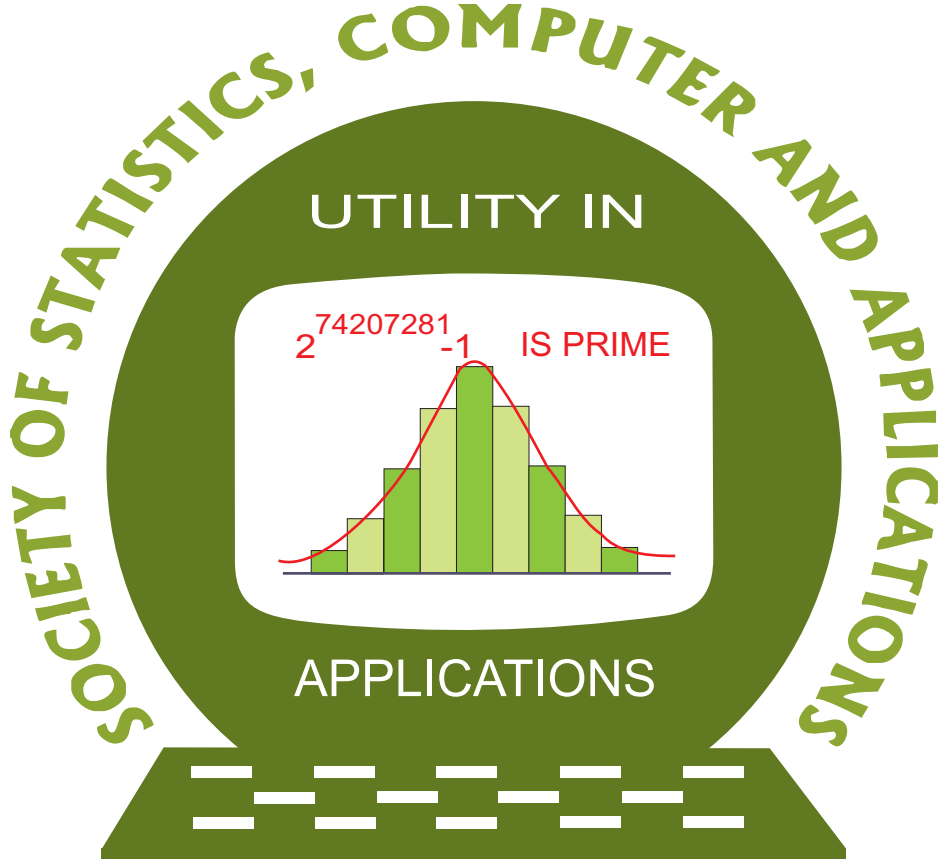
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Vol. 18, No. 1, 2020 (New Series)

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

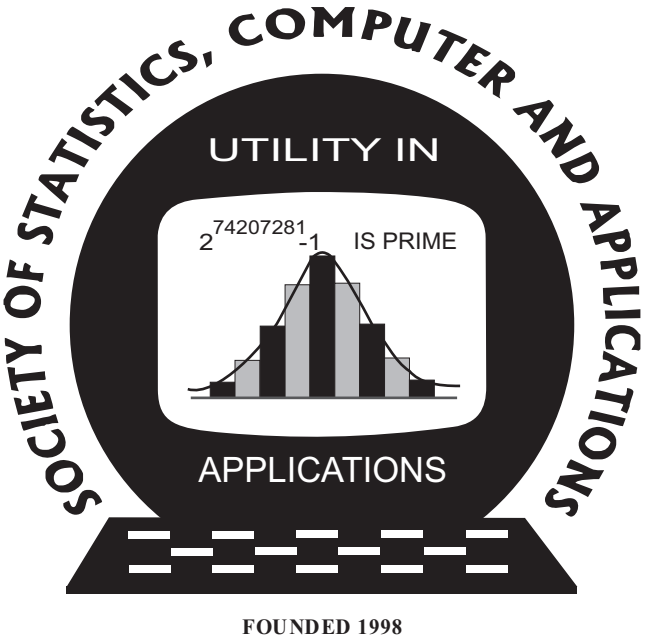
Ex-Officio Members (By Designation)

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

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

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 18, No. 1, 2020 (New Series)

Statistics and Applications

Volume 18, No. 1, 2020 (New Series)

Editorial Panel

Sanjay Chaudhuri, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546; stasc@nus.edu.sg
Sat N. Gupta, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC 27412, USA; sngupta@uncg.edu
Saumyadipta Pyne, Scientific Director, Public Health Dynamics Laboratory, Graduate School of Public Heath, University of Pittsburg, USA; spyne@pitt.edu
Shyamal D. Peddada, Graduate School of Public Health, University of Pittsburgh, PA 15261 USA; Shyamal.Peddada@gmail.com
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA; chatt019@umn.edu
T.V. Ramanathan; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
Tapio Nummi, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
Tathagata Bandyopadhyay, Indian Institute of Management Ahmedabad, Gujarat; tathagata.bandyopadhyay@gmail.com, tathagata@iima.ac.in
Tirupati Rao Padi, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
V. Ramasubramanian, ICAR-IASRI, Library Avenue, PUSA, New Delhi – 110012; ram.vaidhyanathan@gmail.com

CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 18, No. 1 (New Series) 2020

PREFACE

The first case of someone suffering from COVID-19, the disease caused by the novel coronavirus SARS-CoV-2, can be traced back to 17 November 2019, according to media reports on unpublished Chinese government data. However, the "patient zero" is yet to be confirmed. On 31 December 2019, Wuhan Municipal Health Commission, China, reported a cluster of cases of pneumonia in Wuhan, Hubei Province. A novel coronavirus was eventually identified. On January 30, 2020, Director General, WHO, declared coronavirus outbreak as Public Health Emergency of International Concern (PHEIC). On 18 March 2020**,** WHO and partners launched the Solidarity Trial, an international clinical trial, that aims to generate robust data from around the world to find the most effective treatments for COVID-19. The first recorded case of COVID-19 outside of China was confirmed on 13 January 2020 in Thailand. The first case in India was reported on 30 January 2020 in Kerala. Deeply concerned both by the alarming levels of spread and severity, and by the alarming levels of inaction, WHO made the assessment that COVID-19 can be characterized as a pandemic on 11 March 2020. As of now, 215 countries and territories around the world are under the influence of this virus.

COVID-19 has caused severe disruption across the world. The virus has posed an unprecedented threat to both public health and human lives and the economy of the countries globally. There are variations both among and within countries in the number of cases and in case fatality rates. In general, Europe and the United States have borne the brunt of the infection, while Asian and African countries have been relatively less affected so far. It is not yet clear why such geographical differences exist; they cannot be explained by healthcare facilities alone. The impact of COVID-19 on human health and loss of lives has been devastating. The number of persons tested positive is staggering; so is the number of deaths caused by this virus. One of the questions getting asked the most these days is when will the world be able to go back to the way things were in December 2019 before the pandemic struck. However, this appears to be unlikely to happen soon. Several countries are racing towards developing vaccines to fight the disease while re-opening their economies. But without rapid and effective global cooperation, the world may not exit this crisis safely at all.

In March 2020, Saumyadipta Pyne approached V.K. Gupta with great concern and curiosity about the much-needed scientific response to the COVID-19 pandemic, and suggested that SSCA must respond timely to this global crisis. It was then that we decided to devote a large part of the Volume 18, Number 1 (2020) July issue of our journal 'Statistics and Applications' to papers dedicated to the pandemic.

The major purpose of this decision was to motivate the researchers to address the different aspects of this time-sensitive and challenging problem. It was felt that manuscripts based on relevant approaches such as mathematical modelling, biostatistics, bioinformatics, spatial and time series data analysis, epidemiology, health economics, operations research, group testing, etc., will be considered for submission to this issue.

We did have a realization that the authors may present their early research findings in the form of a manuscript that can lead to future major projects and more in-depth studies. Given the time constraints, known techniques of modelling, data mining, correlations with climate and associated variables, available genomic data, geographical positioning, foreign travel, age and other demographic parameters, patients with comorbidity, etc., could form the basis of their study. For instance, even without a definite solution, a study might offer insights into the disease dynamics, its processes and hotspots, etc., especially by analysing the COVID-19 data

as reported publicly by many reliable sources. It would be a necessary scientific endeavour to relate the disease dynamics with demographics, social distancing and lockdown policies, climate, the immunization strategies, genetic variations, antibody testing and early adoption of countrywide preventive measures, and so on. Marking statistically sound hotspots would enable governments to localize the people with positive cases, and then release the rest of the people to return to work and resume normal living. Both human life and livelihood are equally important, and loss of either causes immense suffering that the scientific community must seek to alleviate.

Accordingly, authors from India and abroad were invited to submit their papers for this volume. Despite the pressure on time and the availability of quality data, we note with great pleasure and appreciation that the authors responded with great enthusiasm. However, we do wish to say that the timely availability of quality data and mechanisms and to obtain those will pave the way for better understanding of the dynamics and policy planning.

Thus, the volume 18, number 1 (2020), is divided into two sections. **Section A** contains **8** regular papers submitted to the journal. **Section B** comprises of **13** invited papers focused on the topic of COVID-19, which were also subjected to the usual peer review process.

It may be mentioned here that the views expressed in all the papers on COVID-19 are those of the author(s) only. The journal "Statistics and Applications" and the Editorial Board of the Journal are not responsible for any of the views expressed in these papers.

We would like to express our gratitude to all the authors for responding to our request and contributing to this volume. In fact, this volume has been possible because of the contribution from each one of them. The reviewers have also made a very big contribution by way of finishing the review process in a very short span of time. We are indebted to all the reviewers and thank them profusely for their support. The help received from Dr. Baidya Nath Mandal for bringing the papers in the format of the journal is highly appreciated.

Today, we also recognize that the pandemic COVID – 19 is either re-striking in some countries or continuing unabated in others. The situation is still far from clear and it is possible that many societies may have to re-organize their lives in different ways in a post-COVID world. We also appeal to statisticians, epidemiologists and other related scientists and the authors who contributed to this volume to continue their efforts in learning and unfolding the mysteries about this pandemic. This journal would continue to devote its space on priority to all such researches of current interest. We do hope that the papers in this volume would eventually culminate in some high-profile projects on the most critically important topic of the present.

V.K. Gupta
Saumyadipta Pyne
Rajender Parsad

July 2020

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 1-19

Analysis of Kidney infection Data Using 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: 15 February, 2020; Revised: 05 March, 2020; Accepted: 16 March 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 with three different baseline distributions namely, the Pareto,Burr and the linear failure rate distributions. We introduce the Bayesian estimation procedure using Markov Chain Monte Carlo (MCMC) technique to estimate the parameters involved in these models. We apply these models to a real life bivariate survival data set of McGilchrist and Aisbett (1991) related to the kidney infection data and a better model is suggested for the data.

Key words: Bivariate 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

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

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.

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,

$$h(t \mid z) = z h_0(t) e^{\mathbf{X}\boldsymbol{\beta}}, \tag{1}$$

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

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

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

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

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:

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

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. Baseline Distributions

5.1 Pareto Distribution

The Pareto distribution is a skewed, heavy-tailed distribution that is sometimes used to model the distribution of incomes. This distribution is not limited to describing wealth or income, but to many situations in which an equilibrium is found in the distribution of the ”small” to the ”large”. In insurance applications, heavy-tailed distributions are essential tools for modeling extreme loss, especially for the more risky types of insurance such as medical malpractice insurance. In financial applications, the study of heavy-tailed distributions provides information about the potential for financial fiasco or financial ruin. The Pareto distribution is great way to open up a discussion on heavy-tailed distribution. A continuous random variable T is said to follow the Pareto distribution with the scale parameter λ and the shape parameter α if its survival function is,

$$S(t) = (\lambda t + 1)^{-\alpha}; t > 0, \lambda > 0, \alpha > 0 \tag{19}$$

and the hazard function and the cumulative hazard function as

$$h(t) = (\alpha\lambda)/(\lambda t + 1); t > 0, \lambda > 0, \alpha > 0 \tag{20}$$

$$H(t) = \alpha \log(\lambda t + 1); t > 0, \lambda > 0, \alpha > 0 \tag{21}$$

Observe that $h(t)$ decreases with t ; $\lambda > 0, \alpha > 0$. Hence this distribution belongs to the decreasing failure rate class. The exponential and Rayleigh are the two most commonly used distributions for analyzing lifetime data. These distributions have several desirable properties and nice physical interpretations. Unfortunately the exponential distribution only has constant failure rate and the Rayleigh distribution has increasing failure rate. The linear failure rate distribution generalizes both these distributions. We consider this is the second baseline distribution.

5.2 Linear Failure Rate Distribution

The linear failure rate distribution of a continuous random variable T with the parameters $\alpha > 0$ and $\lambda > 0$, will be denoted by LFRD (α, λ) has the following survival function

$$S(t) = \exp(-\alpha t - \lambda/2t^2); t > 0, \lambda > 0, \alpha > 0 \tag{22}$$

It is easily observed that the exponential distribution (ED(α)) and the Rayleigh distribution (RD (λ)) can be obtained from LFRD(a,b) by putting $\lambda = 0$ and $\alpha = 0$ respectively. Moreover, the probability density function (PDF) of the LFRD (α, λ) can be decreasing or unimodal but the failure rate function is either constant or increasing only. See for example Bain (1974), Sen and Bhattacharya (1995), Lin et al. (2006), Ghitany and Kotz (2007) . The hazard function and the cumulative hazard function of linear failure rate distribution are respectively as stated below:

$$h(t) = \alpha + \lambda t; t > 0, \lambda > 0, \alpha > 0 \tag{23}$$

$$H(t) = \alpha t + \lambda t^2/2; t > 0, \lambda > 0, \alpha > 0 \tag{24}$$

5.3 Burr Distribution (Type XII)

The Burr XII distribution, having logistic and Weibull as special sub-models, is a very popular distribution for modeling life time data and for modeling phenomenon with monotone failure rates. When modeling monotone hazard rates, the Weibull distribution may be an initial choice because of its negatively and positively skewed density shapes. However, it does not provide a reasonable parametric fit for modeling phenomenon with non-monotone failure rates such as the bathtub shaped and the unimodal failure rates that are common in reliability and biological studies. Such bathtub hazard curves have nearly at middle portions and the corresponding densities have a positive anti-mode. Unimodal failure rates can be observed in course of a disease whose mortality reaches a peak after some finite period and then declines gradually. This distribution covers the curve shape characteristics for a large number of distributions. The versatility and flexibility of the Burr-XII distribution turns it quite attractive as a tentative model for data whose underlying distribution is unknown. A continuous random variable T with the parameters $\lambda > 0$ and $\alpha > 0$, will be denoted by Burr(λ, α) has the following survival function

$$S(t) = (1 + t^{\lambda})^{-\alpha}; t > 0, \lambda > 0, \alpha > 0 \tag{25}$$

Hazard function and Cumulative hazard function are

$$h(t) = \frac{\alpha\lambda t^{(\lambda-1)}}{1+t^{\lambda}} \tag{26}$$

$$H(t) = \alpha \log(1+t^{\lambda}) \tag{27}$$

6. Proposed Models

Substituting cumulative hazard functions for the Pareto, linear failure rate (LFR) and Burr baseline distributions in equation (18), we get the unconditional bivariate survival functions at time $t_{1j} > 0$ and $t_{2j} > 0$ as,

$$S(t_{,}t_{2j}) = \exp\left[\rho\frac{1-(1-2\sigma^2\eta_j(\alpha_1\log(\lambda_1+1)+\alpha_2\log(\lambda_2+1)))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j\alpha_1\log(\lambda_1+1))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j\alpha_2\log(\lambda_2+1))^{\frac{1}{2}}}{\sigma^2}\right] \tag{28}$$

$$S(t_{,}t_{2j}) = \exp\left[\rho\frac{1-(1-2\sigma^2\eta_j(\alpha_1 t+\lambda_1 t^2/2+\alpha_2 t+\lambda_2 t^2/2))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j(\alpha_1 t+\lambda_1 t^2/2))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j(\alpha_2 t+\lambda_2 t^2/2))^{\frac{1}{2}}}{\sigma^2}\right] \tag{29}$$

$$S(t_{,}t_{2j}) = \exp\left[\rho\frac{1-(1-2\sigma^2\eta_j(\alpha_1\log(t^{\lambda_1}+1)+\alpha_2\log(t^{\lambda_2}+1)))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j\alpha_1\log(t^{\lambda_1}+1))^{\frac{1}{2}}}{\sigma^2}\right] \exp\left[(1-\rho)\frac{1-(1-2\sigma^2\eta_j\alpha_2\log(t^{\lambda_2}+1))^{\frac{1}{2}}}{\sigma^2}\right] \tag{30}$$

Here onwards we call equation (28), (29) and (30) as Model I, Model II, and Model III respectively and they denote correlated inverse Gaussian frailty model with baseline as Pareto, LFR and Burr distributions respectively.

7. 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,

$$L(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta) = \prod_{j=1}^{n_1} f_1(t_{1j}, t_{2j}) \prod_{j=1}^{n_2} f_2(t_{1j}, c_{2j}) \prod_{j=1}^{n_3} f_3(c_{1j}, t_{2j}) \prod_{j=1}^{n_4} f_4(c_{1j}, c_{2j}) \tag{31}$$

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

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}} \\ f_2(t_{1j}, c_{2j}) &= \frac{\partial S(t_{1j}, c_{2j})}{\partial t_{1j}} \\ f_3(c_{1j}, t_{2j}) &= \frac{\partial S(c_{1j}, t_{2j})}{\partial t_{2j}} \\ and\ f_4(c_{1j}, c_{2j}) &= S(c_{1j}, c_{2j}) \end{aligned} \tag{32}$$

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.

To estimate the 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. Several authors have discussed Bayesian approach for the estimation of parameters of the frailty models. Some of them are, Ibrahim et al.(2001) and references their in, Santos and Achcar (2010). Santos and Achcar (2010) considered parametric models with Weibull and generalized gamma distribution as baseline distributions and gamma, log-normal as frailty distributions. Ibrahim et al. (2001) and references therein considered Weibull model and piecewise exponential model with gamma frailty. They also considered positive stable frailty models.

The joint posterior density function of parameters for given failure times is obtained as,

$$\pi(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \propto L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \times g_1(\alpha_1) g_2(\lambda_1) g_3(\gamma_1) g_4(\alpha_2) g_5(\lambda_2) g_6(\gamma_2) g_7(\theta) \prod_{i=1}^{5} p_i(\boldsymbol{\beta}_i)$$

where $g_i(.)$ $(i = 1, 2, \cdots, 7)$ indicates the prior density function with known hyper parameters of corresponding arguments for baseline parameters and frailty variance; $p_i(.)$ is prior density function for regression coefficient β_i; $\boldsymbol{\beta}_i$ represents a vector of regression coefficients except β_i, $i = 1, 2, \ldots, k$ and likelihood function $L(.)$ is given by equation (31). Here we assume that all the parameters are independently distributed.

To estimate the parameters of the model, we used Metropolis-Hastings algorithm and Gibbs sampler. We monitored the convergence of a Markov chain to a stationary distribution by Geweke test (Geweke 1992)and Gelman-Rubin Statistics (Gelman and Rubin, 1992). Trace plots, coupling from the past plots and sample autocorrelation plots are used to check the behaviour of the chain, to decide burn-in period and autocorrelation lag respectively.

Algorithm consists in successively obtaining a sample from the conditional distribution of each of the parameter given all other parameters of the model. These distributions are known as full conditional distributions. In our case full conditional distributions are not easy to integrate out. So full conditional distributions are obtained by considering that they are proportional to the joint distribution of the parameters of the model.

We have full conditional distribution of the parameter α_1 with frailty as,

$$\pi_1(\alpha_1 \mid \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \propto L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \theta, \boldsymbol{\beta}) \cdot g_1(\alpha_1) \tag{33}$$

We have full conditional distribution of the parameter α_1 without frailty as,

$$\pi_1(\alpha_1 \mid \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \boldsymbol{\beta}) \propto L(\alpha_1, \lambda_1, \gamma_1, \alpha_2, \lambda_2, \gamma_2, \boldsymbol{\beta}) \cdot g_1(\alpha_1) \tag{34}$$

Similarly full conditional distributions for other parameters can be obtained.

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.

8. Analysis of Kidney Infection Data

To illustrate the Bayesian estimation procedure we use kidney infection data of McGilchrist and Aisbett (1991). The data related to recurrence times counted from the moment of the catheter insertion until its removal due to infection for 38 kidney patients using portable dialysis equipment. For each patient, the first and the second recurrence times (in days) of infection from the time of insertion of the catheter until it has to be removed owing to infection is recorded. The catheter may have to be removed for reasons other than kidney infection and this is regarded as censoring. So the survival time for a given patient may be the first or the second infection time or the censoring time. After the occurrence or censoring of the first infection sufficient (ten weeks interval) time was allowed for the infection to be cured before the second time the catheter was inserted. So the first and the second recurrence times are taken to be independent apart from the common frailty component. The data consists of five risk variables age, sex and disease type GN, AN and PKD where GN, AN and PKD are short forms of Glomerulo Neptiritis, Acute Neptiritis and Polycyatic Kidney Disease.

Table 1 in appendix shows the first and second recurrence times with recurrence indicator variable (0-censored, 1-recurrence) and covariates age, sex (0-male, 1-female), and three indicator variables GN, AN, and PKD for six patients only. One can get the entire Table from McGilchrist and Aisbett(1991).

Let T_1 and T_2 be the first and the second recurrence time to infection. Five covariates age, sex and presence or absence of disease type GN, AN and PKD are represented by X_1, X_2, X_3, X_4, and X_5. First we check goodness of fit of the data for the inverse Gaussian frailty distributions with two baseline distributions and then we apply the Bayesian estimation procedure. To check goodness of fit of kidney data set, we consider Kolmogrove-Smirnov (K-S) test for two baseline distributions. Table 2 gives the p-values of goodness of fit test for Model I and Model III. Thus from p-values of K-S test we can say that there is no statistical evidence to reject the hypothesis that data are from the Model I and Model III in the marginal case and we assume that they also fit for bivariate case.

A widely used prior for frailty parameters $\sigma_1 = \sigma_2$ are the gamma distribution $G(0.0001, 0.0001)$. In addition, we assume that the regression coefficients are normal with mean zero and large variance say 1000. Similar types of prior distributions are used in Ibrahim et al. (2001), Sahu et al. (1997) and Santos and Achcar (2010). So in our study we also use same non informative prior for frailty parameters σ_1, σ_2 and regression coefficients $\beta_i, i = 1, .., 5$. Since we do not have any prior information about baseline parameters, $\lambda_1, \alpha_1, \lambda_2$ and α_2, prior distributions are assumed to be flat. We consider two different non-informative prior distributions for baseline parameters, one is $G(a_1, a_2)$ and another is $U(b_1, b_2)$. All the hyper-parameters a_1, a_2, b_1 and b_2 are known. Here $G(a, b)$ is the gamma distribution with the shape parameter a and the scale parameter b and $U(b_1, b_2)$ represents uniform distribution over the interval (b_1, b_2). For correlation parameter we use uniform distribution $U(0, 1)$. We use different values of baseline parameters for Model I, Model II and Model III. We assume the value of the hyper-parameters as $a_1 = 1, a_2 = 0.0001, b_1 = 0$ and $b_2 = 100$.

