et al.* (2012) introduced a non-parametric estimation method for multivariate Hawkes processes based on the spectral factorization of the co-variance matrix and then applied it to tick-by-tick trades data of a futures contract for a total period of 3.5 months. Da Fonseca and Zaatour (2014) proposed an estimation strategy using the method of moments that can be solved almost instantaneously as against the maximum likelihood estimates, and applied to trade arrival times of major stocks for observations of 2 years. Fauth and Tudor (2012) used multivariate marked point processes in order to describe the fluctuation in tick-by-tick data corresponding to trades in currency exchange (EUR,GBP,CHF,JPY). Hawkes (2018), Bacry *et al.* (2015) and Bauwens and Galli (2009) gave excellent reviews of applications of point process in finance. In spite of various applications after Bowsher (2007), there has been no Bayesian study of point processes in finance. Bayesian methods can be advantageous while obtaining the uncertainty about the intensity using the spread of the posterior distribution.

5. Concluding Remarks

In this paper, we have discussed the financial point process associated with the high frequency financial data. With the nonhomogeneous assumption of the count process associated with the durations, it is appropriate to estimate the intensity function $\lambda(t)$ using a nonparametric functional approach. We have addressed the problem using a nonparametric Bayesian method. This review paper is just a first step towards the description of the problem and the associated research. An extensive computational exercise needs to be undertaken by considering different basis combinations for intra-day periodicity adjustment. We also plan to extend this study by investigating various theoretical as well as computa-

tional aspects of Bayesian nonparametric approach of modeling the HFD using Hawkes-type models.

Acknowledgements

Authors are thankful to Professor V. K. Gupta, President, Society of Statistics, Computer and Applications for recommending this paper to the special proceedings of the Society's annual conference. T. V. Ramanathan would like to acknowledge the financial support from the Science and Engineering Research Board, Department of Science and Technology, Government of India, Grant reference SR/S4/MS: 866/13.

References

Andersen, T. G., Thyrsgaard, M., and Todorov, V. (2019). Time-varying periodicity in intraday volatility. *Journal of the American Statistical Association*, **114(528)**, 1695-1707.

Bacry, E., Dayri, K., and Muzy, J.-F. (2012). Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, **85(5)**, 157.

Bacry, E., Mastromatteo, I., and Muzy, J. F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, **1(01)**, 1550005.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, **27(2)**, 536-561.

Bauwens, L. and Galli, F. (2009). Efficient importance sampling for ml estimation of scd models. *Computational Statistics and Data Analysis*, **53(6)**, 1974-1992.

Bauwens, L. and Hautsch, N. (2006). Stochastic conditional intensity processes. *Journal of Financial Econometrics*, **4(3)**, 450-493.

Belitser, E. and Serra, P. (2014). Adaptive priors based on splines with random knots. *Bayesian Analysis*, **9(4)**, 859-882.

Belitser, E., Serra, P. and Zanten, H. V. (2013). Estimating the period of a cyclic non-homogeneous poisson process. *Scandinavian Journal of Statistics*, **40(2)**, 204-218.

Belitser, E., Serra, P. and Zanten, H. V. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *Journal of Statistical Planning and Inference*, **166**, 24-35.

Betancourt, M., Byrne, S., Livingstone, S. and Girolami, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, **23(4A)**, 2257-2298.

Bhogal, S. and Ramanathan, T. (2019). Conditional duration models for high-frequency data: a review on recent developments. *Journal of Economic Surveys*, **33(1)**, 252-273.

Blundell C., Heller, K. A. and Beck, J. M. (2012). Modelling reciprocating relationships with Hawkes processes. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, **25**, 2600-2608.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, **141(2)**, 876-912.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, **24(3)**, 1563-1588.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.,

Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76(1)**, 1-32.

Da Fonseca, J. and Zaatour, R. (2014). Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, **34(6)**, 548-579.

De Boor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.

Donald B. and Percival, A. T. W. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge.

Donnet, S., Rivoirard, V., Rousseau, J. and Scricciolo, C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli*, **24(1)**, 231-256.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, **66(5)**, 1127-1168.

Fauth, A. and Tudor, C. A. (2012). Modeling first line of an order book with multivariate marked point processes. *arXiv: Trading and Market Microstructure*, https://arxiv.org/pdf/1211.4157.pdf .

Florescu, I., Mariani, M., H.E., S., and Viens, F. (2016). *Handbook of High-Frequency Trading and Modeling in Finance*. Wiley, New York.

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

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, Cambridge University Press, Cambridge.

Gregoriou, G. N. (2015). *Handbook of High Frequency Trading*. Academic Press, New York.

Gulddahl, R. J. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, **15(3)**, 623-642.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (2012). *Wavelets, Aapproximation, and Statistical Applications*, Lecture Notes in Statistics, Springer Science and Business Media, New York.

Hautsch, N. (2004). *Modelling Irregularly Spaced Financial Data: Theory and Practice of Dynamic Duration Models*. Springer, New York.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58(1)**, 83-90.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, **18(2)**, 193-198.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, **11(3)**, 493-503.

Kutoyants, Y. A. (1998). *Statistical Inference for Spatial Poisson Processes*. Springer-Verlag, New York.

Mishra, A. and Ramanathan, T. (2017). Nonstationary autoregressive conditional duration models. *Studies in Nonlinear Dynamics and Econometrics*, **21(4)**, https://doi.org/10.1515/snde-2015-0057.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, **30(1)**, 243-261.

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, **31(1)**, 145-155.

Pacurar, M. (2008). Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. *Journal of Economic Surveys*, **22(4)**, 711-751.

Polasek, W. (2012). Handbook of Markoc Chain Monte Carlo edited by Steve Brooks, Andrew Gelman, Galin Jones, Xiao-Li Meng. *International Statistical Review*, **80(1)**, 184-185.

Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, **7(2)**, 311-334.

Russell, J. R. (1999). Econometric modeling of multivariate irregularly-spaced high-frequency data. Manuscript, GSB, University of Chicago.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **4(1)**, 10-26.

Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, **42(4)**, 1194-1213.

Viens, F., Mariani, M., and Florescu, I. (2011). *Handbook of Modeling High-Frequency Data in Finance*. Wiley, New York.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics*, **32(5)**, 2028-2043.

Weinberg, J., Brown, L. D., and Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, **102(480)**, 1185-1198.

Wu, Z. (2012). On the intraday periodicity duration adjustment of high-frequency data. *Journal of Empirical Finance*, **19(2)**, 282-291.

Publisher
Society of Statistics, Computer and Applications
B - 133, Ground Floor, C.R. Park, New Delhi - 110019
Tele: 011 - 40517662
https://ssca.org.in/
statapp1999@gmail.com
2020

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203