We run two parallel chains for all four models using two sets of prior distributions with the different starting points using the Metropolis-Hastings algorithm and the Gibbs sampler based on

normal transition kernels. We iterate both the chains for 100000 times. As seen in the simulation study here also we got nearly the same estimates of parameters for both the set of prior, so estimates are not dependent on the different prior distributions. The convergence rate of the Gibbs sampler for both the prior sets is almost the same. Also both the chains shows somewhat similar results, so we present here the analysis for only one chain with $G(a_1, a_2)$ as prior for the baseline parameters, for all the four models.

The Gelman-Rubin convergence statistic values are nearly equal to one and the Geweke test statistic values are quite small and the corresponding p-values are large enough to say that the chains attain stationary distribution. The posterior mean and the standard error with 95% credible intervals, the Gelman-Rubin statistics values and the Geweke test values with p-values for Model I to III are presented in Table 3, 4, and 5. The AIC, BIC and DIC values for all three models are given in Table 6. The Bayes factors for all models are given in Table 7.

In order to compare the proposed models we use the Akaike information criteria (AIC), Bayesian information criteria (BIC) and deviance information criteria (DIC). The comparison between three proposed models is done using AIC, BIC and DIC values given in Table 6. The smallest AIC value is Model-II (linear failure distribution with frailty). Same result hold for BIC and DIC value. To take the decision about Model I, Model II, and Model III, we use the Bayes factor. The Bayesian test based on the Bayes factors for Model II against Model I is 40.4254 and Model II against Model III is 48.6518 which are high and strongly support Model II for kidney infection data set. Some patients are expected to be vary prone to infection compared to others with same covariate value. This is not surprising, as seen in the data set there is a male patient with infection time 8 and 16, and there is also male patient with infection time 152 and 562. Table 6 shows that Model II is better then other two models. From Table 6 and 7, we can observe that, Model II is best. We can observe that the regression coefficients for all the three models are different. The credible interval of the regression coefficient β_2 does not contain zero which indicates that the covariate sex is significant for all the models. But in Model I and Model III β_5 is significant. Negative value of β_2 indicates that the female patients have a slightly lower risk for infection. Negative value of β_5, the regression coefficient corresponding to the covariate X_5 (the disease type PKD) indicates the absence of the disease type PKD in the patients have lover risk of infection in Model I and Model II.

9. Conclusions

In this paper we discuss results for inverse Gaussian correlated frailty models with three different base line distributions. We use the Pareto, LFR and Burr as a baseline distributions. Main aim of our study is to check which distribution with inverse Gaussian correlated frailty fits better. To estimate the parameters in the inverse Gaussian frailty models, we use Bayesian approach.

Different prior gives the same estimates of the parameters. The convergence rate of the Gibbs sampling algorithm does not depend on these choices of the prior distributions in our proposed model for kidney infection data. The estimate of σ from the correlated frailty models show that there is a strong evidence of high degree of heterogeneity in the population of patients. The covariate sex is the only covariate which is significant for all models. Negative value of regression coefficient (β_2) of covariate sex indicates that the female patients have a slightly lower risk of

infection. Negative value of β_5 indicates that the absence of the disease type PKD in the patients have lover risk of infection in Model I and Model II

The comparison between three proposed models is done using AIC, BIC and DIC values. The smallest AIC value is for Model II (linear failure rate distribution with correlated frailty). The same result holds for BIC and DIC values. We observe from Tables 8 and 9 that the Model II is best. Also we can conclude that the correlated inverse Gaussian correlated frailty with the linear failure rate distribution as the baseline distribution is a better fit than other correlated inverse Gaussian correlated frailty models. We compare also with correlated gamma frailty and correlated inverse Gaussian frailty models suggested by Hanagal et al. (2017) and Hanagal and Pandey (2020) and observe that correlated inverse Gaussian frailty with linear failure rate baseline distribution performs better than correlated gamma frailty and correlated inverse Gaussian frailty models proposed by Hanagal et al. (2017) and Hanagal and Pandey (2020) for kidney infection data set. By referring all the above analysis, now we are in a position to say that, we have suggested a new correlated inverse Gaussian frailty model with the linear failure rate distribution as the baseline distribution which is the best in the proposed models for modeling of kidney infection data.

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 Royal Statistical Society*, Ser., A, **148**, 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 and Janssen (2008). *The Frailty Model.* Springer. New York.

Gelman, A. and Rubin, D. B. (1992). A single series from the Gibbs sampler provides a false sense of security. In *Bayesian Statistics 4* (J. M. Bernardo, J. 0.Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: Oxford Univ. Press. pp 625-632.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press, pp 169-193.

Ghitany, M. E. and Kotz, S. (2007). Reliability properties of extended linear failure-rate distributions, *Probability in the Engineering and Information Sciences*, **21**, 441 – 450.

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 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 & 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 & 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 & 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 & 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. (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 & 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 & 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.

Ibrahim, J.G., Ming-Hui C. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer, Verlag. New York.

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.

Lin, C. T., Wu, S. J. S., and Balakrishnan, N. (2006). Monte carlo methods for Bayesian inference on the linear hazard rate distribution, *Communications in Statistics Theory and Methods*, **35**, 575–590.

McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461–466.

Sahu, S.K., Dey, D.K. Aslanidou, H. and Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Life time data analysis*, **3**, 123–137.

Santos, C. A. and Achcar, J. A. (2010). A Bayesian analysis for multivariate survival data in the presence of covariates. *Journal of Statistical Theory and Applications*, **9**, 233–253.

Sen, A. and Bhattacharya, G. K. (1995). Inference procedure for the linear failure rate model, *Journal of Statistical Planning and Inference*, **46**, 59 – 76.

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.

Appendix : Summary of Tables

Table 1: Kidney infection data

Pat	Time1	Ind1	Time2	Ind2	Age	Sex	GN	AN	PKD
1	8	1	16	1	28	0	0	0	0
2	23	1	13	0	48	1	1	0	0
3	22	1	28	1	32	0	0	0	0
4	447	1	318	1	31.5	1	0	0	0
5	30	1	12	1	10	0	0	0	0
6	24	1	245	1	16.5	1	0	0	0

Table 2: p-values of K-S statistics for goodness of fit test for Kidney infection data set

	Recurrence time	
Distribution	first	second
Model I	0.6641	0.7349
Model II	0.7523	0.8386
Model III	0.6256	0.7256

Table 3: Posterior summary for Kidney infection data set Model I

Parameter	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 3150; autocorrelation lag = 300							
α_1	7.2734	0.32204	6.8439	7.8768	0.005102	0.50203	1.0014
λ_1	0.0879	0.0179	0.0094	0.1076	-0.001360	0.49945	1.0089
α_2	2.0142	0.10723	1.6334	2.51025	-0.003313	0.49867	1.0043
λ_2	0.0943	0.02331	0.0096	0.2722	-0.014091	0.49437	0.9999
ρ	0.6594	0.05703	0.5540	0.76740	-0.018870	0.49247	1.0009
σ	0.6009	0.04958	0.5072	0.69349	-0.004090	0.49836	1.0008
β_1	0.0134	0.00802	0.0092	0.02820	-0.003226	0.49871	0.9999
β_2	-1.0468	0.41285	-1.8688	-0.81025	0.005089	0.50203	1.0008
β_3	0.2210	0.22927	0.3472	0.30237	0.006850	0.50273	1.2086
β_4	0.5262	0.03882	0.3940	0.64252	-0.002433	0.49902	1.0013
β_5	-0.7012	0.17155	-0.6208	-0.81960	0.005508	0.50219	0.9999

Table 4: Posterior summary for Kidney infection data set Model II

Parameter	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 3500; autocorrelation lag = 300							
α_1	0.0318	1.59643	0.01178	0.0563	0.00874	0.5034	1.0010
λ_1	0.0316	0.01825	0.0191	0.0495	-0.00126	0.4994	1.0089
α_2	0.0222	0.01799	0.0139	0.0304	-0.00242	0.4990	1.0064
λ_2	0.0093	0.0019	0.0076	0.0229	-0.00082	0.4996	1.0110
ρ	0.8053	0.09019	0.62666	0.9720	0.00595	0.5023	1.0005
σ	0.7014	0.30422	0.16458	1.2912	-0.00405	0.4983	1.0026
β_1	0.0166	0.00960	-0.00129	0.0345	-0.00481	0.4980	1.0041
β_2	-1.9902	0.37287	-2.71766	-1.2694	0.00942	0.5037	1.0002
β_3	0.2212	0.21856	-0.16428	0.2657	-0.01161	0.4953	1.0015
β_4	0.7261	0.03457	-0.08878	-0.0388	-0.02241	0.4910	1.0058
β_5	-0.9701	0.2743	-2.14538	-0.7748	0.00394	0.5015	0.9999

Table 5: Posterior summary for Kidney infection data set Model III

Parameter	Estimate	Standard Error	Lower Credible Limit	Upper Credible Limit	Geweke values	p values	Gelman & Rubin values
burn in period = 3500; autocorrelation lag = 300							
α_1	2.2412	0.19643	2.0217	2.4363	0.00874	0.5034	1.0010
λ_1	0.0071	0.0009	0.0041	0.0098	-0.00126	0.4994	1.0089
α_2	0.0272	0.0199	0.0143	0.0404	-0.00242	0.4990	1.0064
λ_2	0.0113	0.0009	0.0069	0.0242	-0.00082	0.4996	1.0110
ρ	0.7045	0.08419	0.62686	0.8720	0.00595	0.5023	1.0005
σ	0.7324	0.2002	0.56458	0.8912	-0.00405	0.4983	1.0026
β_1	-0.016	0.0082	-0.0712	0.0215	-0.00481	0.4980	1.0041
β_2	-1.6413	0.1737	-1.9766	-1.4694	0.00942	0.5037	1.0002
β_3	0.0867	0.0185	-0.16428	0.2243	-0.01161	0.4953	1.0015
β_4	0.2182	0.03247	-0.0987	0.3368	-0.02241	0.4910	1.0058
β_5	-0.1284	0.0274	-0.2453	0.0848	0.00394	0.5015	0.9999

Table 6: Comparison of AIC, BIC and DIC

Model	AIC	BIC	DIC
Model I	678.3812	781.5814	660.5836
Model II	656.4315	665.3621	641.2143
Model III	793.2195	799.5715	774.8951

Table 7: Bayes Factor for three models

Numerator model against denominator model	Bayes factor	Range	Evidence against model in denominator
Model II against Model I	40.4254	> 10	very strong
Model I against Model III	12.3912	> 10	very strong
Model II against Model III	48.6518	> 10	very strong

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 21–33

Estimation and Spatial Mapping of Incidence of Indebtedness in the State of Karnataka in India by Combining Survey and Census Data

Priyanka Anjoy, Hukum Chandra and Rajender Parsad
ICAR-Indian Agricultural Statistics Research Institute, New Delhi.

Received: 28 December 2019; Revised: 17 January 2020; Accepted: 20 March 2020

Abstract

Information about the household debt behaviour in different occupational categories is of key importance to the Governmental organization for taking effective policy measures targeting the vulnerable groups. This paper illustrates small area estimation (SAE) methodology to estimate proportion of indebted households in rural areas for the two major occupation categories- rural cultivator and rural non-cultivator as well as for both categories combined together across all the 30 districts of Karnataka state in India using the data of All India Debt and Investment Survey 2012-13 and population census 2011. The findings show that the district-level estimates of incidence of indebtedness obtained from SAE are more precise than the direct survey estimates. A spatial map has also been produced to observe the inequality in distribution of indebtedness within districts and in each occupational category across districts. Such maps are definitely useful for framing consistent policy actions and fund disbursement for the indebted household mass.

Key words: Small area estimation; Generalized linear mixed model; Indebtedness; Spatial Map.

1. Introduction

Agriculture plays an important role in the economy of Karnataka and it is the main occupation for more than 60% of population. Karnataka is a drought-prone region with a large proportion of wasteland and having the second largest arid zone in the country after Rajasthan. And due to these factors, the state has been facing agrarian distress with increasing incidence of farmers' suicides since 1997. In fact, the rate of farmer suicides in Karnataka has hit the highest level in a decade, topping the list after Maharashtra, highlighting agrarian distress in the state, according to the report Accidental Deaths and Suicides in India 2015 published by National Crime Records Bureau (NCRB). According to NCRB 2015 data, about 1,197 farmers committed suicides in Karnataka during 2014-15; the state was just behind Maharashtra and Telangana. The NCRB also found that about 79% suicides (946 out of 1,197) in Karnataka were due to bankruptcy or indebtedness. The pre-requisite for any effective policy approach taken in this regard is a proper statistical and economic framework that allows for an effective analysis and monitoring of farmers'

Corresponding Author: Priyanka Anjoy
E-mail: anjoypriyanka90@gmail.com

distress. Measure of disaggregated level indebtedness can be an important tool to the policy makers to mark certain region or group for upliftment and reduce the situation of agrarian distress or farmers' suicides. In this study we attempt to estimate such micro or disaggregated level incidence of indebtedness at micro or local level using the area level small area model.

Most of the large scale surveys are planned to produce reliable estimates at macro or higher geographical (e.g. national and state) level, and cannot be used directly to generate reliable micro or local (also referred to as small area) level estimates because of the small sample sizes (Rao and Molina 2015). This is because, large scale survey designed for a large population (*e.g.* national and state level) may select a small number of units or even no unit from the small area of interest (*e.g.* district or further disaggregation of district). Hence, sample sizes from small areas (or small domains) are too small to justify the use of traditional direct survey estimates. The underlying theory in the literature of survey sampling that helps in resolving the problem of smaller sample sizes is referred as small area estimation (SAE) technique. The technique is model-based methods that links the variable of interest from survey with the auxiliary information available from other data sources for small areas and hence increase the overall (effective) sample size and precision. In this paper we employ area level SAE technique to produce reliable estimates of the incidence of indebtedness among cultivators and non-cultivators categories as well as for both the categories combined in different districts of rural areas of Karnataka in India by linking data from the All-India Debt and Investment Survey (AIDIS) 2012-13 of National Sample Survey Office (NSSO), and the Population Census 2011. This work will enable us to obtain spatial distribution of incidence of indebtedness as well as regional inequality in such measure of indebtedness among the farm families and other families of rural areas in Karnataka. The rest of the paper has been organized into five sections. In Section 2, we discuss the data used in the paper. Section 3 provides an overview of SAE technique that has been used to generate incidence of indebtedness among occupational category by districts in Karnataka. In Section 4, we present diagnostic procedures to examine model assumptions and validate small area estimates including discussion about the results. Finally, Section 5 provides concluding remarks and some recommendations.

2. Data Description

This Section describes about data used in this analysis. In particular, the SAE analysis is based on the AIDIS 2012-13 data for rural areas of the State of Karnataka in India and the Population Census 2011. The sampling design used in the AIDIS 2012-13 data is stratified multi-stage random sampling with districts as strata, the census villages in the rural sector as first stage units and households as the ultimate stage units. For the state of Karnataka, there are a total of 2340 surveyed rural households (including both indebted and non-indebted) spread over 30 districts. The rural households are broadly classified into two types; namely; cultivator and non-cultivator households. As per the concepts and definitions of AIDIS, all rural households operating at least 0.002 hectare of land during the 365 days preceding the date of survey are treated as 'cultivator households'. On the other hand, all rural households operating no land or land less than 0.002 hectare are considered to be non-cultivator households. What follows, based on land holding size (LHS), we denote three categories of households: (i) LHS-A: All households (ii) LHS-C: Cultivator-households with LHS greater than 0.002 ha, and (iii) LHS-NC: Non cultivator-households with LHS less or equal to 0.002 ha. Here, the districts and district by household categories are small areas of interest. Table 1 presents the distribution of district-wise sample sizes

for three categories of households. Across all the districts (*i.e.* LHS-A), the sample size ranges between a minimum of 55 households to a maximum of 112 with an average of 78 households. The sample sizes become too small if sub-grouped further by land holding size categories (i.e. district by cultivator and non-cultivator categories). That is, the sample size of rural cultivators (LHS-C) varies from a minimum of 23 to a maximum of 90 households across the 30 districts with an average of 49 households. And for non-cultivators (LHS-NC), the sample size varies from a minimum of 11 to a maximum of 51 households across the districts with an average of 29 households. Such small samples from the districts pose a challenge in deriving reliable direct estimates of indebtedness. Thus, SAE is an obvious choice to address this problem.

Table 1: Distribution of sample size by occupational categories across districts in rural Karnataka

District	All	Cultivator	Non-Cultivator	District	All	Cultivator	Non-Cultivator
Belgaum	112	67	45	Tumkur	112	90	22
Bagalkot	84	57	27	Kolar	56	45	11
Bijapur	112	85	27	Bangalore	56	23	33
Gulbarga	98	60	38	Bangalore Rural	56	34	22
Bidar	84	49	35	Mandya	112	85	27
Raichur	84	55	29	Hassan	84	63	21
Koppal	84	63	21	Dakshina Kannada	84	41	43
Gadag	56	31	25	Kodagu	56	35	21
Dharwad	56	28	28	Mysore	112	71	41
Uttara Kannada	56	32	24	Chamarajanagar	56	39	17
Haveri	84	52	32	Ramanagara	55	24	31
Bellary	112	72	40	Chikkaballapura	56	42	14
Chitradurga	84	33	51	Yadgir	56	44	12
Davanagere	84	58	26	Minimum	55	23	11
Shimoga	87	50	37	Maximum	112	90	51
Udupi	56	28	28	Average	78	49	29
Chikmagalur	56	27	29	Total	2340	1483	857

Two types of variables are utilized in SAE technique, the variable of interest and the auxiliary variable. As noticed in Section 1, the auxiliary (covariates) variables play an important role in SAE. The auxiliary variables for this analysis are available at district level from the Census 2011. The Population Census 2011 provides a number of covariates at district level that can be utilized for small area modeling. We therefore carried out a preliminary data analysis in order to define appropriate covariates for SAE modeling, using Principal Component Analysis (PCA) to derive composite scores for selected groups of variables. In particular, we carried out PCA separately on three groups of variables, all measured at district level and identified as P1, P2 and P3 below. The first group (P1) consisted of literacy rates by gender and proportions of worker population by gender. The first principal component (P11) for this group explained 61% of the variability, while adding the second principal component (P12) increased explained variability to 85%. The second group (P2) consisted of the proportions of main worker by gender, proportions of main cultivator by gender and proportions of main agricultural labourer by gender. The first principal component (P21) for this second group explained 48% of the variability in the P2 group, while adding the second component (P22) increased explained variability to 62%. Finally, the third group (P3) consisted of proportions of marginal cultivator by gender and proportions of marginal agriculture labourers by gender. The first principal component (P31) for this third group explained

37% of the variability in the P3 group, while adding the second component (P32) increased explained variability to 60%. Finally, three variables, P11, P21 and P31 that significantly explained the model with AIC value 51.59, are identified for the use in SAE analysis. In this paper, the *Y*-variable of interest is the indebted households, i.e. whether a household is in debt or not. A household is defined to be indebted if it has outstanding loan (from respective source) as on 30.06.2012. The target is to estimate the proportion of indebted household (*i.e.* the incidence of indebtedness) at the district (LHS-A) and district by household category (LHS-C and LHS-NC) level. Incidence of indebtedness (IOI) is defined as number of households with any one loan (from respective source) divided by all households in that population segment.

3. Methodological Framework

This Section describes the methodology used in the small area analysis considered in this paper. To begin with, we assume a finite population *U* of size *N* which is consisting of D non-overlapping and mutually exclusive small areas (or district in this paper). We assume that a sample *s* of size *n* is drawn from this population using a probability sampling method. Here, a subscript *d* has been used to denote quantities related to small area d. Let U_d and s_d be the population and sample of sizes N_d and n_d in small 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 i^{th} household is in debt and 0 otherwise) in area d, 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 survey estimator (hereafter denoted by DIR) for P_d is, $p_{dw} = \sum_{i \in s_d} \tilde{w}_{di} y_{di}$ where $\tilde{w}_{di} = w_{di} \Big/ \sum_{i \in s_d} w_{di}$ is the normalized survey weight with $\sum_{i \in s_d} \tilde{w}_{di} = 1$ and w_{di} is the survey weight for unit *i* in area *d*. The estimated design-based variance of DIR is approximated by $v(p_{dw}) \approx \sum_{i \in s_d} \tilde{w}_{di}(\tilde{w}_{di} - 1)(y_{di} - p_{dw})^2$, with the simplifications $w_{di} = a_{di}^{-1}$, $a_{di,di} = a_{di}$ and $a_{di,dj} = a_{di} a_{dj}, i \neq j$, where a_{di} is the first order inclusion probability of unit *i* in area *d* and $a_{di,dj}$ is the second order inclusion probability of units *i* and *j* in area *d*. Under simple random sampling (SRS), $w_{di} = N_d n_d^{-1}$ and DIR is then $p_d = n_d^{-1} y_{sd}$, with estimated variance $v(p_d) \approx n_d^{-1} p_d (1 - p_d)$, 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 DIR may be biased. Furthermore, DIR is based on area-specific sample data and can therefore be very imprecise when the area specific sample size is small or may even be impossible to compute if this sample size is zero. However, model-based SAE procedures that 'borrow strength' via a common statistical model for all the small areas can be used to address this problem. If we ignore the sampling design, the sample count y_{sd} in area (*i.e.* district) *d* can be assumed to follow a Binomial distribution with parameters n_d and π_d, i.e. $y_{sd} \sim \text{Bin}(n_d, \pi_d)$, where π_d is the probability of occurrence of an event for a population unit in area *d* or the probability of prevalence in area *d*.

Similarly, for the non-sample count, $y_{rd} \sim Bin(N_d - n_d, \pi_d)$. Further, y_{sd} and y_{rd} are assumed to be independent binomial variables with π_d being a common success probability.

Let $\mathbf{x}_d$ be the *k*-vector of covariates for area *d* from available data sources. Following Chandra *et al.* (2011) the model linking the probability π_d with the covariates $\mathbf{x}_d$ is the logistic linear mixed model of the 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 $\pi_d = \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)\left\{1 + \exp(\mathbf{x}_d^T \boldsymbol{\beta} + u_d)\right\}^{-1}$. Here $\boldsymbol{\beta}$ is the *k*-vector of regression coefficients, often known as fixed effect parameters, and u_d is the area-specific random effect that captures the area dissimilarities. We assume that u_d's are independently and normally distributed with mean zero and variance σ_u^2. Here, we observe that model (1) relates the area level proportions (direct estimates) from the survey data to the area level covariates. The Fay and Herriot (FH) method for SAE is based on area level linear mixed model and their approach is applicable to a continuous variable. Model (1), a special case of a generalized linear mixed model (GLMM) with logit link function, is suitable for modelling discrete data, particularly the binary variables. (Chandra, 2013; Chandra *et al.*, 2017). Under model (1), an empirical predictor (EP) of the population count y_d in area *d* is

$$\hat{y}_d^{EP} = y_{sd} + \hat{y}_{rd} = y_{sd} + \left(N_d - n_d\right)\left[\exp(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{u}_d)\left(1 + \exp(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{u}_d)\right)^{-1}\right]. \tag{2}$$

An estimate of the corresponding proportion in area *d* is obtained as $\hat{p}_d^{EP} = N_d^{-1} \hat{y}_d^{EP}$. 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 can observe that the parameters $\boldsymbol{\beta}$ and σ_u^2 are the same for every area; i.e., they can be estimated using the data from all small areas. We use an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\beta}$ and $\mathbf{u}$ with REML estimation of σ_u^2 to estimate unknown parameters (Chandra *et al.*, 2011).

The mean squared error (MSE) estimates are computed to assess the reliability of estimates and also to construct the confidence interval (CI). The MSE estimate of (2) is:

$$mse(\hat{p}_d^{EP}) = M_1(\hat{\sigma}_u^2) + M_2(\hat{\sigma}_u^2) + 2M_3(\hat{\sigma}_u^2). \tag{3}$$

Following Chandra *et al.* (2011) we define few notations to express different components of (3). We denote by $\hat{\mathbf{V}}_s = diag\left\{n_d \hat{p}_d^{EP}(1 - \hat{p}_d^{EP})\right\}$ and $\hat{\mathbf{V}}_r = diag\left\{(N_d - n_d)\hat{p}_d^{EP}(1 - \hat{p}_d^{EP})\right\}$ the diagonal matrices defined by the corresponding variances of the sample and non-sample parts, respectively. We then define $\mathbf{A} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_r$, $\mathbf{B} = \left\{diag(N_d^{-1})\right\}\hat{\mathbf{V}}_{rd}\mathbf{X} - \mathbf{A}\hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}$ and $\hat{\mathbf{T}} = \left(\hat{\sigma}_u^2 \mathbf{I}_D + \hat{\mathbf{V}}_s\right)^{-1}$, where $\mathbf{X} = (\mathbf{x}_1^T, ..., \mathbf{x}_D^T)^T$ is a $D \times k$ matrix, and $\mathbf{I}_D$ is an identity matrix of order *D*. We further write

$\hat{\mathbf{T}}_{11} = \left\{\mathbf{X}^T\hat{\mathbf{V}}_s\mathbf{X} - \mathbf{X}^T\hat{\mathbf{V}}_s\hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}\right\}^{-1}$ and $\hat{\mathbf{T}}_{22} = \hat{\mathbf{T}} + \hat{\mathbf{T}}\hat{\mathbf{V}}_s\mathbf{X}\hat{\mathbf{T}}_{11}\mathbf{X}^T\hat{\mathbf{V}}_s^T\hat{\mathbf{T}}$. Under model (1), the components of MSE estimate are: $M_1(\hat{\sigma}_u^2) = \mathbf{A}\hat{\mathbf{T}}\mathbf{A}^T$, $M_2(\hat{\sigma}_u^2) = \mathbf{B}\hat{\mathbf{T}}_{11}\mathbf{B}^T$ and $M_3(\hat{\sigma}_u^2) = trace\left(\hat{\nabla}_i\hat{\mathbf{\Sigma}}\hat{\nabla}'_j v(\hat{\sigma}_u^2)\right)$ with $\hat{\mathbf{\Sigma}} = \hat{\mathbf{V}}_{sd} + \hat{\phi}\mathbf{I}_D\hat{\mathbf{V}}_{sd}\hat{\mathbf{V}}_{sd}^T$. Let us write $\Delta = \mathbf{A}\hat{\mathbf{T}}$ and $\hat{\nabla}_i = \partial(\Delta_i)/\partial\phi\big|_{\phi=\hat{\phi}} = \partial(\boldsymbol{A}_i\hat{\mathbf{T}})/\partial\sigma_u^2\Big|_{\sigma_u^2=\hat{\sigma}_u^2}$, where $\boldsymbol{A}_i$ is the i^{th} row of the matrix $\boldsymbol{A}$. Here $v(\hat{\sigma}_u^2)$ is the asymptotic covariance matrix of the estimate of variance component $\hat{\sigma}_u^2$, which can be evaluated as the inverse of the appropriate Fisher information matrix for $\hat{\sigma}_u^2$. This term also depends upon whether we use ML or REML estimate of $\hat{\sigma}_u^2$. We use REML estimates for $\hat{\sigma}_u^2$ and where $v(\hat{\sigma}_u^2) = 2\left((\hat{\sigma}_u^2)^{-2}(D - 2t_1) + (\hat{\sigma}_u^2)^{-4}t_{11}\right)^{-1}$ with $t_1 = (\hat{\sigma}_u^2)^{-1}trace(\hat{\mathbf{T}}_{22})$ and $t_{11} = trace(\hat{\mathbf{T}}_{22}\hat{\mathbf{T}}_{22})$.

4. Results

The estimation of district level estimates of indebted household for cultivators, non-cultivators and their combined category has been carried out by using direct and model-based methods. In the present study, two types of diagnostics measures are employed: (i) the model diagnostics, and (ii) the diagnostics for the small area estimates. The model diagnostics have been applied to verify model assumptions. The second diagnostics have been applied to validate reliability of the model-based small area estimates.

In model (2), the random area specific effects $u_d (d = 1,..,D)$ have been assumed to have a normal distribution with mean zero and fixed variance σ_u^2. If the model assumptions are satisfied, then the area (or district) level residuals are expected to be randomly distributed and not significantly different from the regression line $y = 0$, where under model (2), the area level residuals are defined as $r_d = \hat{\eta}_d - \mathbf{x}_d^T\hat{\boldsymbol{\beta}}$. The histogram and q-q plot are used to examine the normality assumption. Figure 1 presents the histogram of the district-level residuals, distribution of the district-level residuals and normal q-q plot of the district-level residuals. Besides these graphical methods for checking normality, Shapiro-Wilk (SW) test (*i.e.* test based on uncertainty measurement in terms of p-value) has been performed. The p-value from SW test indicates the chance that the sample comes from a normal distribution. Typically, if p-value is less than 0.05 we can conclude that the sample deviates from normality. Table 2 reports the results of SW test.

Table 2: Shapiro-Wilk (SW) test result for the occupational categories

Occupational category	SW statistic	p-value
All	0.991	0.996
Cultivator	0.986	0.946
Non-Cultivator	0.961	0.332

For assessing validity and reliability of the model-based small area estimates, we applied a set of diagnostics. The values for the model-based small area estimates derived from the fitted model should be consistent with the unbiased direct survey estimates. In other words, these should

provide an approximation to the direct survey estimates that is consistent with these values being "close" to the expected values of the direct estimates. Again the model-based small area estimates should have mean squared errors significantly lower than the variances of the corresponding direct survey estimates. For this purpose, we consider three commonly used diagnostics, *viz*. the bias diagnostics, percentage coefficient of variation (CV %) and 95% confidence intervals for the small area estimates. We compute bias between average value of direct and model estimates (Bias) and average relative difference between direct and model estimates (RE) as:

$$Bias = D^{-1}\left(\sum_{d=1}^{D} Direct\ estimate_d\right) - D^{-1}\left(\sum_{d=1}^{D} Model\ based\ estimate_d\right) \text{ and}$$

$$RE = D^{-1}\left(\sum_{d=1}^{D} \frac{Direct\ estimate_d - Model\ based\ estimate_d}{Direct\ estimate_d}\right).$$

The values of Bias and RE are given in Table 3. The diagnostic results in Table 3 reveal that model-based small area estimates are consistent with the direct survey estimates.

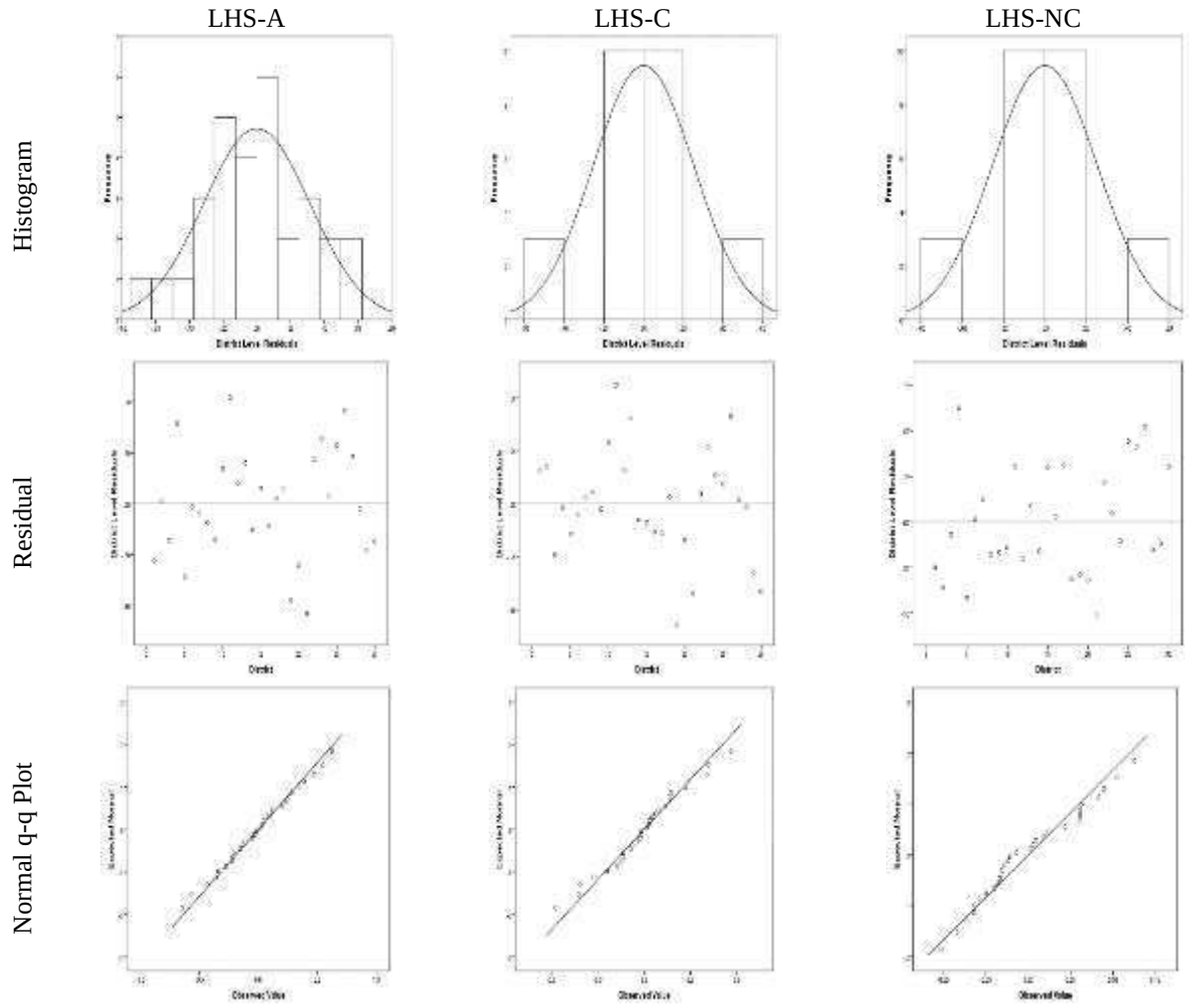
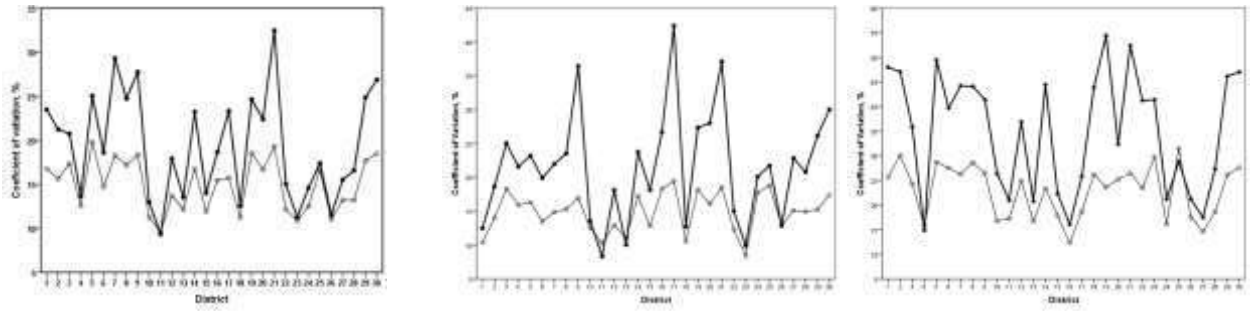


Figure 1: Histogram, distribution and normal q-q plots of the district-level residuals for model based SAE estimates of incidence of indebtedness

Table 3: Bias diagnostics for sample districts

Occupation Category	Bias	RE
All	-0.004	-0.042
Cultivator	-0.006	-0.057
Non-Cultivator	-0.003	-0.131

We compute %CV to assess the improved precision of the model-based estimates compared to the direct survey estimates. Estimates with large CVs are considered unreliable. The average (minimum, maximum) values of CV of direct and model-based (*i.e.* EP) estimates of indebtedness are 19.44% (9.35%, 32.45%) and 14.96% (9.56%, 19.82%), respectively. Similarly, the average (minimum, maximum) values of CV of direct and model-based estimates for cultivators and non-cultivators are 21.61% (8.33%, 42.38%) and 14.88% (8.42%, 19.46%); 35.11% (14.89%, 54.41%) and 22.94% (12.3%, 31.53%), respectively. The district-wise distribution of percentage CV of the model-based estimates and the direct estimates for cultivator and non-cultivator as well as their combined category is shown in Figure 2. These plots show that model-based estimates have a higher degree of reliability as compared to the direct estimates. In general, 95% CIs for the direct estimates are wider than the 95% CIs for the model-based estimates. 95% CIs for the model-based estimates are more precise and contain both direct and model-based estimates of the incidence of indebtedness. The districts-wise estimates of proportion of indebted households along with 95% CIs for the 30 districts of Karnataka are presented in Table 4. The district-wise estimates of proportion of indebted households generated by EP method range between 31.5 to 60.7 % with an average of 46.8%. Similarly, the estimates of proportion of indebted households by occupational categories within districts ranges between 39.7 to 70.2% with an average of 53.3% for cultivators and 24 to 72.9% with average of 39.6% for non-cultivators (Table 4). The maximum proportion of indebted cultivator households (0.70) is reported in Hassan while Udupi (0.73) in case of non-cultivator households. Overall, the maximum incidence of indebtedness (0.61) is found to be in district Haveri. In about 25 out of 30 districts, the incidence of indebtedness is higher among cultivator households as compared to non-cultivator households. The spatial mapping of the incidence of indebtedness among occupational categories (cultivators and non-cultivators) and also for their combined category is shown in Figure 3. Such mapping is useful in microscopic identification of location as well as extent of indebtedness of occupationally differentiated indebted households.



LHC-A LHS-C LHS-NC

Figure 2: District-wise coefficient of variation (%) for the direct (solid line,●) and the model-based SAE estimates (thin line, o) of the incidence of indebtedness

5. Conclusions

The Census in India, like in other countries, usually has limited scope in collection of data. It focuses mainly on basic social and demographic information and that too at decennial interval. On the other hand, NSSO conducts regular surveys on a number of socio-economic indicators, but their utility is restricted to generate national and state level estimates, but not administrative units below state because of small sample sizes for such units. Due to emphasis on disaggregate level Sustainable Development Goal indicators, Government of India as well as different State Governments are now struggling with generation of disaggregate level statistics. The SAE is only indispensable alternative to meet the growing demand for such disaggregated level statistics needed for decentralized policy planning. Using SAE method to link data from the AIDIS 2012-13 and the Population Census 2011, we have derived district level estimates of incidence of indebtedness among cultivators and non-cultivators categories as well as for both the categories combined in different districts of rural areas of Karnataka in India and mapped them to show the spatial variability in incidence of indebtedness at district level. The results might be useful for the program managers and policy planners to implement their policy and interventions effectively.

The use of the diagnostic measure *e.g.* coefficient of variation and the comparison with direct estimates confirm that the model-based district level estimates are robust and provide reliable district level estimates of incidence of indebtedness. The results further confirm that the state level estimates of incidence of indebtedness reported in the AIDIS 2012-13 report mask the district level heterogeneity in rural areas of Karnataka. In particular, this study uncovers the district level incidence of indebtedness in rural areas of Karnataka with their accuracy measures. The region wise picture of indebtedness depicts that southern Karnataka is having higher cases of farm indebtedness, which may be due to more dependence on informal source of credit in this region. Cultivator households need credit on a continuous basis for meeting their working capital needs, hence limited formal source of credit may lead to rising chances of farm indebtedness in this category. It is noteworthy that the AIDIS data used in this study is based on reference year 2012-13 which is almost seven years old. Obvious question arises that the present scenario would be different from what emerges from this study. But, AIDIS is the only regular source to obtain unit level data pertaining to farm indebtedness and the AIDIS-2012-13 is the latest available data for this purpose. Since there is no other recent and updated data available, the estimates generated based on this data is expected to be used as recent information by policy and research analyst and Government departments.

Acknowledgements

The authors express their gratefulness to the reviewer and the editor for their invaluable comments and suggestions, which have helped us to improve the paper substantially.

References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-component model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

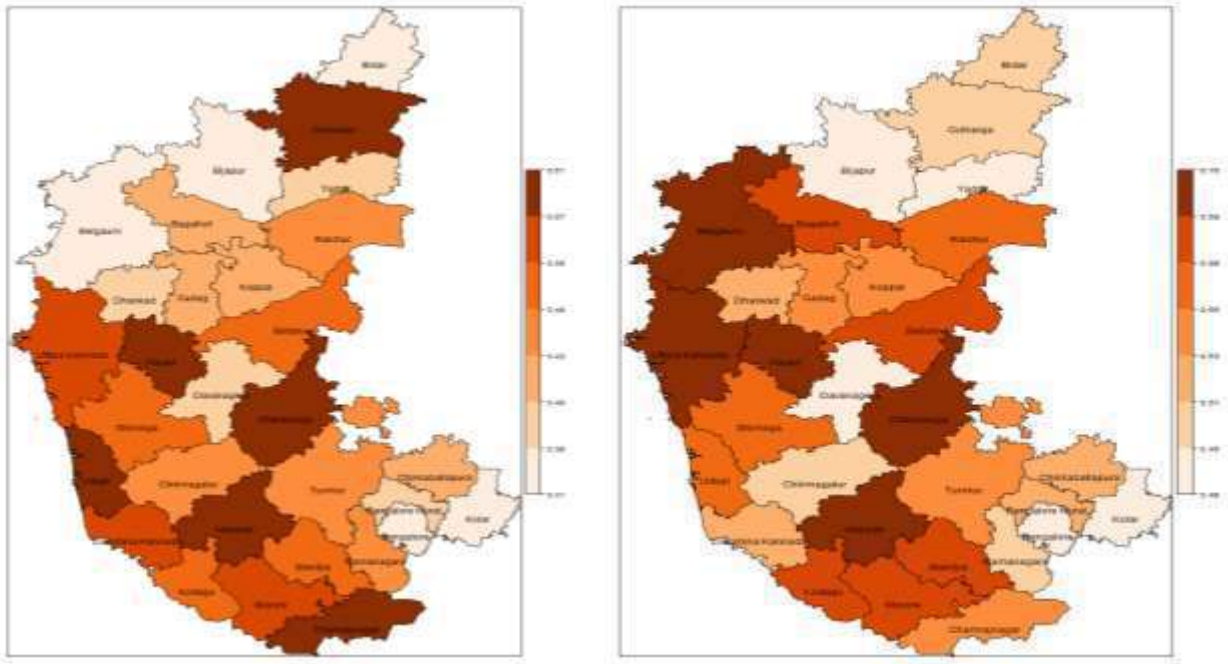
Chandra, H., Salvati, N. and Sud, U. C. (2011). Disaggregate-level estimates of indebtedness in the state of Uttar Pradesh in India – An application of small area estimation Technique. *Journal of Applied Statistics*, **38**, 2413–2432.

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.

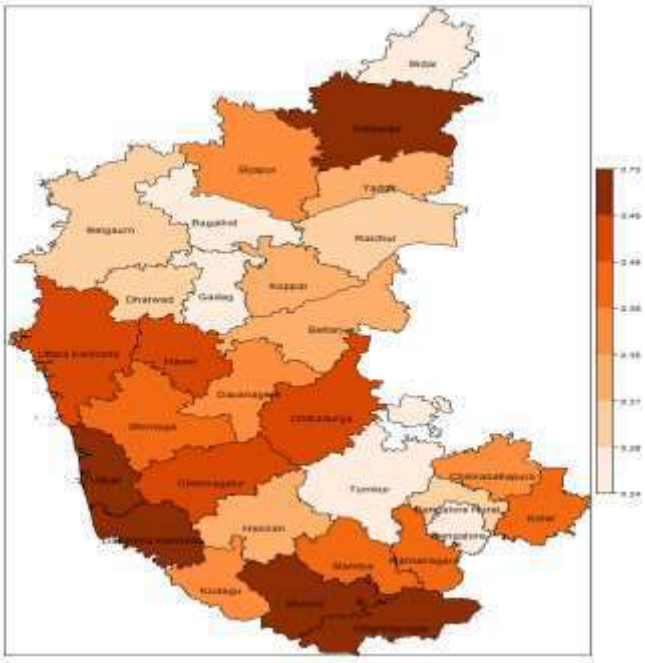
Chandra, H., Salvati, N. and Chambers, R. (2017). Small area prediction of counts under a non-stationary spatial model. *Spatial Statistics*, **20**, 30-56.

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.

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



All Cultivator



Non-Cultivator

Figure 3: Maps of the incidence of indebtedness in rural Karnataka

Table 4: District and occupational category wise estimates of incidence of indebtedness in rural Karnataka along with 95% confidence interval (Lower and Upper) for the direct and model-based small area (EP) method

Category	Districts	Direct			EP		
		Estimate	Lower	Upper	Estimate	Lower	Upper
All	Belgaum	0.34	0.18	0.49	0.38	0.25	0.50
	Bagalkot	0.43	0.25	0.61	0.43	0.30	0.56
	Bijapur	0.34	0.20	0.48	0.38	0.25	0.51
	Gulbarga	0.65	0.48	0.83	0.57	0.43	0.71
	Bidar	0.27	0.14	0.40	0.33	0.20	0.46
	Raichur	0.43	0.27	0.58	0.43	0.31	0.56
	Koppal	0.42	0.18	0.66	0.42	0.27	0.58
	Gadag	0.38	0.20	0.57	0.41	0.27	0.55
	Dharwad	0.36	0.16	0.55	0.40	0.25	0.54
	Uttara Kannada	0.58	0.44	0.73	0.56	0.44	0.68
	Haveri	0.67	0.55	0.80	0.61	0.49	0.72
	Bellary	0.49	0.32	0.66	0.48	0.35	0.61
	Chitradurga	0.61	0.45	0.77	0.57	0.43	0.70
	Davanagere	0.37	0.20	0.53	0.40	0.27	0.54
	Shimoga	0.52	0.38	0.67	0.51	0.39	0.63
	Udupi	0.55	0.35	0.76	0.58	0.40	0.75
	Chikmagalur	0.46	0.25	0.66	0.46	0.32	0.60
	Tumkur	0.48	0.37	0.60	0.48	0.37	0.58
	Kolar	0.26	0.14	0.39	0.35	0.22	0.47
	Bangalore	0.32	0.18	0.46	0.39	0.26	0.52
	Bangalore Rural	0.22	0.08	0.37	0.32	0.20	0.43
	Mandya	0.56	0.40	0.73	0.53	0.41	0.66
	Hassan	0.62	0.48	0.76	0.58	0.46	0.71
	Dakshina Kannada	0.55	0.39	0.70	0.55	0.41	0.68
	Kodagu	0.61	0.41	0.82	0.54	0.37	0.72
	Mysore	0.63	0.49	0.77	0.56	0.44	0.68
	Chamarajanagar	0.64	0.44	0.83	0.57	0.42	0.71
	Ramanagara	0.46	0.31	0.60	0.46	0.34	0.58
	Chikkaballapura	0.35	0.18	0.53	0.42	0.27	0.56
	Yadgir	0.34	0.16	0.52	0.38	0.24	0.52
Cultivator	Belgaum	0.62	0.46	0.77	0.60	0.48	0.72
	Bagalkot	0.61	0.39	0.84	0.56	0.41	0.71
	Bijapur	0.35	0.18	0.52	0.42	0.27	0.58
	Gulbarga	0.50	0.29	0.71	0.49	0.33	0.64
	Bidar	0.41	0.23	0.60	0.46	0.32	0.61
	Raichur	0.53	0.33	0.74	0.55	0.40	0.69
	Koppal	0.55	0.31	0.79	0.54	0.39	0.70
	Gadag	0.54	0.29	0.80	0.53	0.37	0.69
	Dharwad	0.45	0.13	0.78	0.52	0.35	0.69
	Uttara Kannada	0.69	0.51	0.88	0.60	0.45	0.74
	Haveri	0.79	0.66	0.92	0.65	0.52	0.79
	Bellary	0.64	0.41	0.86	0.58	0.43	0.73
	Chitradurga	0.83	0.67	0.99	0.69	0.54	0.84
	Davanagere	0.42	0.22	0.61	0.44	0.29	0.59

	Shimoga	0.54	0.35	0.73	0.55	0.41	0.69
	Udupi	0.49	0.23	0.75	0.55	0.36	0.75
	Chikmagalur	0.38	0.06	0.69	0.46	0.28	0.63
	Tumkur	0.55	0.41	0.69	0.55	0.44	0.66
	Kolar	0.26	0.12	0.40	0.40	0.26	0.54
	Bangalore	0.44	0.20	0.69	0.51	0.35	0.68
	Bangalore Rural	0.26	0.07	0.45	0.41	0.26	0.56
	Mandya	0.59	0.42	0.76	0.57	0.43	0.70
	Hassan	0.73	0.59	0.88	0.70	0.59	0.82
	Dakshina Kannada	0.57	0.34	0.79	0.52	0.34	0.71
	Kodagu	0.62	0.36	0.89	0.56	0.36	0.77
	Mysore	0.66	0.49	0.82	0.56	0.42	0.70
	Chamarajanagar	0.55	0.30	0.80	0.55	0.39	0.71
	Ramanagara	0.51	0.30	0.71	0.51	0.36	0.65
	Chikkaballapura	0.41	0.20	0.62	0.51	0.36	0.67
	Yadgir	0.32	0.13	0.51	0.45	0.30	0.61
Non-Cultivator	Belgaum	0.22	0.01	0.42	0.29	0.14	0.43
	Bagalkot	0.14	0.01	0.27	0.24	0.10	0.38
	Bijapur	0.31	0.09	0.54	0.35	0.19	0.52
	Gulbarga	0.77	0.55	1.00	0.58	0.40	0.76
	Bidar	0.13	0.00	0.25	0.28	0.12	0.43
	Raichur	0.27	0.06	0.48	0.29	0.13	0.45
	Koppal	0.36	0.05	0.67	0.34	0.16	0.51
	Gadag	0.22	0.03	0.40	0.28	0.12	0.44
	Dharwad	0.27	0.05	0.48	0.30	0.15	0.46
	Uttara Kannada	0.45	0.22	0.68	0.49	0.33	0.65
	Haveri	0.52	0.30	0.73	0.46	0.31	0.62
	Bellary	0.28	0.08	0.47	0.32	0.16	0.47
	Chitradurga	0.50	0.29	0.70	0.49	0.33	0.64
	Davanagere	0.31	0.04	0.58	0.38	0.20	0.55
	Shimoga	0.51	0.28	0.73	0.46	0.30	0.62
	Udupi	0.71	0.49	0.94	0.73	0.55	0.91
	Chikmagalur	0.53	0.26	0.79	0.47	0.30	0.64
	Tumkur	0.21	0.03	0.39	0.28	0.14	0.43
	Kolar	0.30	-0.02	0.62	0.39	0.21	0.57
	Bangalore	0.24	0.09	0.39	0.30	0.15	0.45
	Bangalore Rural	0.13	0.00	0.26	0.28	0.14	0.43
	Mandya	0.48	0.09	0.86	0.41	0.22	0.59
	Hassan	0.33	0.06	0.60	0.32	0.13	0.50
	Dakshina Kannada	0.52	0.31	0.74	0.56	0.38	0.73
	Kodagu	0.60	0.26	0.93	0.37	0.14	0.60
	Mysore	0.59	0.35	0.84	0.50	0.33	0.67
	Chamarajanagar	0.78	0.51	1.05	0.62	0.44	0.80
	Ramanagara	0.40	0.19	0.61	0.42	0.27	0.57
	Chikkaballapura	0.29	0.03	0.55	0.35	0.17	0.54
	Yadgir	0.46	0.04	0.88	0.34	0.16	0.53

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 35-53

Application of Parametric Survival Model and Multinomial-Dirichlet Bayesian Model within a Multi-state Setup for Cost-Effectiveness Analysis of Two Alternative Chemotherapies for Patients with Chronic Lymphocytic Leukaemia

Gurprit Grover[1] and **Vishal Deo**[1, 2]

[1]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi*
[2]*Department of Statistics, Ramjas College, University of Delhi*

Received: 19 September 2019; Revised: 21 March 2020; Accepted: 25 March 2020

Abstract

Estimation of transition probabilities between disease states and determination of length of stay in each state are two major concerns underlying the multi-state model based approach for cost-effectiveness analysis (CEA). The objective of this study is to apply and compare two different methods to estimate transition probabilities between three disease states *viz.*, progression-free, progression, and death, for performing CEA of chemoimmunotherapy, as compared to chemotherapy, for treating patients suffering from Lymphocytic Leukaemia. In the first method, we fit a parametric survival model to the events progression and death, and as an alternative approach, we fit a Multinomial- Dirichlet Bayesian model to the number of transitions between different states. In the first method a Weibull clock-forward time-inhomogeneous semi-markov model is used, while in the second method the transition probabilities are assumed to be time-independent and are estimated through simulations from their posterior distributions using MCMC implementation.

Results from both methods suggest that chemoimmunotherapy is cost-effective over chemotherapy. However, a comparison between the predictions of long term transitions from the two methods suggests that the method based on Weibull time-inhomogeneous semi-markov model provides more reliable estimates, especially when the time horizon of the study is long. Chemoimmunotherapy is cost effective when patients are willing to pay an additional cut off cost of around 13,000-15,000 GBP (by first method) for per unit additional gain in QALY.

Key words: ICER; Multi-state model; IPD reconstruction; Rituximab; Willingness to pay; Health economics; Total length of stay.

1. Introduction

Scientific comparison of alternative treatments for a disease, both in terms of desired outcome and costs, is imperative for optimal decision making in medical sciences. Cost-effectiveness analysis (CEA) is an important aspect of Health Economics and deals with the evaluation of cost per outcome gained. Outcome is usually defined in terms of survivability and hence, cost-effectiveness analysis evaluates both survival data and costs data simultaneously.

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

Markov, semi-Markov and non-Markov multi-state models provide a comprehensive approach towards CEA of interventions for diseases for which discrete progression states can be defined based on certain clinical and pathological markers. Briggs and Sculpher (1998) provided a comprehensive structure of markov modelling for health economic evaluation and also discussed the importance of scrutinizing the Markovian assumption before estimating the transition probabilities. In a significant number of work based on Markov decision-analytic models in health economics, authors have preferred to obtain transition probabilities from published literature, refer Gharaibeh *et al*. (2015), Veldhuijzen *et al.* (2010), Lee *et al.* (2013), and Yeh *et al.* (2010). However, as the time horizon of the study from which published estimates are obtained is generally different from the time horizon of the study being conducted, these estimates are unlikely to act as reliable estimates of the true transition probabilities. This is due to the fact that, in most cases of disease progressions, transition probabilities are expected to be time dependent owing to the impact of changes in various covariates.

In the presence of individual patient data (IPD), parametric and semi-parametric survival models can be fitted to the survival data to estimate the transition probabilities of multi-state models. Exponential, Gamma, Weibull, Lognormal, and Generalized Gamma survival models are popularly found to be appropriate parametric options for the purpose of estimating transition probabilities, see for example Wu *et al.* (2014), Speight *et al.* (2006), Coon *et al.* (2010), and Diaby *et al.* (2013). Use of Cox proportional hazards models has been suggested by some authors, like Malehi *et al.* (2015) and Mihaylova *et al.* (2011) among others, especially when our interest also lies in estimating hazards associated with the covariates, given the validity of the proportional hazards assumption. Flexible semi-parametric survival models, like partitioned Cox models, can be adopted to allow for flexibility in case of violation of the proportional hazards assumption, refer Jackson *et al.* (2010) and Williams *et al.* (2017 b). Application of Bayesian parametric models to estimate the transition probabilities of multi state models has also been discussed to some extent in the literature of Health Economics; see for example Welton and Ades (2005) and Baio (2013).

In the absence of IPD, survival data can be reconstructed from published Kaplan-Meier (KM) curves by incorporating the published information about risk sets at different time points of the study. The method is discussed in detail by Hoyle and Henley (2011), Guyot *et al.* (2012), and Wan *et al.* (2015).

This paper aims to apply and compare two different methods to estimate transition probabilities between the three states of chronic lymphocytic leukaemia *viz.*, progression-free (PF), progression, and death, and carry out CEA of two types of chemotherapies used for treating patients. We have reconstructed survival data of two groups of patients suffering from chronic lymphocytic leukaemia; one group was treated with the combination of fludarabine and cyclophosphamide (chemotherapy group), and the other group was treated with the combination of fludarabine, cyclophosphamide, and rituximab (chemoimmunotherapy group). Transition probabilities between different states are estimated by fitting a a) time-inhomogeneous Weibull semi-markov model and b) Multinomial-Dirichlet Bayesian model for number of transitions. Quality-adjusted life years (QALYs) and incremental cost-effectiveness ratios (ICERs) are calculated to compare the economic and survival utility of the two interventions.

2. Survival Data: Reconstruction of IPD

KM curves for overall survival (OS) and progression-free survival (PFS) for both chemotherapy and chemoimmunotherapy groups are obtained from the published work of Hallek *et al.* (2010) based on a randomized phase 3 trial of patients with chronic lymphocytic leukaemia. PFS is defined as the time between randomization and the date of the first documented disease progression, relapse, or death by any cause, and OS is defined as the time between randomization and the date of death from any cause; refer Roche (2008). Summary of the actual (published) data is shown in Table 1.

Enguage Digitizer software is used to extract coordinates from the four KM curves. The algorithm (R code) for reconstructing IPD from the extracted coordinates of KM curves developed by Guyot *et al.* (2012) is applied to reconstruct the survival data with right censoring for both treatment groups. KM curves based on actual data from Hallek *et al.* (2010) and those based on reconstructed data (with 95% confidence bounds) are shown in Figures 1 and 2 respectively. The *x*-axis in these curves represents time since the start of the study in months.

3. Methodology and Results

A three state multi-state model is conceived with possible transitions between states as described in Figure 3. An overview of the methodological structure of this study is outlined in Figure 4. A lifetime time horizon of 15 years is taken for base cost-effectiveness analyses as only 1.3% of the cohort are estimated to be surviving beyond this period as reported by Roche (2008). However, QALYs and ICERs have also been calculated for a time horizon of 20 years to evaluate the effect of choice of time horizon on QALYs and ICERs. This is necessary to account for the uncertainty underlying the choice of lifetime time horizon, see Jackson *et al.* (2017).

3.1. Estimation of transition probabilities of the multi-state model using parametric survival model

To start with, Gamma, Exponential, Weibull, Log-logistic, Log-normal, and Generalized Gamma survival models are fitted to the transitions data of both groups. Based on the Akaike Information Criterion (AIC), and size of confidence intervals, survival models with Weibull distribution for time to events (progression and death) are found to exhibit best fits for both chemotherapy and chemoimmunotherapy groups. AIC values of the fitted models are provided in Tables 2 and 3. Plots of estimated survival functions, along with corresponding confidence intervals and KM estimates, are presented in Appendix-A, Figures A.1 and A.2. A combination of R functions available in the packages ‘flexsurv’ by Jackson (2016) and ‘mstate’ by de Wreede *et al.* (2010) are used to obtain these results. The chosen Weibull model is a clock-forward time-inhomogeneous semi-markov model. This suggests that the transition probabilities are assumed to vary with time. As an illustration, estimated transition probability matrices at the times $t = 50$ months, $t = 100$ months, $t = 180$ months and $t = 240$ months from the start of the study ($t = 0$) are presented in Table 4 (for chemotherapy arm) and Table 5 (for chemoimmunotherapy arm). These transition probabilities, say, P_{rs}^{T}, $r = 1,2,3$, $s = 1,2,3$, and $T = 1,2,3\ldots$, represent the probability that a patient is in state s at time $t = T$, given that he/she was in state r at time $t = 0$. These transition probabilities are used to calculate expected total length of stay (TLOS) in each state s, when a patient starts from a

particular state r at time $t = 0$. TLOS matrices calculated for both groups, at 15 years' and 20 years' time horizon, are provided in Table 6.

3.2. Costs data

In both treatment regimes, some of the costs are fixed, while some of them are variable and depend on the length of the treatment. Cost of supportive care in PF state, cost of supportive care in progression, and cost of second-line and subsequent therapy are dependent on the duration of treatment/ care in the respective states and so, total expected costs under these heads have been calculated using estimated TLOS in the respective states. Monthly mean costs for these heads are obtained from Roche (2008). Expected total costs for the variable heads are calculated at discount rates of 3.5% and 5% per annum over the lengths of stay. Estimated mean total costs of both treatment regimes, and the expected cost incremental for chemoimmunotherapy with respect to chemotherapy, are shown in Table 7 (15 years time horizon) and Table 8 (20 years time horizon). Discount rate of 3.5% has been advised by the National Institute of Health and Care Excellence (NICE), UK, and discount rate of 5% is taken to analyze the sensitivity of the results towards the choice of discount rates. Remaining mean costs, which are not related to the length of stay in any state and are essentially fixed costs, are taken from Williams *et al.* (2017 a). Since a patient is not expected to go back to the PF state after entering progression state, it is safe to use the generic formula given in equation (1) to calculate discounted costs.

$$PV = \sum_i \frac{V_i}{(1+d)^i} \tag{1}$$

Here, d is per unit time discount rate, PV is present value of the total cost and V_i is actual cost incurred at i^{th} time point (with base period at $i = 0$).

3.3. QALY and ICER

For calculation of mean QALYs, utility values of 0.8 and 0.6 have been considered for the PF health state and the progression health state respectively; refer Roche (2008). QALYs are discounted at 3.5% and 5% rates, in concurrence with the rates of discount for costs, and using the formula given in equation (1) after replacing costs with lengths of stay. ICER, which represents the cost per unit increase in QALY, is calculated for each time horizon at both discount rates. Calculated values of discounted mean QALYs for both treatment groups, QALY incremental, cost incremental and ICERs are reported in Table 9. QALY incremental, cost incremental, and ICERs have been calculated taking chemotherapy as the base intervention. Figure 5 exhibits the cost-effectiveness plane, showing acceptability of the chemoimmunotherapy over chemotherapy at two different values of willingness to pay, *viz*. K = 15,000 GBP and K = 13,000 GBP, for a unit additional gain in QALY.

3.4. Estimation of transition probabilities of the multi-state model using Multinomial-Dirichlet Bayesian model

In this method, instead of getting into the realm of survival models, we define the observed number of transitions between states as a vector of random variables following Multinomial distribution and estimate the parameters under Bayesian framework using Dirichlet as the prior distribution; refer Baio (2013) and Welton and Ades (2005). Dirichlet

distribution is a conjugate prior for Multinomial distribution. The Multinomial- Dirichlet Bayesian model for our multi-state set up is defined as follows:

Notations:

$r_{st}^{(I)}$: Total number of observed transitions from state *s* to state *t* for intervention *I.*

$\lambda_{st}^{(I)}$: Transition probability from state *s* to state *t* for intervention *I.*

$n_s^{(I)}$: Total number of transitions from state *s* to all other states for intervention *I.*

$r_s^{(I)} = (r_{s1}^{(I)}, r_{s2}^{(I)}, r_{s3}^{(I)})$

$\lambda_s^{(I)} = (\lambda_{s1}^{(I)}, \lambda_{s2}^{(I)}, \lambda_{s3}^{(I)})$

Here, $s = 1, 2, 3$ and $I = 1, 2$.

$$r_s^{(I)} | \lambda_s^{(I)} \sim Multinomial(\lambda_s^{(I)}, n_s^{(I)})$$

$$= \frac{n_s^{(I)}}{r_{s1}^{(I)}! r_{s2}^{(I)}! r_{s3}^{(I)}!} \lambda_{s1}^{(I)^{r_{s1}^{(I)}}} \lambda_{s2}^{(I)^{r_{s2}^{(I)}}} \lambda_{s3}^{(I)^{r_{s3}^{(I)}}} \quad (2)$$

And prior distribution of the transition probabilities is defined as,

$$\lambda_s^{(I)} | \alpha^{(I)} \sim Dirichlet(\alpha_1^{(I)}, \alpha_2^{(I)}, \alpha_3^{(I)})$$

$$= \frac{\Gamma(\alpha_1^{(I)} + \alpha_2^{(I)} + \alpha_3^{(I)})}{\Gamma(\alpha_1^{(I)})\Gamma(\alpha_2^{(I)})\Gamma(\alpha_3^{(I)})} \lambda_{s1}^{(I)(\alpha_1^{(I)} - 1)} \lambda_{s2}^{(I)(\alpha_2^{(I)} - 1)} \lambda_{s3}^{(I)(\alpha_3^{(I)} - 1)} \quad (3)$$

Unknown parameter of the Multinomial distribution in (2) is nothing but the vector of transition probabilities from state *s* to all other states, whose prior distribution is defined by the Dirichlet distribution with density function given in (3). It should be noted that while specifying this model, the transition probabilities are assumed to be constant, *i.e.* independent of time, unlike in the case of previous method based on Weibull clock-forward semi-markov model. Markov Chain Monte Carlo (MCMC) method is implemented through JAGS (Just Another Gibbs Sampler) within R session for simulating posterior distributions of the vectors of transition probabilities. R and JAGS codes are adopted from Baio (2013), and implemented with necessary modifications, corrections and additions to estimate transition probabilities from their posterior realizations. Beyond this point, two approaches are implemented to carry out CEA.

In the first approach, CEA is performed using the function bcea( ) available in the R package 'BCEA'. In this method, ICERs are determined from vectors of cost incremental values and benefit incremental values calculated at all simulated values of the transition probabilities. Here, benefit is defined as the total number of time units (months) spent by patients in the first state, *i.e.* PF state; which is mathematically equal to the total number of patients in PF state summed over the entire time horizon. Calculations of costs and benefits are done at discount rates of 3.5% and 5% each. ICERs and Cost Effectiveness Acceptability Curves (CEACs) are obtained for time horizons of 15 and 20 years. Results at discount rates of 3.5% and 5% are presented in the Figures 6 and 7, respectively.

The second approach has been adopted to gain better insight into the comparative assessment between the Weibull semi-markov method and the Multinomial-Dirichlet

Bayesian method. Expected TLOS in each state and respective 95% confidence intervals are calculated on the basis of transition probabilities simulated from their posterior distribution. Further, at utility values of 0.8 for PF state and 0.6 for progression state, QALYs are obtained using TLOS matrices at 3.5% and 5% discount rates. Estimated time-independent probability transition matrices (with 95% CIs), and expected TLOS matrices, for both treatment groups, are provided in Tables 10 and 11 respectively.

4. Summary and Discussions

4.1. CEA based on Weibull semi-markov model

The estimated transition probability matrices based on Weibull semi-markov models for chemotherapy group and chemoimmunotherapy group exhibit notable differences in the probability of a patient in PF state to remain in the same state after a time interval $t\ (>0)$. A resultant major impact of this finding from the CEA point of view is the significant difference between the expected TLOS in the first state, PF, of the two intervention groups. PF state has maximum utility value among the three states and contributes the most to gain in QALYs. From Table 6, for time horizons of both 15 years and 20 years, the expected TLOS in the PF state for the chemoimmunotherapy group is around 52 months as opposed to that of just 35 months for the chemotherapy group; a difference of around 17 months. However, difference in the expected TLOS in the progression state between the two intervention groups is least prominent.

QALY results in Table 9 show additional/ incremental gains of 0.84 (d = 3.5%) and 0.94 (d = 5 %) QALYs for 15 years' time horizon, and of 0.79 (d = 3.5%) and 0.90 (d = 5%) QALYs for 20 years' time horizon, for the chemoimmunotherapy group over the chemotherapy group. At both discount rates, QALY incremental is lower and cost incremental is higher in case of 20 years' time horizon, as compared to those for 15 years' time horizon. As a result, the ICERs for 20 years' time horizon are on the higher side as compared to those for 15 years' time horizon. For a fixed time horizon, ICERs corresponding to the discount rate of 3.5% are significantly higher than those corresponding to the discount rate of 5%. It can also be noted that the ICERs are more sensitive towards the choice of discount rates (keeping time horizon fixed), than towards the choice of time horizon (keeping discount rate fixed). ICERs corresponding to 5% discount rate are below the willingness to pay line of K = 13,000 GBP, while both ICERs corresponding to 3.5% discount rate are above that line. At the willingness to pay of 15,000 GBP or more for a unit increase in QALY, choosing chemoimmunotherapy over chemotherapy accounts for an optimal decision as all the four points A,B,C and D lie below the line corresponding to K = 15,000 GBP (Figure 5).

4.2. CEA based on Multinomial- Dirichlet Bayesian model

First approach: Optimal decisions derived from this approach find chemoimmunotherapy to be cost-effective over chemotherapy for willingness to pay parameter more than or equal to around 227,000 GBP, for both time horizons, and 3.5% discount rate (Figure 6). At 5% discount rate, chemoimmunotherapy is cost-effective over chemotherapy if a patient is willing to pay around 306,000 GBP for an additional gain of QALY, for both 15 years' and 20 years' time horizons (Figure 7). In this case also, ICERs are found to be significantly sensitive towards the choice of discount rates, but not towards the

choice of time horizon. These values of ICERs are strikingly and absurdly higher than those obtained from the first method.

Second approach: As is apparent from the results reported in Table 11, the expected TLOS in the PF state are much higher and in progression state are unreliably low, raising speculations of wrong predictions. However, because of the drastic underestimation of TLOS in progression state and overestimation in PF state, the ICER comes out to be balanced. At 3.5% discount rate and for 15 years time horizon, while the cost incremental is approximately 11,560 GBP, QALY incremental is only 0.66, rendering ICER to around 17,515 GBP for one unit additional gain in QALY.

Another notable difference in these two approaches is that utility values have not been used for calculation of benefits in the first approach. Also, in the first approach, using the function bcea( ), the benefit incremental is calculated as difference in the total number of months spent by patients in PF state for the two intervention groups; which is equivalent to expected TLOS incremental in months in PF state. Thus, the first approach completely ignores the gain in utility because of stay in progression state. Extremely high ICER from the first approach suggests that the difference in expected number months spent by patients in PF state is very small and does not capture the actual difference in gain in QALY between the two groups.

4.3. Comparing predictions from the two models

As the study involves long lifetime time horizons, accuracy of prediction of transitions is of utmost importance for conducting CEA. Remarkable differences in the results of QALYs and ICERs obtained from the two methods suggest that at least one of them may not be reliable. To compare and examine the predictions from the two methods, graphs of estimated proportion of patients in each state at different time points (virtual follow-up times) are plotted for both intervention groups (Figures 8 and 9). From the graphs in Figure 8, pertaining to first method, we can see that only few patients are expected to remain in PF state till around 150 months in chemotherapy group, and till 180 months in chemoimmunotherapy group. While based on the transition probabilities of the second method, it is apparent from the graphs in Figure 9 that more than 25% of patients are expected to live even after 180 months in both intervention groups. This is in clear contrast with the observed survival data and the contradiction can be visualized easily on comparing the original KM curves in Figure 1with the graphs in Figure 9. However, we can safely claim that the shapes of the graphs in Figure 8 conform to those of the original KM curves till the observed time period of 5 years (or 60 months). According to the KM curves, at the end of 60 months, around 25% patients in chemotherapy group and 40% patients in chemoimmunotherapy group were free of progression (*i.e.* in PF state), while around 60% patients in chemotherapy group and more than 50% patients in chemoimmunotherapy group were still alive. Around same proportions are depicted by the patient proportion graphs based on Weibull time-inhomogeneous semi-markov model.

Failure of the Multinomial-Dirichlet Bayesian model in this study can be attributed to the assumption of time-homogeneous (or constant) transition probabilities over the entire time horizon of the study. In a long-term study, transition probabilities from a state to other states are expected to change with time, especially when patients stay in the state for a longer duration. So, it is safe to conclude that this method, or any other method with the assumption of constant transition probabilities, should be avoided for CEA in long-term studies.

For further comparison, plots of estimated transition probabilities against time for the two methods are provided in Appendix-B (Figures B.1 and B.2).

5. Conclusion

We have applied two different methods for analyzing cost-effectiveness of chemoimmunotherapy over chemotherapy for treating patients with chronic lymphocytic leukaemia. Although results from both methods find chemoimmunotherapy to be cost effective over chemotherapy, values of QALYs and ICERs from the Weibull time-inhomogeneous semi-markov model are found to be more reliable. To be precise, chemoimmunotherapy is cost effective over chemotherapy if the patients are willing to pay around 15,000 GBP or more for a unit additional gain in QALY. Also, it can be inferred from the results that administration of chemoimmunotherapy in place of chemotherapy is expected to result in a patient to stay for a much longer period (over a year on an average) in the PF state, which is the state of highest utility.

For one-way sensitivity analysis of cost-effectiveness towards the choice of lifetime time horizon and the choice of discount rate, analyses are carried out for two different lifetime time horizons and at two different discount rates for cost and QALY calculations. Results from both models confirm that ICERs are more sensitive towards the choice of discount rate than the choice of lifetime time horizon. This suggests that discount rate should be chosen carefully after consulting relevant economic parameters of the region of study to avoid biased and misleading results.

Since a reconstructed data has been used for the analyses, the data consists of only survival times for the events progression and death, and no information is available on covariates and factors affecting survivability. Inclusion of data on covariates, like pathological and clinical factors, demographic variables *etc.*, will make such cost-effectiveness studies more comprehensive and informative.

Acknowledgements

We are extremely grateful to the reviewers and the editors for their invaluable comments and suggestions, which have helped us to improve the paper substantially.

References

Baio, G. (2013). *Bayesian Methods in Health Economics.* Chapman & Hall / CRC Biostatistics Series.

Briggs, A. H., Claxton, K. and Sculpher, M. J.(2006). *Decision Modelling for Health Economic Evaluation.* Oxford University Press.

Briggs, A. H., and Sculpher, M. J. (1998). An introduction to Markov modelling for economic evaluation. *PharmacoEconomics*, **13**(**4**), 397-409.

Coon, J. T., Hoyle, M., Green, C., Liu, Z., Welch, K. *et al.* (2010). Bevacizumab, sorafenib tosylate, sunitinib and temsirolimus for renal cell carcinoma: a systematic review and economic evaluation. *Health Technology Assessment*, **14**(**2**), doi: 10.3310/hta14020.

de Wreede, L. C., Fiocco, M. and Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, **99**, 261-274.

Diaby, V., Adunlin, G. and Montero, A. J. (2013). Survival modeling for the estimation of transition probabilities in model-based economic evaluations in the absence of individual patient data: a tutorial. *PharmacoEconomics*, **32**(**2**), 101-108.

Gharaibeh, M., McBride, A., Bootman, J. L. and Abraham, I. (2015). Economic evaluation for the UK of nab-paclitaxel plus gemcitabine in the treatment of metastatic pancreas cancer. *British Journal of Cancer*, **112**, 1301-1305.

Guyot, P., Ades, A. E., Ouwens, M. J. and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, **12(9)**, doi:10.1186/1471-2288-12-9.

Hallek, M., Fischer, K., Fingerle-Rowson, G., Fink, A. M., Busch, R. *et al.* (2010). Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet*, **376,** 1164-1174.

Hoyle, M. W. and Henley, W. (2011). Improved curve fits to summary survival data: application to economic evaluation of health technologies. *BMC Medical Research Methodology*, **11(139)**, doi:10.1186 /1471-2288-11-139.

Jackson, C. H. (2016). flexsurv: A platform for parametric survival modelling in R. *Journal of Staistical Software*, **70(8),** doi: 10.18637/jss.v070.i08.

Jackson, C. H., Sharples, L. D. and Thompson, S. G. (2010). Survival models in health economic evaluations: balancing fit and parsimony to improve prediction. *The International Journal of Biostatistics*, **6(1)**, Article 34, doi: 10.2202/1557-4679.1269.

Jackson, C. H., Stevens, J., Ren, S., Latimer, N., Bojke, L. *et al.* (2017). Extrapolating survival from randomized trials using external data: a review of methods. *Medical Decision Making*, **37**, 377-390.

Lee, Y. Y., Veerman, J. L. and Barendregt, J. J. (2013). The cost-effectiveness of Laparoscopic adjustable gastric banding in the morbidly obese adult population of Australia. *PLoS ONE*, **8(5),** e64965, doi:10.1371/journal.pone.0064965.

Malehi, A. S., Pourmotahari, F. and Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. *Health Economics Review*, **5(11)**, doi: 10.1186/s13561-015-0045-7.

Mihaylova, B., Briggs, A. H., O'Hagan, A. and Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, **20**, 897-916.

Roche Products Limited. (2008). *Rituximab for the 1st line Treatment of Chronic Lymphocytic Leukaemia*. National Institute For Health And Clinical Excellence, London, UK. Available from: http://www.nice.org.uk/guidance/ta174/resources/roche-submission2.

Speight, P. M., Palmer, S., Moles, D. R., Downer, M. C., Smith, D. H. *et al.* (2006). The cost-effectiveness of screening for oral cancer in primary care. *Health Technology Assessment*, **10(14**), doi:10.3310/hta10140.

Veldhuijzen, I. K., Toy, M., Hahne', S. J. M, De Wit, G. A., Schalm, S. W. *et al*. (2010). Screening and early treatment of migrants for chronic Hepatitis B virus infection is cost-effective. *Gastroenterology*, **138**, 522-530.

Wan, X., Peng, L. and Li, Y. (2015). A review and comparison of methods for recreating individual patient data from published kaplan-meier survival curves for economic evaluations: a simulation study. *PLoS ONE*, **10(3)**, e0121353, doi:10.1371/ journal.pone.0121353.

Welton, N. J., and Ades, A. E. (2005). Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making,* **25**, 633-645.

Williams, C., Lewsey, J. D., Briggs, A. H. and Mackay, D. F. (2017). Cost-effectiveness analysis in R using a multi-state modeling survival analysis framework: a tutorial. *Medical Decision Making*, **37**, 340-352.

Williams, C., Lewsey, J. D., Mackay, D. F. and Briggs, A. H. (2017). Estimation of survival probabilities for use in cost-effectiveness analyses: a comparison of a multi-state modeling survival analysis approach with partitioned survival and markov decision-analytic modeling. *Medical Decision Making*, **37(4)**, 427-439.

Wu, B., Li, T., Cai, J., Xu, Y. and Zhao, G. (2014). Cost-effectiveness analysis of adjuvant chemotherapies in patients presenting with gastric cancer after D2 gastrectomy. *BMC Cancer*, **14**, doi: 10.1186/1471-2407-14-984.

Yeh, J. M., Hur, C., Kuntz, K. M., Ezzati, M. and Goldie, S. J. (2010). Cost-effectiveness of treatment and endoscopic surveillance of precancerous lesions to prevent gastric cancer. *Cancer*, **116(12)**, 2941-2953.

TABLES AND FIGURES

Table 1: Summary of original data

Total no. of patients enrolled: 817	**Chemotherapy Group**	**Chemoimmunotherapy Group**
Total assigned to group	409	408
Lost to follow up	40	14
No. of PFS events observed	227	162
No. of death events observed	86	65
Total follow-up period	5 year	5 years

Source: *Hallek et al.* (2010)

Table 2: Chemotherapy data fit

Distribution	**AIC**
Exponential	2597.634
Gamma	2597.693
Weibull	2585.173
Log-logistic	2600.635
Log-normal	2620.277
Generalized Gamma	2596.533

Table 3: Chemoimmunotherapy data fit

Distribution	**AIC**
Exponential	3286.647
Gamma	3281.338
Weibull	3269.976
Log-logistic	3287.074
Log-normal	3323.709
Generalized Gamma	3280.407

Table 4: Chemotherapy—estimated transition probabilities (and 95% confidence intervals))

At t = 1 month	**Progression Free**	**Progression**	**Death**
Progression Free	0.990 (0.984; 0.993)	0.006 (0.004; 0.009)	0.004 (0.003; 0.007)
Progression	0	0.997 (0.995; 0.998)	0.003 (0.002; 0.005)
Death	0	0	1
At t= 50 months	**Progression Free**	**Progression**	**Death**
Progression Free	0.237 (0.192; 0.277)	0.348 (0.295; 0.399)	0.415 (0.365; 0.476)
Progression	0	0.658 (0.574; 0.726)	0.342 (0.274; 0.426)
Death	0	0	1
At t= 180 months	**Progression Free**	**Progression**	**Death**
Progression Free	0.001 (0; 0.003)	0.101 (0.043; 0.174)	0.898 (0.824; 0.957)
Progression	0	0.123 (0.05; 0.215)	0.877 (0.785; 0.950)
Death	0	0	1

At t= 240 months	**Progression Free**	**Progression**	**Death**
Progression Free	0	0.041 (0.011; 0.093)	0.959 (0.907; 0.989)
Progression	0	0.049 (0.012; 0.115)	0.951 (0.885; 0.988)
Death	0	0	1

Table 5: Chemoimmunotherapy—estimated transition probabilities (and 95% confidence intervals)

At t= 1 month	**Progression Free**	**Progression**	**Death**
Progression Free	0.994 (0.99; 0.997)	0.003 (0.002; 0.005)	0.003 (0.001; 0.005)
Progression	0	0.998 (0.996; 0.999)	0.002 (0.001; 0.004)
Death	0	0	1
At t= 50 months	**Progression Free**	**Progression**	**Death**
Progression Free	0.425 (0.371; 0.469)	0.253 (0.211; 0.304)	0.322 (0.277; 0.375)
Progression	0	0.705 (0.613; 0.779)	0.295 (0.221; 0.387)
Death	0	0	1
At t= 180 months	**Progression Free**	**Progression**	**Death**
Progression Free	0.012 (0.003; 0.032)	0.138 (0.061; 0.230)	0.850 (0.748; 0.933)
Progression	0	0.166 (0.063; 0.296)	0.834 (0.704; 0.937)
Death	0	0	1
At t= 240 months	**Progression Free**	**Progression**	**Death**
Progression Free	0.002 (0; 0.009)	0.065 (0.014; 0.143)	0.933 (0.853; 0.986)
Progression	0	0.074 (0.013; 0.174)	0.926 (0.826; 0.987)
Death	0	0	1

Table 6: Expected total length of stay in months (and 95% confidence intervals)

Chemotherapy 15-Year Horizon=180 months	**Progression Free**	**Progression**	**Death**
Progression Free	34.80 (31.74; 38.08)	41.66 (31.80; 51.73)	103.54 (92.78; 114.25)
Progression	0	85.14 (70.64; 99.43)	94.86 (80.57; 109.36)
Death	0	0	180
Chemotherapy 20-Year Horizon=240 months	**Progression Free**	**Progression**	**Death**
Progression Free	34.81 (31.56; 38.45)	45.67 (33.24; 59.04)	159.52 (145.35; 73.57)
Progression	0	90.01 (71.29; 108.45)	149.99 (131.55; 168.71)
Death	0	0	240
Chemoimmunotherapy 15-Year Horizon=180 months	**Progression Free**	**Progression**	**Death**
Progression Free	51.92 (46.25; 57.51)	37.48 (27.47; 47.48)	90.60 (79.56; 102.46)
Progression	0	93.52 (76.61; 108.83)	86.48 (71.17; 103.39)
Death	0	0	180
Chemoimmunotherapy 20-Year Horizon=240 months	**Progression Free**	**Progression**	**Death**
Progression Free	52.24 (46.28; 59.46)	43.38 (30.90; 58.34)	144.38 (127.39; 160.44)
Progression	0	100.40 (78.68; 122.97)	139.60 (117.03; 161.32)
Death	0	0	240

Table 7: Mean costs (in GBP) for 15-year lifetime time horizon

Cost Head	Chemoimmunotherapy		Chemotherapy		Incremental (*d*=3.5%)	Incremental (*d*=5%)
	Mean total cost (*d*=3.5%)	Mean total cost (*d*=5%)	Mean total cost (*d*=3.5%)	Mean total cost (*d*=5%)		
Mean cost of PFS	**18645.28**	**18605.12**	**6650.10**	**6634.65**	**11995.18**	**11970.47**
Costs of rituximab	10113	10113	0	0	10113	10113
Administration costs of rituximab	1224	1224	0	0	1224	1224
Cost of fludarabine	2776	2776	2790	2790	−14	−14
Administration costs of fludarabine	1109	1109	1115	1115	− 6	− 6
Costs of cyclophosphamide	21	21	22	22	−1	−1
Administration costs of cyclophosphamide	1109	1109	1115	1115	− 6	− 6
*Cost of supportive care in PFS	1061.28	1021.12	741.10	725.65	320.18	295.47
Cost of bone marrow transplantation	592	592	360	360	232	232
Cost of blood transfusions	640	640	507	507	133	133
Mean cost of progression	**7329.60**	**7178.44**	**8061.20**	**7893.01**	**− 731.60**	**− 714.57**
*Cost of supportive care in progression	1802.04	1764.88	1981.91	1940.56	− 179.87	− 175.68
*Cost of second-line & subsequent therapy	5527.55	5413.56	6079.28	5952.45	− 551.73	− 538.89
Mean total cost	**25974.88**	**25783.57**	**14711.30**	**14527.66**	**11263.58**	**11255.90**
*Calculated and discounted with respect to the total length of stay in the given state						

Table 8: Mean costs (in GBP) for 20-year lifetime time horizon

Cost Head	Chemoimmunotherapy		Chemotherapy		Incremental (*d*=3.5%)	Incremental (*d*=5%)
	Mean total cost (*d*=3.5%)	Mean total cost (*d*=5%)	Mean total cost (*d*=3.5%)	Mean total cost (*d*=5%)		
Mean cost of PFS	**18648.36**	**18611.09**	**6650.10**	**6634.65**	**11998.26**	**11976.45**
Costs of rituximab	10113	10113	0	0	10113	10113
Administration costs of rituximab	1224	1224	0	0	1224	1224
Cost of fludarabine	2776	2776	2790	2790	−14	−14
Administration costs of fludarabine	1109	1109	1115	1115	− 6	− 6
Costs of cyclophosphamide	21	21	22	22	− 1	− 1
Administration costs of cyclophosphamide	1109	1109	1115	1115	− 6	− 6
*Cost of supportive care in PFS	1064.36	1027.09	741.10	725.65	323.26	301.45
Cost of bone marrow transplantation	592	592	360	360	232	232
Cost of blood transfusions	640	640	507	507	133	133

Mean cost of progression	**8377.23**	**8160.49**	**8769.82**	**8547.54**	**− 392.60**	**− 387.05**
*Cost of supportive care in progression	2059.61	2006.33	2156.14	2101.49	− 96.52	− 95.16
*Cost of second-line & subsequent therapy	6317.61	6154.16	6613.69	6446.06	− 296.07	− 291.89
Mean Total cost	**27025.58**	**26771.58**	**15419.92**	**15182.19**	**11605.66**	**11589.39**

*Calculated and discounted with respect to the total length of stay in the given state

Table 9: QALYs and ICERs (in GBP/ QALY)

State	**15-year Horizon**		**20-year Horizon**	
	***d* = 3.5%**	***d* = 5%**	***d* = 3.5%**	***d* = 5%**
Gain in QALY for a patient in PF state at randomization−chemotherapy				
Progression Free	2.25	2.22	2.25	2.22
Progression	2	1.97	2.18	2.14
Death	0	0	0	0
Total QALY	4.25	4.19	4.43	4.36
Gain in QALY for a patient in PF state at randomization−chemoimmunotherapy				
Progression Free	3.28	3.20	3.29	3.22
Progression	1.81	1.78	2.08	2.04
Death	0	0	0	0
Total QALY	5.09	4.98	5.37	5.26
QALY Incremental	0.84	0.94	0.79	0.90
Cost Incremental	11263.58	11255.90	11605.66	11589.39
ICER	13409.02	11974.37	14690.71	12877.10

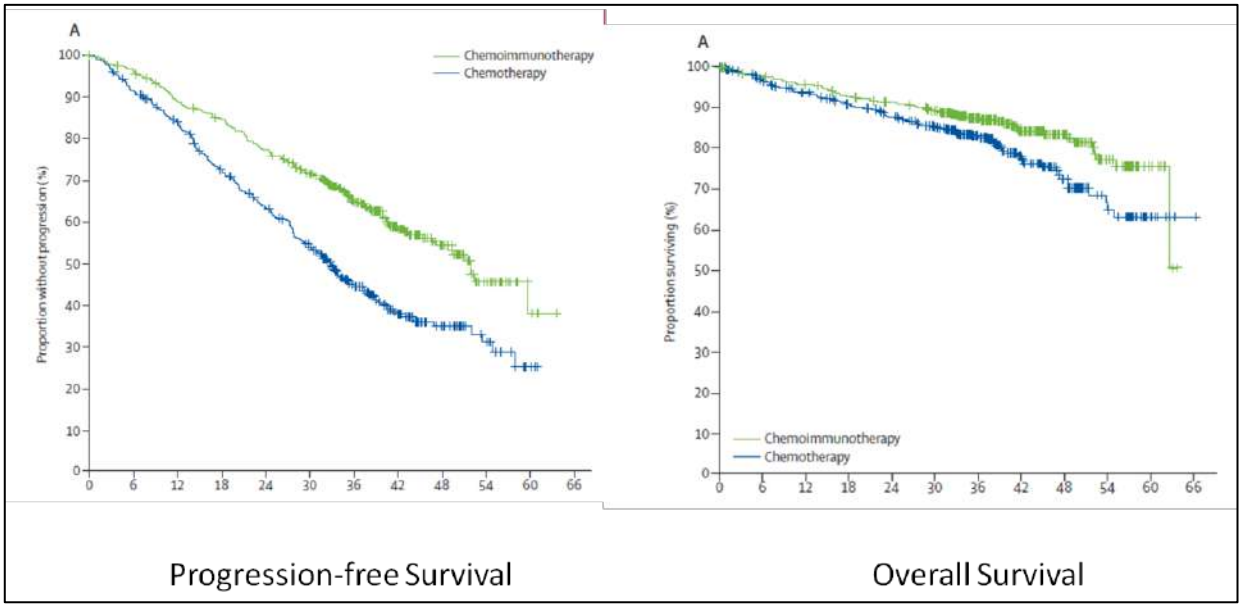
Table 10: Bayesian model—estimated transition probabilities (with 95% confidence limits)

Chemotherapy	**Progression Free**	**Progression**	**Death**
Progression Free	0.99 (0.977; 0.998)	0.01 (0.002; 0.023)	0
Progression	0	0.794 (0.401; 0.0.993)	0.206(0.007; 0.599)
Death	0	0	1
Chemoimmunotherapy	**Progression Free**	**Progression**	**Death**
Progression Free	0.993 (0.982; 0.998)	0.007 (0.002; 0.018)	0
Progression	0	0.798 (0.414; 0.993)	0.202 (0.007; 0.586)
Death	0	0	1

Table 11: Bayesian model—expected total length of stay in each state for 15-year lifetime time horizon (in months)

Chemotherapy	**Progression Free**	**Progression**	**Death**
Progression Free	90.91 (72.97; 145.91)	10.17 (2.45; 56.33)	78.92 (61.29; 127)
Progression	0	15.40 (3.30; 102.15)	164.60 (160.12; 178.33)
Death	0	0	180
Chemoimmunotherapy	**Progression Free**	**Progression**	**Death**
Progression Free	106.70 (85.76; 157.38)	8.54 (2.11; 45.74)	64.77 (44.11; 118.99)
Progression	0	15.54 (3.32; 103.76)	164.46 (160.12; 178.29)
Death	0	0	180

Figure 1: Kaplan-Meier curves used for data reconstruction



Source: *Hallek et al.* (2010)

Figure 2: Kaplan-Meier curves from reconstructed data

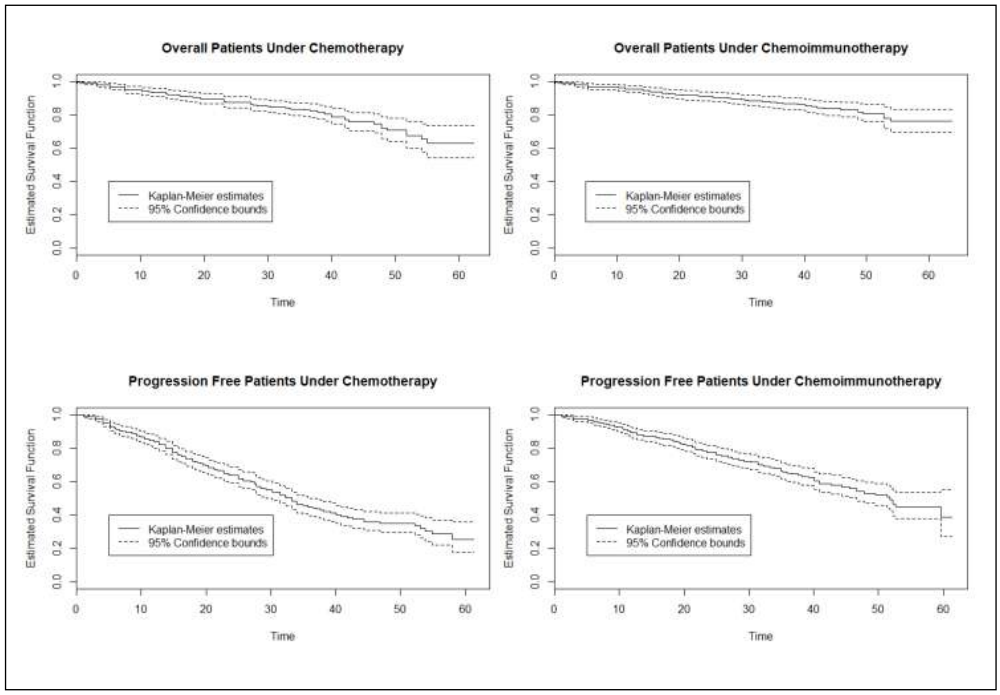


Figure 3: Transition map between three states of the multi-state model

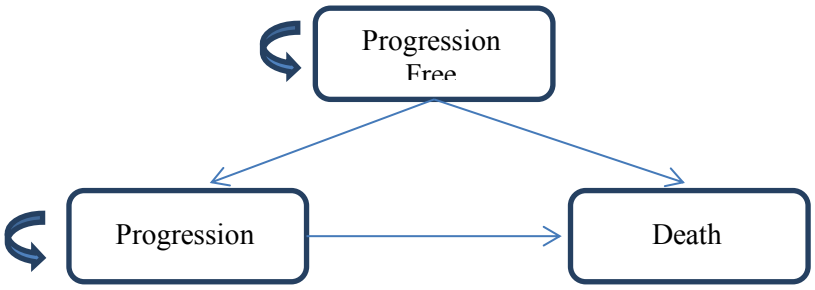


Figure 4: Methodological structure

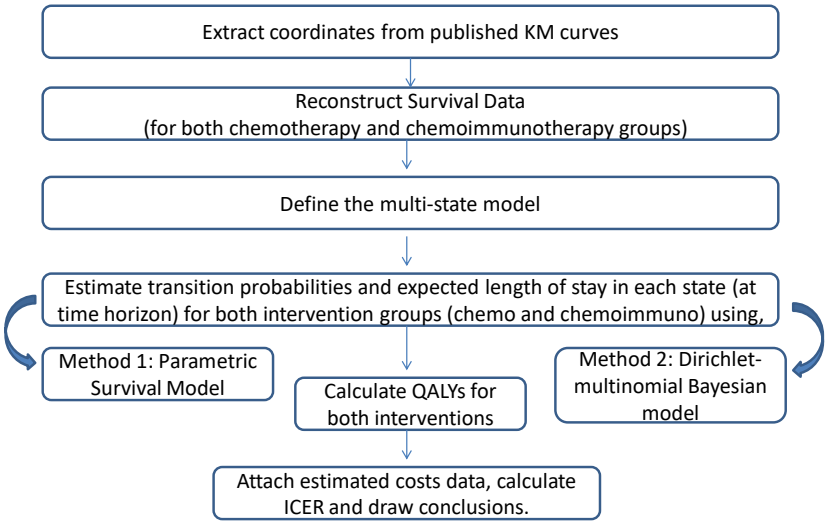


Figure 5: Cost-effectiveness plane from the first method

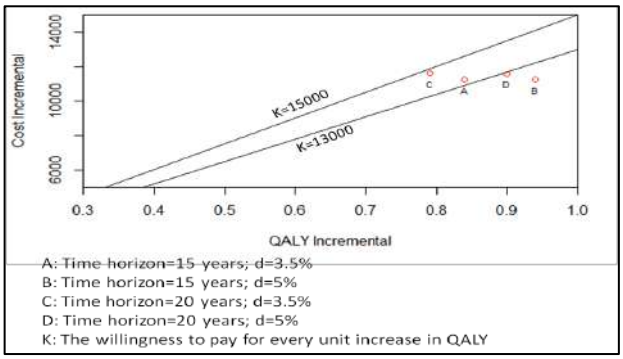
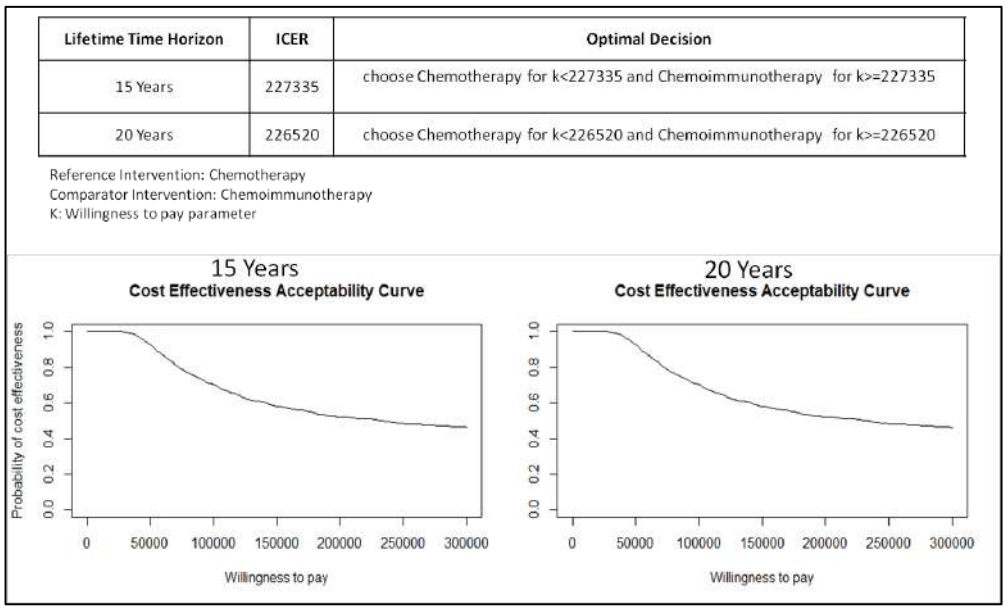


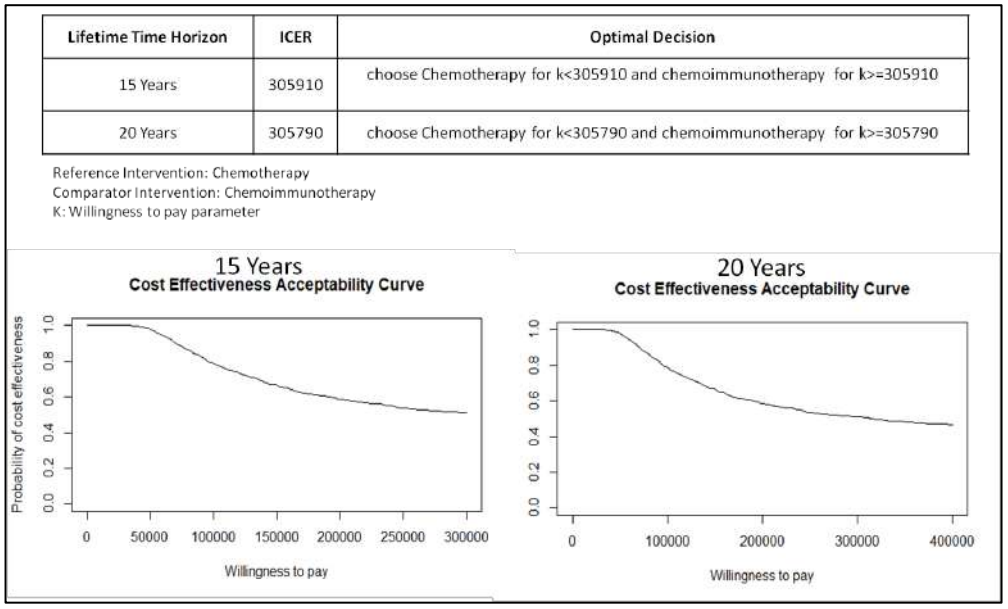
Figure 6: Results of Bayesian cost-effectiveness analysis using BCEA (d = 3.5%)



Lifetime Time Horizon	ICER	Optimal Decision
15 Years	227335	choose Chemotherapy for k<227335 and Chemoimmunotherapy for k>=227335
20 Years	226520	choose Chemotherapy for k<226520 and Chemoimmunotherapy for k>=226520

Reference Intervention: Chemotherapy
Comparator Intervention: Chemoimmunotherapy
K: Willingness to pay parameter

Figure 7: Results of Bayesian cost-effectiveness analysis using BCEA (d = 5%)



Lifetime Time Horizon	ICER	Optimal Decision
15 Years	305910	choose Chemotherapy for k<305910 and chemoimmunotherapy for k>=305910
20 Years	305790	choose Chemotherapy for k<305790 and chemoimmunotherapy for k>=305790

Reference Intervention: Chemotherapy
Comparator Intervention: Chemoimmunotherapy
K: Willingness to pay parameter

Figure 8: Proportion of patients expected to be in each state at different virtual follow-up time points—Weibull time-inhomogeneous semi-markov model

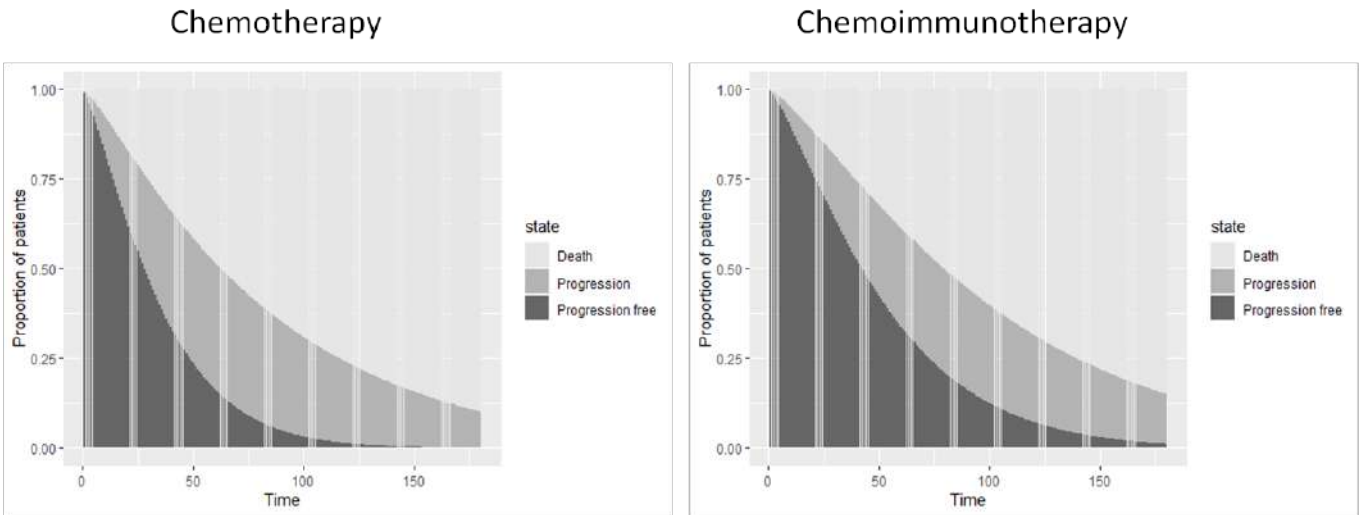
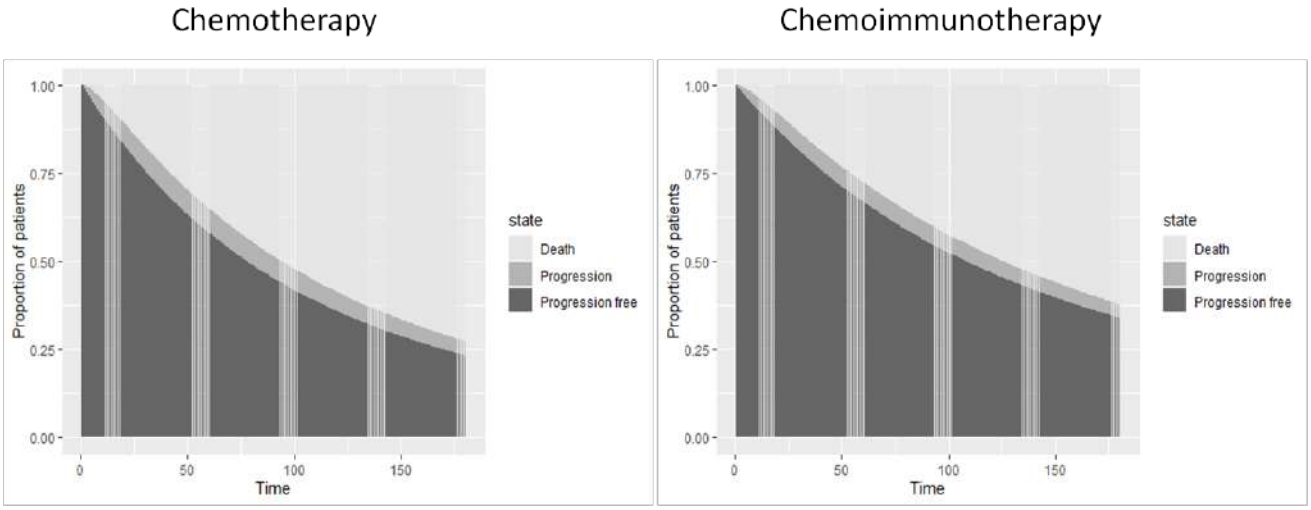


Figure 9: Proportion of patients expected to be in each state at different virtual follow-up time points—Multinomial-Dirichlet Bayesian model



Appendix-A

Figure A.1: Estimated survival functions of fitted models—Chemotherapy group

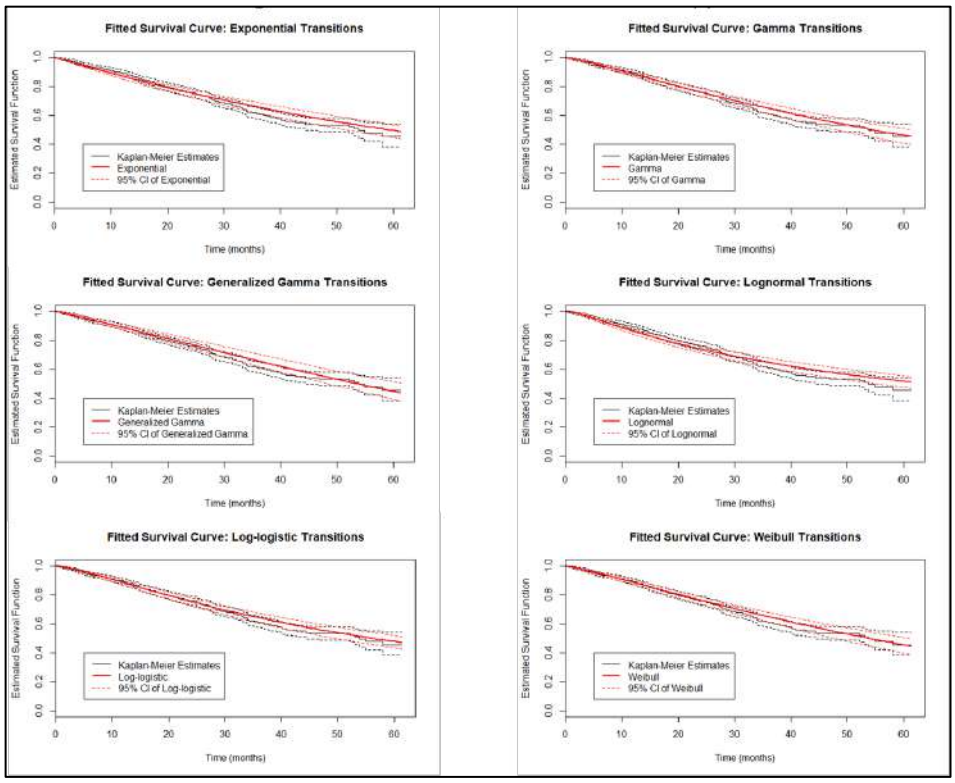
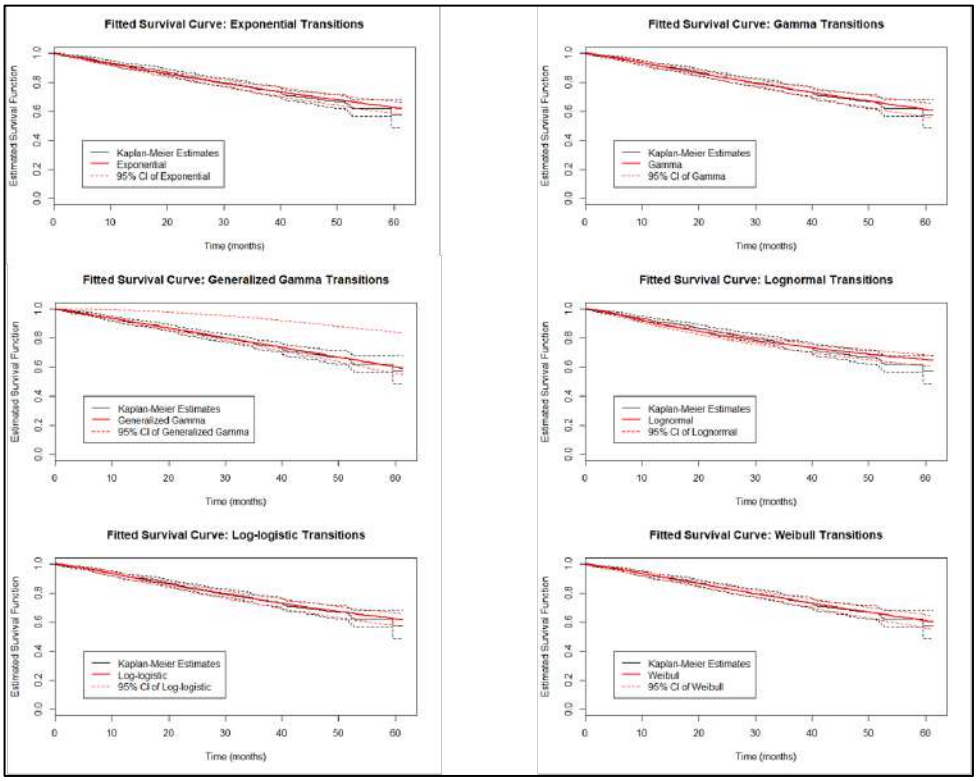


Figure A.2: Estimated survival functions of fitted models—Chemoimmunotherapy group



Appendix-B

Figure B.1: Transition probabilities plotted against time—Weibull semi-markov model

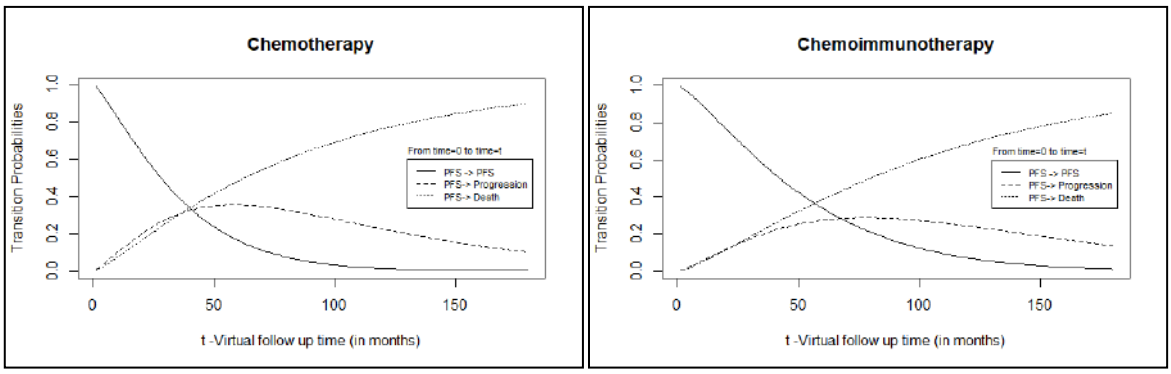
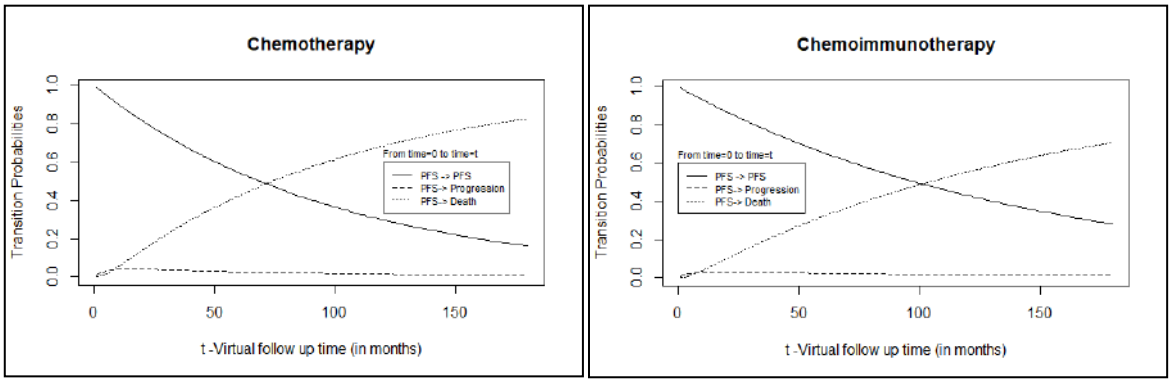


Figure B.2: Transition probabilities plotted against time—Multinomial-Dirichlet Bayesian model



Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 55-62

Weighted Sustainable Development Goal Index

A. K. Nigam[1] and Manoj Kumar Pant[2]

[1]*Consultant Advisor, IASDS*
[2]*Additional Chief Executive Officer, Centre for Public Policy and Good Governance, Department of Planning, Government of Uttarakhand*

Received: 13 March 2020; Revised: 26 March 2020; Accepted: 30 March 2020

Abstract

Sustainable Development Goal (SDG) India Index, developed by NITI AAYOG in 2018, is to oversee the progress of the nation and the states on the achievements of Sustainable Development Goals (SDG). The document presents SDG Indices of different major goals. A major limitation of these SDG indices is their unweighted form because these are developed using arithmetic mean as average even in the presence of extreme values. The purpose of the present work is essentially to propose weights to develop weighted index numbers for SDGs. We follow the methodology on developing weights for different SDG indicators developed recently by Nigam (2019). The utility of the proposed methodology is demonstrated by applying it to SDG data on poverty.

Key words: Sustainable development goals; Weighted SDG index; Relative gap; Choice of weights.

1. Introduction

The Sustainable Development Goals (SDGs) are a set of 17 global goals set by the United Nations General Assembly in 2015 for the year 2030. The SDGs are part of Resolution 70/1 of the United Nations General Assembly, the 2030 Agenda. The SDGs are: No Poverty (End poverty in all its forms everywhere), Zero Hunger (End hunger, achieve food security and improved nutrition and promote sustainable agriculture), Good Health and Well-being (Ensure healthy lives and promote well-being for all at all ages), Quality Education (Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all), Gender Equality(Achieve gender equality and empower all women and girls), Clean Water and Sanitation (Ensure availability and sustainable management of water and sanitation for all), Affordable and Clean Energy (Ensure access to affordable, reliable, sustainable and modern energy for all), Decent Work and Economic Growth (Promote sustained, inclusive and sustainable economic growth, fill and productive employment and decent work for all), Industry, Innovation, and Infrastructure (Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation), Reducing Inequality (Reduce inequality within and among countries), Sustainable Cities and Communities (Make cities and human settlements inclusive, safe, resilient and sustainable), Responsible Consumption and Production (Ensure sustainable consumption and production patterns), Climate Action (Take urgent action to combat climate change and its impacts), Life Below Water, Life On Land (Protect, restore and promote sustainable use of terrestrial ecosystems,

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

sustainably manage forests, combat desertification and halt and reverse land degradation and halt biodiversity loss), Peace, Justice and Strong Institutions (Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels), and Partnerships for the Goals (Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development).

The goals are broad based and interdependent. Each of the 17 SDGs has a list of targets that are measured with indicators. Given the importance accorded by the Government of India to achieving SDGs, NITI Aayog decided to come out with the progress through a single measurable index that would serve as an advocacy tool and trigger action at the State level. With this purpose, SDG India Index report was launched by NITI Aayog in December 2018.

The SDG India Index is an aggregate measure which can be understood and used by everyone - planners, policymakers, academicians, businesses, civil society and the general public. It has been designed to provide an aggregate assessment of the performance of all Indian States and UTs, and to help leaders and change makers evaluate their performance on social, economic and environmental parameters. It aims to measure India and its States' progress towards the SDGs for 2030.

There are already several index numbers in related areas, like food insecurity by M.S. Swaminathan Research Foundation (MSSRF), global hunger index (GHI) by IFPRI, Yale's environmental performance index, sustainable environmental performance index by DES Uttarakhand and IIFM, SDG indices by NITI Aayog and some related work by Nigam (2018a, b).

Work on GHI and SDG indices by NITI Aayog follows MSSRF methodology and hence suffers from all those limitations as in food insecurity indices by MSSRF. A major limitation is the unweighted form of these indices and also use of arithmetic mean as average even in the presence of extreme values.

The present work aims at proposing weights to develop weighted index numbers for SDGs. The methodology followed on developing weights for different SDG indicators is essentially the one developed recently by Nigam (2019). We demonstrate the utility of the proposed methodology by applying it to data on poverty SDG.

2. Niti Aayog Methodology

Just as a preliminary and for the sake of completeness, it is essential to describe briefly the SDG India Index. For details one may refer to SDG Index India Report, 2018 by Niti Aayog.

To make data comparable across indicators, State-wise data values of each of the Priority Indicators were rescaled from its raw form into a score ranging from 0 to 100, with 0 denoting lowest performer and 100 indicating that the target has been achieved.

For indicators where increasing value means better performance (*e.g.* forest area coverage), score was computed as follows:

$$x' = \frac{x - min\,(x)}{T(x) - min\,(x)} \times 100 \qquad (1)$$

where, x = raw data value, $min(x)$ = minimum observed value of the indicator in the dataset, $T(x)$ = national target value of the indicator, x' = normalized value after rescaling.

For indicators where increasing value means worse performance (*e.g*. Poverty rate), score was computed as follows:

$$x' = \left[1 - \frac{x - T(x)}{max(x) - T(x)}\right] \times 100 \quad (2)$$

where, x = raw data value, $max(x)$ = maximum observed value of the indicator in the dataset, $T(x)$ = national target value of the indicator, x' = normalized value after rescaling.

In instances where States and Union Territories (UTs) performed better than the target, their Index Score has been capped at 100.

SDG Index Score: For each of the Goals under SDGs (except Goals 12, 13, 14 and 17), SDG India Index Score was computed for each State/UT. This was calculated as the arithmetic mean of the normalised values of all the Priority Indicators within the Goal. In calculating the average, equal weights were assigned to each indicator and the arithmetic mean was rounded off to the nearest whole number. The Goal scores for the respective states are computed using

$$I_{ij}\left(N_{ij}, I_{ijk}\right) = \sum_{k=1}^{N_{ij}} \frac{1}{N_{ij}} \times I_{ijk} \quad (3)$$

where, I_{ij} = Goal score for State i under SDG j, N_{ij} = Number of non-null indicators for State I under SDG j, I_{ijk} = Normalised value for State i of indicator k under SDG j.

The Goal Score I_{ij} for State i under SDG j was then rounded off to the nearest whole number to give the SDG Index Score. Based on the SDG India Index, States and UTs were classified into 4 categories under each of the SDGs (except Goals 12, 13, 14 and 17):

- Achiever – when SDG India Index score is equal to 100
- Front Runner – when SDG India Index score is less than 100 but greater than or equal to 65
- Performer – when SDG India Index score is less than 65 but greater than or equal to 50
- Aspirant – when SDG India Index score is less than 50

Composite SDG India Index Score: Every State's and UT's composite SDG India Index score was finally computed to quantify the overall progress of the States and UTs towards the SDGs. This was calculated as the arithmetic mean of the Goal scores across 13 out of the 17 Goals. This was done by assigning equal weight to every Goal score and the arithmetic mean was rounded off to the nearest whole number.

$$I_i\left(N_i, N_{ij}, I_{ijk}\right) = \frac{1}{N_i} \sum_{j=1}^{N_i} I_{ij} \times \left(N_{ij}, I_{ijk}\right) \quad (4)$$

where, I_i = Composite SDG index score of State i, N_i = Number of Goal scores for which State i has nonnull data, I_{ij} = Goal score for State i under SDG j.

The arithmetic mean of Goal scores are then rounded off to the nearest whole number to give the composite SDG India Index score for each State/UT. The States/UTs are again classified into the four categories (as described above): Achiever, Front Runner, Performer and Aspirant.

3. Proposed Methodology

When all the commodities are not of equal importance, we assign weight to each commodity relative to its importance and the index number computed from these weights is called a weighted index number. If the base year quantities are used as weights, then it also called the base year weighted index. Generally, planners and policymakers are entrusted to monitor the progress of those indicators which are more beneficial for the people. Therefore, identification of such indicators with their weights might be useful for measuring the real progress. For assigning the weights to different indicators, many methods can be adopted which depend upon the availability of time series data (Regression Method), Group Exercises (Analytical Hierarchy Process) and so on. Here we have proposed gap between target and current progress of base year as the weight. It is not easy to assign weights to respective indicators by regression and other related methods as the availability of time series data is a big challenge. We have adopted a new method where we take the base year data of target and achievement and assign the weights as their gap. More the gap, less the weight is the simple phenomenon of this gap-based weight assignment methodology. In contrast, NITI AAYOG SDG India Index methodology provides the index score with equal weight to all indicators.

Let w_{jk}, be the weight assigned to the kth indicator in the jth goal, there being p goals and n_j indicators in the jth goal, $j = 1,\ldots,p$, $k= 1,\ldots, n_j$. It may be noted that the total number of weights is the sum of n_j over j.

We first describe how to work out the weights for the SDG index.

Step 1. Compute the gap between current/baseline figure (raw data) and the target. If the target has already been achieved, fix the weight as 100. If the target is yet to be achieved compute the relative gap. Relative gap is the ratio of the gap and the current value. The weight is then the inverse of the relative gap. Using the inverse as the weight ensures that poor performing indicators (with larger relative gap) are assigned lower weight and contribute little to the overall index score.

Step 2. The next step is to normalize the weights and make them vary between 0 and 1. The normalized weights are obtained as the ratio of the inverse of relative gap and the total of inverse of relative gaps.

Step 3. The overall index number of the indicators, which are yet to achieve their targets, is the product of weights and the current values of indicators.

Step 4. The overall weighted SDG index number is given by the arithmetic mean of 100 and the index number of the indicators which are yet to achieve their targets (as computed in Step 3).

The above method can easily be extended to include minor variants of the case discussed above.

4. Example

Consider the all India SDG raw data on poverty as given in Niti Aayog's SDG Index. Table 1 gives raw data, unweighted and weighted SDG indices of Indian States on poverty. First five columns starting from third column give raw data, and the next five columns give their normalized values using formulae (1) and (2). These also give unweighted SDG indices and their average (arithmetic mean) gives the unweighted SDG index (SDG 1 in Table 1) as reported in the India SDG index, 2018. The next column gives the rank of states labeled as rank (1). The following column gives weighted SDG indices, weights based upon the gap and target values as per the proposed methodology in steps 1-5, and their average gives the weighted SDG index (SDG 2 in Table 1). The last column gives the rank of states labeled as rank (2).

A comparison between ranks (1) and (2) show that the ranks of states are changed with no definite pattern. Nevertheless, the ranks (2) are more logical and realistic. An examination of the two ranks reveals that Tamil Nadu which ranked 1 in the unweighted category with the index score 68 is also top ranked with score 99 in the weighted index category. On the other hand, Haryana which ranked 25 in the unweighted category with the score 25, has rank 1 in the weighted category with a score of 99. Uttarakhand's weighted index is also ranked on top though it ranked 7 in the unweighted category. As a matter of fact, besides Haryana and Uttarakhand, Gujarat, Mizoram and Daman & Diu are also on the top with identical scores of 99.

The above comparison opens up the debate on how the ranks of SDG 2 are changing in a haphazard manner *vis-a-vis* SDG 1. The issue was a point of thorough discussion on first author's Special lecture on Hunger and Related SDGs in a 2-days Workshop on SDGs at Hyderabad (23-24 January 2020). The Workshop was organized by Ministry of Statistics and Programme Implementation (MoSPI), Government of India, with the objective of Capacity Building of Senior Indian Statistical Services officers of Deputy Director General and above level. The discussion led to a recommendation that it needs a full-fledged research study of around 2 - years duration to find answers to different types of questions bothering us. The authors are already working on it and the results obtained will be reported in a separate communication in the future.

Yet another problem that needs to be examined is in the normalization of Priority Indicators, which were rescaled from its raw form into a score ranging from 0 to 100, with 0 denoting lowest performer and 100 indicating that the target has been achieved. If we consider the Percentage of population living below the national poverty line, the first indicator of the goal Poverty given in the first column of Raw data in Table 1, we note that for all raw data values 9.2, 5.09, 8.06, and few others the Index Score is 100. This seems illogical and requires rescaling modification. Similar is the scenario in case of Index Scores 0. It requires again substantial efforts to modify formulae (1) and (2).

5. Choice of Other Weights

Besides gaps between current/baseline value, there can be other choices of weights. We may pick up inverse of the number of observations used to compute the value for a given indicator.

Another choice could be the coefficient of variation for each indicator value. If we choose more than one weight, then the overall weight can be taken as the geometric mean of all the weights.

While computing weighted index numbers usually geometric mean is preferred against arithmetic mean or harmonic mean. In most cases, we may have weights with extreme values and this makes geometric mean superior to arithmetic mean. The easier interpretability of geometric mean makes it a better choice than harmonic mean. While computing weighted index numbers usually geometric mean is preferred against arithmetic mean or harmonic mean.

Biggest hurdle in choosing more than one weight is the non-availability of data on these weights. However, it is advisable to use as many weights as possible to derive a good weighted index.

Acknowledgements

We thank the reviewer and the editors for their suggestions that helped considerably improve the contents.

References

Nigam, A. K. (2018a). Improving global hunger index. *Agricultural Research*, **7**, e-print (2019) **8**, 132-139.

Nigam, A. K. (2018b). Global hunger index – Revisited. *Journal of the Indian Society of Agricultural Statistics*, **72**, 225–230.

Nigam, A. K. (2019). *Weighted SDG Index*. Unpublished Manuscript. Also, part of the Key Note Address on Zero Hunger at a National Seminar on Statistics and SDGs, organized at Statistics Department, Lucknow University.

Nigam, A. K. (2020). *Hunger and Related SDGs*. Workshop for capacity building of Senior ISS officers of DDG and above level on SDGs 23-24 January 2020 organized at Hyderabad.

SDG India Index. Baseline Report (2018). Niti Aayog.

Table 1: Unweighted and Weighted SDG Indices of Indian States

(*The description of column numbers is given in Note 2 at the end of the table*)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
		Raw Data					**Index Score**								
1	AnP	9.2	74.6	87.77	17.4	15.34	100	74	72	16	72	67	**5**	86	**24**
2	ArP	34.67	58.3	85.56	20.5	0.23	18	57	67	19	100	52	**22**	79	**31**
3	AS	31.98	10.4	87.1	66.1	3.42	27	7	71	66	94	53	**21**	94	**10**
4	BI	33.74	12.3	75.63	53.9	3.13	21	9	44	53	94	45	**30**	94	**10**
5	CH	39.93	68.5	77.25	66.2	15.77	0	67	48	66	71	50	**25**	83	**28**
6	Goa	5.09	15.9	98.15	7.4	1.62	100	13	96	6	97	62	**9**	98	**6**
7	GU	16.63	23.1	80.92	8.9	14.06	80	20	57	8	74	48	**27**	86	**24**
8	Hy	11.16	12.2	80.16	13.5	14.52	99	9	55	12	73	50	**25**	99	**1**
9	HP	8.06	25.8	91.12	13.1	9.37	100	23	80	12	83	60	**12**	89	**18**
10	J&K	10.35	4.2	88.79	54	12.33	100	1	74	53	77	61	**10**	86	**24**
11	JR	36.96	13.3	70.34	41.6	4.38	10	10	32	41	92	37	**35**	93	**12**
12	UK	11.26	19.5	90.1	49.4	8.88	99	17	77	49	84	65	**7**	99	**1**
13	KN	20.91	28.1	84.26	19.9	8.45	66	26	64	19	85	52	**22**	87	**21**
14	KL	7.05	47.7	87.98	20.4	3.4	100	46	73	19	49	66	**6**	80	**29**
15	MP	31.65	17.7	79.68	61.1	21.42	29	15	54	61	61	44	**31**	77	**32**
16	MH	17.35	15	86.4	8.7	17.18	78	12	69	7	68	47	**29**	86	**24**
17	MN	36.89	3.6	96.46	26.2	2.95	10	0	92	25	95	44	**31**	96	**9**
18	ML	11.87	34.6	97.3	28	4.24	97	32	94	27	92	68	**4**	97	**8**
19	MZ	20.4	45.4	99.94	47.5	0.62	67	43	100	47	99	71	**2**	99	**1**
20	NL	18.88	6.1	98.91	29.7	3.03	73	3	98	29	94	59	**13**	98	**6**
21	OR	32.59	47.7	84.76	72.6	7.11	25	46	65	72	87	59	**13**	88	**19**
22	PB	8.26	21.2	81.63	19.1	6.59	100	18	58	18	88	56	**19**	90	**15**
23	RJ	14.71	18.7	85.07	56.1	16.51	87	16	66	55	70	59	**13**	88	**19**
24	SK	8.19	30.3	94.16	29.4	11.75	100	28	87	28	78	64	**8**	90	**15**
25	TN	11.28	64.1	98.83	29.5	4.56	99	63	97	28	92	76	**1**	99	**1**
26	TG	Null	66.4	77.06	12.2	8.92	Null	65	48	11	84	52	**22**	87	**21**
27	TR	14.05	58.1	94.38	32.6	5.34	89	57	87	32	90	71	**2**	93	**12**
28	UP	29.43	6.1	84.3	48.7	5.36	36	3	64	48	90	48	**27**	90	**15**
29	WB	19.98	33.4	87.63	28.7	8.28	69	31	72	28	85	57	**17**	87	**21**
30	AN	1	5.7	92.46	1.4	0.97	100	2	83	0	98	57	**17**	51	**36**
31	CD	21.81	21.3	Null	13.7	19.7	63	19	Null	12	64	39	**34**	80	**29**
32	DN	39.31	30.8	Null	2.6	24.94	2	28	Null	1	54	21	**37**	73	**33**

33	DD	9.86	17	Null	19.3	0.67	100	14	Null	18	99	58	**16**	99	**1**
34	DL	9.91	16.4	Null	7.9	54.52	100	13	Null	7	0	30	**36**	56	**35**
35	LK	2.77	3.4	56.13	17.5	0	100	0	0	16	100	43	**33**	58	**34**
36	PD	9.69	32.8	93.78	21.4	3.97	100	30	63	20	93	61	**10**	93	**12**
	IN	**21.92**	**28.7**	**84.75**	**36.4**	**10.39**	**62**	**26**	**65**	**35**	**81**	**54**	**20**	**70**	
	Tt	**10.95**	**100**	**100**	**100**	**0**	**100**	**100**	**100**	**100**	**100**	**100**			

Note 1: AnP – Andhra Pradesh; ArP – Arunachal Pradesh; AS – Assam; BI – Bihar; CH – Chhattisgarh; GU – Gujarat; HY – Haryana; HP – Himachal Pradesh; J&K – Jammu and Kashmir; JR – Jarkhand; UK – Uttarakhand; KN – Karnataka; KL – Kerala; MP – Madya Pradesh; MH – Maharashtra; MN – Manipur; ML – Meghalaya; MZ – Mizoram; NL – Nagaland; OR – Orissa; PB – Punjab; RJ – Rajasthan; SK – Sikkim; TN – Tamil Nadu; TG – Telangana; TR – Tripura; UP – Uttar Pradesh; WB – West Bengal; AN – Andaman and Nicobar Islands; CD – Chandigarh; DN – Dadra and Nagar Haveli; DD – Daman and Diu; DL – Delhi; LK – Lakshadweep; PD – Pondicherry; IN – India; Tt – Target.

Note 2: The columns are

(1) Serial Number
(2) State
(3) % population living below National Poverty line
(4) % households with any usual member covered by any health scheme or health insurance
(5) Persons provided employment as % of persons who demanded employment under MGNREGA
(6) % of the population (out of total eligible population) receiving social protection benefits maternity benefits
(7) Number of homeless households per 10,000 households
(8) % of population living below National Poverty line
(9) % of households with any usual member covered by any health scheme or health insurance
(10) Persons provided employment as a percentage of persons who demanded employment under MGNREGA
(11) Proportion of the population (out of total eligible population) receiving social protection benefits maternity benefits %
(12) Number of homeless households per 10,000 households
(13) SDG 1 Index score
(14) Rank (1)
(15) SDG 2 Index score
(16) Rank (2)

Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 1, 2020 (New Series), pp 63–83

A Reward-Earning Quaternary Random Walk on a Parity Dial

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

Received: 26 March 2020; Revised: 15 April 2020; Accepted: 20 April 2020

Abstract

A casino offers a game which involves a symmetric quaternary random walk on a parity dial with twelve nodes labeled as $(1, 11, 3, 9, 5, 7, 6, 8, 4, 10, 2, 0)$, reading clockwise. A player begins at Node 0; she tosses a copper coin to decide whether to move clockwise (if heads) or counterclockwise (if tails); simultaneously she tosses a silver coin to decide whether she will move one step (if tails) or two steps (if heads) in the direction determined by the copper coin. Whenever she lands at a new node she is said to have 'captured' it. If a player intends to capture c nodes and she wishes to toss the coins k times, then her admission fee is $(25 + 25c + k)$ cents (one quarter to play, one quarter per node to capture and one penny per toss). The game ends as soon as either c nodes (other than Node 0) are captured or k tosses are over, whichever event happens earlier; and the player earns as many nickels as the sum of the labels of the captured nodes. How should the player determine c and k?

The player's optimal choices can be derived from the theory of stochastic processes. Alternatively, optimal choices can be anticipated through a computer simulation. Lessons learned from the game empower entrepreneurs and consumers behave optimally to determine when and how to intervene to benefit from an opportunity and/or to prevent a catastrophe.

Key words: Probability mass function; Stopping time; Optimal strategy; Central limit theorem; Law of large numbers.

AMS Subject Classifications: 60G50, 05C81

1. Introduction

When you agree to play a game of chance offered by a casino, you should expect to lose money on average. You accept this loss in anticipation of some entertainment, and a rare possibility of winning big. When millions of players play the game multiple times, the casino makes a positive profit even after paying occasional windfalls, administrative costs, staff salaries, discounts and government taxes. When a game appears to be favorable to the player, it attracts many participants. Of course, if a game were truly favorable to the player,

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

the casino would stop offering the game. But if a game only *appears* to favor the player, the casino can entice more players play it more often, and earn more profit for itself. The casino must know ahead of time the exact long-run performance of each game it offers, while the player is oftentimes attracted by the lure of apparent gain. Sarkar (2020 a) introduced such a game of a random walk on a parity dial, and proposed a wide variety of modifications to the game. In this paper, we change the rules of the random moves—from binary walk to quaternary walk—and find the optimal decision for the player.

The game serves as a model for decisions made by entrepreneurs and customers—both parties maximize their gains while abiding by some rules and coping with inherent uncertainty. The optimal decisions for each party may be derived using the theory of stochastic processes. See Ross (1996) for the general theory, and see Sarkar (2006) and Maiti and Sarkar (2019) for random walks on a circle. However, the theory being generally inaccessible to the common person, one can take recourse to a computer simulation involving repeated plays of the game. Lessons learned from the game equip all parties engaged in the marketplace to determine when and how to participate to benefit from an opportunity and/or to prevent a catastrophe. For an optimization problem of a different flavor (investing the smallest amount of input to extract a desired quality of output), see Sarkar (2020 b).

In Section 2, we describe the game of quaternary walk on the parity dial. In Section 3, for $c \leq 3$, we discover the optimal number of tosses k using exact probability distributions. In Section 4, for $4 \leq c \leq 11$, we find the optimal k via simulation. In Section 5, we give some theoretical results and beckon the reader to discover more. Section 6 compares the game of quaternary walk with that of binary walk. In Section 7, we pose some modified games and invite the reader to discover new optimal decisions.

All computations are done using the freeware R. Some codes are given in the Appendix.

2. Rules of the Game

Consider a network of twelve nodes arranged in a circle. The nodes are labeled $\mathbf{l} = (1, 11, 3, 9, 5, 7, 6, 8, 4, 10, 2, 0)$ reading clockwise. See Fig. 1. Note that the labels l_i $(1 \leq i \leq 12)$ are distinct non-negative integers, obtained from the usual dial of a clock by changing the top node from 12 to 0; and interchanging nodes within pairs $(2, 11)$, $(4, 9)$, and $(6, 7)$. Note that all odd values are on the right half (going clockwise from the top), while all even values are on the left half (going counterclockwise from the top) of the dial. Therefore, we call this network the parity dial.

Sarkar (2020 a) studied the following: "A player pays an admission price to play a game of random walks on the parity dial by repeatedly tossing a fair coin. Starting from Node 0, after each toss the player moves one position clockwise (if heads) or one position counterclockwise (if tails); and she captures the visited node. The player has total liberty to determine c, the number of nodes she intends to capture, and k, the number of times she wishes to toss. The game ends as soon as either c nodes (other than Node 0) are captured or k tosses are over. The player pays an admission price of $(25c + k)$ cents, and earns as reward as many nickels as the sum of the labels on the captured nodes. How should the player determine (c, k) to maximize her expected net reward?"

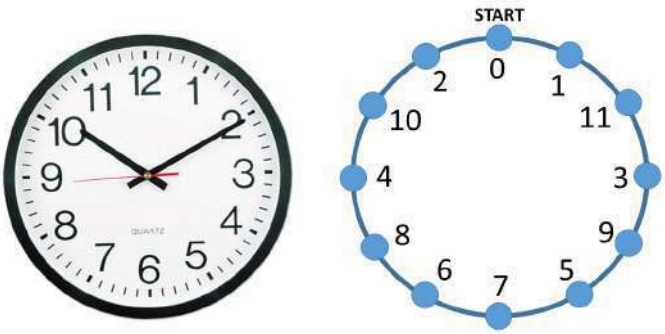


Figure 1: The usual dial of a clock and the parity dial

The answer to that problem is $(6, 28)$ for which the player pays 178 cents and earns 160.7961 cents on average (with a standard deviation (SD) of 0.0243 cents). Therefore, she loses on average 9.66% of her wager. With any other (c, k) game, she will lose even more. In this optimal game, the player tosses on average 16.2 times.

There are many possible modifications to the game. In this paper, we change the nature of the random walk: Instead of going one step clockwise or one step counterclockwise according as the outcome of a toss of a fair coin is heads or tails, we allow each move to be one or two steps clockwise, or one or two steps counterclockwise according to the outcome of tossing two coins simultaneously. We call this modified game the quaternary random walk game and the original game the binary random walk game. Since in the quaternary walk game the player has more opportunities of capturing a new node in each move than in the binary walk game, the admission fee of the quaternary walk game is one quarter more than the admission fee of the binary walk game. The stopping rule and the reward amount remain the same as before. We still ask the same question: How should the player determine (c, k) to maximize her expected net reward?"

More specifically, the player begins at Node 0. She tosses a copper coin and a silver coin simultaneously. The outcome of the copper coin determines whether to move clockwise (if heads) or counterclockwise (if tails). The outcome of the silver coin determines whether she will move one step (if tails) or two steps (if heads) in the direction determined by the copper coin. Whenever she lands at a new node for the first time she is said to have 'captured' it. She does not capture the node she skips over. If a player intends to capture c nodes and she wishes to toss the coins k times, then she must pay an admission fee of $(25 + 25c + k)$ cents (one quarter to avail the quaternary walk game, one quarter per node she wishes to capture and one penny per toss or the pair of coins). At the random time T, when either c nodes (other than Node 0) are captured or k tosses are over, whichever event happens earlier, the game stops; and the player earns as many nickels as the sum of the labels of the captured nodes.

Here's how the game is played: The player begins at Node 0; and after each toss of the copper and the silver coins, the player moves one step clockwise if the outcome is HT, two steps clockwise if HH, or one step counterclockwise if the outcome is TT or two steps counterclockwise if TH; that is, each of the four moves happens with probability 1/4. She is said to have captured a node on the first visit to it. She does not visit or capture the node

she skips over. When the c nodes (other than Node 0) are captured or when the k tosses are over, whichever event happens earlier, the player must stop. Here is a simple way to think about the stopping time T: Toss the coin k times; let Z_t denote the number of nodes captured (other than Node 0) after t tosses (for $t = 1, 2, \ldots, k$). Either the game stops after k tosses, if fewer than c nodes are captured; or it stops as soon as c nodes are captured. That is,

$$T = \begin{cases} \min\{1 \le t \le k : Z_t = c\}, & \text{if } Z_k \ge c; \\ k, & \text{if } Z_k < c. \end{cases} \tag{1}$$

Let S_T denote the random set of nodes visited and captured by the random walk on the parity dial when the game ends. The player will earn as many nickels N as the sum of the labels in S_T. Thus, her reward is $N = \sum_{i \in S_T} l_i$ nickels or $5N$ cents, where l_i is the label of Node i. How should the player choose (c, k) to maximize her expected net reward?

3. Analyzing the (c, k) Games for $c = 1, 2, 3$

In this section we study the (c, k) games for $c = 1, 2, 3$, using the exact probability mass function (PMF) of N, the number of nickels earned when the game ends.

3.1 The $(1, 1)$ game

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!

3.2 Prospects of the $(2, k)$ games

How about choosing $c = 2$? Surely, in this case $k \ge 2$, since with only one toss, it is not possible to capture two nodes. But with only two tosses, there is 3/4 chance of capturing two distinct nodes and a pretty high chance of 1/4 that the player will return to 0 after capturing just one node. With three tosses there is a 1/16 or 6.25% chance of revisiting the already captured node and earn no additional reward. Consider a simple-minded player, Amber, who is contemplating tossing $k = 4$ tosses. She reasons as follows:

> "I will toss the coin $k = 4$ times. There is a very small chance (less than 2%) that I will capture only one node from $\{1, 2, 10, 11\}$. With a high chance I will capture two nodes out of $\{1, 2, 3, 4, 8, 9, 10, 11\}$. Since this set has an average of 6, the two nodes I will capture are worth 12 nickels or 60 cents, on average. Since I have to pay $25 + 25(2) + 4 = 79$ cents, my loss will be about 24.05%. The game is highly unfair! I won't play."

Later that day Amber wondered: "Why is the average sum of all possible pairs 12?" She listed the $\binom{8}{2} = 28$ pairs, computed the within-pair sums and averaged them. Indeed, the average turned out to be 12. (We encourage the reader to verify the same.) However, Amber did not stop there. As she stared at the list, all at once it dawned on her that not all possible pairs of nodes are admissible: In fact, 16 pairs are inadmissible and only 12 pairs

are admissible. More precisely, with Node 8 we can pair up only Node 10; with Node 4 we can pair up only Nodes 2 and 10. Similarly, with Node 9 we can pair up only Node 11; with Node 3 we can pair up only Nodes 1 and 11. Finally, we can have all 6 pairs from among nodes $\{1, 2, 10, 11\}$. Therefore, when we capture two nodes, the total number of nickels we will earn are

$$8+10, 4+2, 4+10;\ 9+11, 3+1, 3+11;\ 1+2, 1+10, 1+11, 2+10, 2+11, 10+11.$$

Thus the sums (after sorting) are $3, 4, 6, 11, 12, 12, 13, 14, 14, 18, 20, 21$. Hence, the average earning is $12 + 1/3$ nickels, or $61 + 2/3$ cents; and the player's loss is about 21.94%. Amber was fascinated with her finding. "What can I do with my discovery?" she asked herself while tossing and turning that night.

> Next day, Amber went to the casino and told the manager: "The $(c = 2, k = 4)$ game allows me to earn 12 nickels on average. So it will be a fair game if you charge 60 cents, instead of 79 cents."
>
> The manager said: "Look, we don't let players dictate games. But I will make an exception for you, Amber, and only for today. Tell you what? I will even give you one free toss. Go ahead, and play the $(c = 2, k = 5)$ game on payment of 60 cents."
>
> Amber was ecstatic! She said to herself: "This is my lucky day! I can earn about 5/3 of a penny per game; or about \$1.66 per 100 games."

Amber jumped to action. However, after playing 100 times, she lost about 5 dollars! What went wrong? Did the casino tamper with the random walk? Amber quit the game; and visited her statistician friend, Staci, for an explanation.

Staci explained that Amber was correct in thinking she will capture two nodes with a high probability. In fact, with 4 tosses the chance of not capturing two nodes is only 1/64; and with 5 tosses it is 1/256. She was also correct in identifying the admissible pairs, whether she tossed 4 times or 5. However, she had blundered in assuming that all admissible pairs are equally likely. They are not! To demystify the reason for her loss, Staci must help Amber understand the exact probability distribution of N, the number of nickels captured.

We already noted that when the copper and the silver coins are tossed for the first time, then N is equally likely to be in $\{10, 2, 1, 11\}$. If the coins are tossed twice, then enumerating all $4^2 = 16$ possible outcomes we see that N takes values

$$18, 14, 12, 10;\quad 6, 12, 2, 3;\quad 3, 1, 12, 4;\quad 11, 12, 14, 20.$$

For $k \geq 3$ tosses, manually enumerating all 4^k outcomes becomes tedious. However, one can write a small program (see the Appendix for a code in our favorite software R) to do the job efficiently, and tabulate the values of N in Table 1. Some theoretical properties of the frequencies in Table 1 are discussed in Section 5. Table 1 shows that the possible values of N are not equally likely, as Amber was prone to assume.

Based on the exact distribution of N, the number of nickels earned, under the $(2, k)$ game, we can compute the expected loss under the publicly available admission price of $75 + k$ cents. This is shown in Table 2. When $c = 2$, the optimal number of tosses where

Table 1: The distribution of number of nickels earned when $c = 2$ and $k \geq 1$

	deno-	N nickels												
k	minator	1	2	3	4	6	10	11	12	13	14	18	20	21
1	4	1	1	0	0	0	1	1	0	0	0	0	0	0
2	16	1	1	2	1	1	1	1	4	0	2	1	1	0
3	64	1	1	10	4	4	1	3	20	2	8	4	4	2
4	256	1	1	42	17	17	1	9	84	8	34	17	17	8
5	1024	1	1	170	68	68	1	35	340	34	136	68	68	34
6	4096	1	1	682	273	273	1	137	1364	136	546	273	273	136

the percentage loss is minimized is $k_* = 4$. Moreover, in Section 5 we will show that as $k \to \infty$, the expected reward increases monotonically (but at a progressively slower rate), until it approaches an asymptotic value of $55 + 1/3$ cents, but the price keeps on increasing linearly. Hence, as k increases, the percentage loss initially decreases steadily (though at a progressively slower rate), and later it monotonically increases until it approaches the asymptotic value of one. As we shall see in Section 4, this property holds for all $c \geq 2$.

Table 2: Expected loss when $c = 2$ and $k \geq 1$

	nickels		cents			
k	mean	SD	E[Reward]	E[Loss]	Price	E[% Loss]
1	6.00	5.52	30.000	46.000	76	60.52
2	9.63	5.81	48.125	28.875	77	37.50
3	10.75	5.54	53.75	24.25	78	31.09
4	10.98	5.44	54.90	24.10	79	30.51
5	11.05	5.42	55.25	24.75	80	30.94
6	11.06	5.41	55.30	25.70	81	31.73
15	11.07	5.40	55.33	34.66	90	38.51

When the casino manager offered the $(c = 2, k = 5)$ game to Amber for an admission fee of 60 cents, he knew quite well that Amber's expected loss will be $60 - 5(11.047) = 4.77$ cents, or 7.95%. Hence, in view of the central limit theorem [see Dudewicz and Mishra (1988), for example], after playing the $(c = 2, k = 5)$ game 100 times, Amber should have expected an approximately normally distributed net loss with a mean of 4.77 dollars and a standard deviation of 0.542 dollars. Amber's actual experience seems to be less than half a standard deviation below the expected value. There is no reason to suspect any foul play on part of the casino. With Staci's expert guidance and some self-study using Wikipedia (2020), Amber learned a whole lot about the central limit theorem. (Readers will act wisely to do the same.)

Let us return to the publicly available $(2, k)$ game with an admission fee of $75+k$ cents. Regarding the optimality of $k = 4$, we have two additional messages for our simple-minded gambler friend Amber.

First, we should explain to her that $k = 4$ is better than $k = 5$. If after 4 tosses she already captures two nodes, she cannot use her fifth toss at all. The only time she can make

use of the fifth toss, is when she captures only one node after four tosses. This means after four tosses she has earned 1, 2, 10, or 11 nickels with probability 4^{-4} each, and returned to Node 0. Using the fifth toss, she can capture a new node with an average node label of $23/4, 22/4, 14/4, 13/4$ respectively for the above four cases. Thus, over and above what she has earned with four tosses, the additional expected earning with the fifth toss is only $4^{-4} \times (23 + 22 + 14 + 13)/4 = 0.07032$ nickels, or 0.3516 cents. This is exactly the amount the casino manager had offered Amber when permitting a free fifth toss. Why should any other player (and Amber on any other day) pay an extra penny at the beginning of the game knowing that on average they will earn about one-third of a penny more?

Second, consider a make-belief scenario to convince Amber why she should not pay for any more than 4 tosses. Suppose that after 4 tosses Amber has captured only one node (and returned to Node 0); and she has earned 1, 2, 10, or 11 nickels, which events happen with probability 4^{-4} each. Suppose also that the casino very generously offers her *at no cost* an *unlimited* number of tosses until she captures a new node! Then Amber is expected to earn an additional $11.07 - 10.98 = .09$ nickels, or 0.45 cents. See a more detailed reason in Section 5. If Amber had to pay even one penny more for these infinitely-many tosses she would certainly lose even more than 7.95% of her wager. If this make-belief scenario is too incredible to be true, we can transform it into a more realistic scenario: At the outset when Amber agrees to pay 60 cents, the casino makes this offer: "Should you fail to capture two nodes with your 4 tosses, we will let you toss an unlimited number of times (until you capture a second node) if you will pay just one penny more right now." However, we have already reasoned that accepting this offer is more disadvantageous to the player than to simply toss 4 times. For there is a high chance that she will forfeit her unlimited number of tosses anyways!

When our gambler friend Amber learns all these truth, having chosen $c = 2$, she should pay for exactly 4 tosses and be prepared to lose roughly 8% of her wager. On any other day, her admission price will be 79 cents, just like for any other player. But will she have the appetite to lose 30.5% per play? Although choosing $(c = 2, k = 4)$ is surely better than choosing $(c = 1, k = 1)$, which had an expected loss of 41.2%, it is not an attractive offer to a gambler. Games that are so unattractive to the gambler are not conducive to the casino's business prospect either. The gambler must inspect other choices.

3.3 Prospects of the $(3, k)$ games

Amber, slightly more enlightened by now, continues to investigate other alternative choices of (c, k) games. She has learned that it is not enough to simply list the possible number of nickels she will earn. It is important to know the associated probabilities also. Correcting her flawed logic in case of $c = 2$ and paying heed to our above messages, Amber might reason as follows:

> "For $c = 1$, I choose $k = 1$ toss; I pay 51 cents, earn 30 cents on average and lose about 41%. For $c = 2$, I just learned that I should choose $k = 4$ tosses, or pay a total of 79 cents. I expect to earn 55 cents on average; and so I lose about 30%. Perhaps for $c = 3$, I ought to choose $k = 3^2 = 9$ tosses, or pay \$1.09. But how can I calculate the actual probability distribution of the reward? I know, I will imitate the R codes for

the $(c = 2, k = 4)$ game that Staci gave me and write the codes for the $(c = 3, k = 9)$ game."

Amber modified our R codes for the $(c = 2, k = 4)$ game and constructed the exact distribution of N for the $(c = 3, k = 9)$ game involving $4^9 = 262144$ possible sequences of outcomes shown in Table 3. With PMF so constructed, Amber computed the mean and the SD of the expected reward. (We urge the reader to do the same.)

Table 3: Exact reward distribution, mean and SD for the $(c = 3, k = 9)$ game

	N nickels									
value	1	2	3	4	5	6	7	8	9	10
freq	1	1	10	15	0	10223	10208	0	4680	1
value	11	12	13	14	15	16	17	18	19	20
freq	31	5700	33694	35249	31331	31331	1555	15	6235	14903
value	21	22	23	24	25	26	27	28	29	30
freq	10238	21419	19864	6235	10915	0	4680	0	1555	1555

mean(N)=16.68917 SD(N)=5.285686 Total frequency=$4^9 = 262144$

After hanging out with her statistician friend, Staci, long enough, Amber has learned to ask some critical questions. This is what she asked next:

"Why am I using $k = 9$ tosses? Am I being duped to thinking $k = c^2$ because the formula held true for $c = 1, 2$? But the optimal k may be different. If I choose a higher k, computing the PMF is time consuming. But if I choose a smaller k, I can compute the PMF quickly. In this latter case, although I may earn less, I will also pay less. Maybe I will reduce my percentage loss! Let me try various values of $5 \leq k \leq 8$ for $c = 3$."

What she found is documented in Table 4.

Table 4: Expected loss when $c = 3$ and $5 \leq k \leq 9$

	nickels		cents			
k	mean(N)	SD(N)	E[Reward]	E[Loss]	Price	E[% Loss]
5	15.99121	5.698083	79.96	25.04	105	23.85
6	16.36841	5.499275	81.84	24.16	106	22.79
7	16.56256	5.376299	82.81	24.19	107	22.60
8	16.64687	5.317425	83.23	24.77	108	22.93
9	16.68917	5.285686	83.45	25.55	109	23.44

Thus, Amber discovered that for $c = 3$, the optimal choice for k (that minimizes the percentage loss) is not 3^2, rather it is $k = 7$. For the $(c = 3, k = 7)$ game, we show the exact PMF in Table 5. In particular, the mean reward is 82.81 cents with a SD of 26.88 cents. Moreover, she wins (earns over \$1.09) with probability $P\{N \geq 22\} = .2487$; she never wins

more that half-a-dollar; but she loses over half-a-dollar with probability $P\{N \leq 11\} = .1031$. The player wins over a quarter with probability only $P\{N \geq 27\} = .0294$; but she loses over a quarter with a high probability of $P\{N \leq 16\} = .6266$.

Table 5: Exact reward distribution, mean and SD for the $(c = 3, k = 7)$ game

	N nickels									
value	1	2	3	4	5	6	7	8	9	10
freq	1	1	126	7	0	627	620	0	292	1
value	11	12	13	14	15	16	17	18	19	20
freq	15	544	2050	2145	1919	1919	95	7	387	919
value	21	22	23	24	25	26	27	28	29	30
freq	634	1311	1216	387	679	0	292	0	95	95

mean(N)=16.56256 SD(N)=5.376299 Total freq. $=4^7 = 16384$

4. The Best Choice of (c, k) for $c \geq 4$

Amber has learned the inevitable truth that the reward random walk game on the parity dial is unfavorable to her if she chooses $c = 1, 2, 3$. Perhaps she has already resigned to accepting that every (c, k) game will be unfavorable to her. Consequently, she is willing to tolerate a 10% loss per play, in exchange for the entertainment and thrill she experiences during the game. What if she chooses $c \geq 4$? Amber dived deeper into her thoughts.

> "I see my percentage loss keeps on reducing as I try higher values of c. I must try other values $4 \leq c \leq 11$. Perhaps I can reduce my loss further; maybe I can even earn a positive return on investment! I won't bet on it though; I will be happy if my loss is under 10%. Nonetheless, for each c, finding the corresponding optimal k requires computing the exact PMF of the number of nickels captured in the (c, k) game. However, such computations become exceedingly time consuming as $k \geq 10$ becomes large, since 4^k increases exponentially. What can I do? I suppose I need some help from superwoman. Let me visit Staci once more."

And so she did. Incredibly, Staci had another trick up her sleeve. She said:

> "Amber, you do not need to know the exact PMF. You only need to know the mean (and perhaps the SD) of the reward you will earn. There is a law of large numbers [see Dudewicz and Mishra (1988), for example] that says: "If from any distribution (with a finite expectation), you take many, many (independent) observations, then the long-run sample mean will be close to the expected value of the distribution." So simply replicate the game many times (say, 10^{2m} times); and then compute the mean and the SD of the rewards earned in these plays. That mean will approximate the long run expected reward, with the precision of approximation given by the standard error (SE), which equals SD/10^m. You can learn more about this law of large numbers by reading Wikipedia (2020)."

Amber was so overjoyed to learn about this wonderful trick that she forgot to ask Staci anything about the optimal choice of k. Left to her own devices, she reasoned: "For $c = 2$ and 3, the corresponding optimal k were respectively 4 and 7. I think $k = 1+c(c+1)/2$ may hold true, in general." Amber modified the codes again to play the game 10^4 times, where each game consists of tossing a pair of fair copper and silver coins (or equivalently, choosing from a discrete uniform random variable which takes values $-2, -1, 1, 2$ equally likely) until either c nodes are captured or $k = 1 + c(c + 1)/2$ tosses are over, whichever event happens first. See the codes for simulation in the Appendix. The average reward and the SD of the rewards earned in these repeated games suffice to approximate the expected loss and the associated SE. Thereafter, it is easy to calculate the expected percentage loss as a fraction of the admission fee. The results of her simulation study are summarized in Table 6.

Table 6: Expected reward and loss to capture $c \geq 3$ nodes in $k = 1 + c(c + 1)/2$ tosses, via simulation based on 10^4 iterations

		cents				
c	k	Price	E[Reward]	SE	E[Loss]	E[% Loss]
2	4	79	54.97	0.27	24.03	30.42
3	7	107	82.80	0.27	24.20	22.61
4	11	136	112.90	0.26	23.10	16.99
5	16	166	143.63	0.26	22.37	13.48
6	22	197	174.83	0.25	22.17	11.25
7	29	229	205.74	0.25	23.26	10.16
8	37	262	236.93	0.24	25.07	†9.57
9	46	296	267.74	0.22	28.24	†9.55
10	56	331	297.44	0.18	33.56	10.14
11	67	367	326.01	0.14	40.99	11.17

Amber concluded that all (c, k) games are unfavorable to her. Only two games were within her tolerance limit of 10% loss—$(8, 37), (9, 46)$—with the latter being slightly preferable. Are Amber's above reasoning justified?

Amazingly, our friend Amber has reasoned very wisely. We applaud her quick understanding of the law of large numbers and her smart implementation of the simulation. Nonetheless, she could have done a little better: Corresponding to each c, instead of relying on her conjecture $k = 1 + c(c + 1)/2$, she should have searched for the optimal k, again via a more thorough simulation study. Then she could discover the best available choice.

While we could simulate the game for all values of $k \geq c$, we follow a *smart search algorithm*. For $c = 2$, we already know the optimal choice is $k = 4$. For $c = 3$, we successively tried $k = 5, 6, 7, 8, 9$. Since the optimal value turns out to be $k = 7$, for the next choice $c = 4$, we should successively try $k \geq 8$. When the optimal value for k is found (by continuing as long as expected percentage loss decreases, and as soon as it begins to increase, by trying out one more value of k to verify that the increasing trend continues), we stop the search. Then we repeat the process for the next value of c starting with the value of k greater than the optimal value for the previous c. In Table 7, we report the performance of the optimal k for each $c \geq 2$, together with the performance of two values of k below and and two values of k above the optimal as demonstration. In summary, the optimal values are

$k_*(c) = (1, 4, 7, 10, 14, 18, 22, 27, 32, 39, 48)$ for $c = 1, 2, \ldots, 11$.

Table 7: Expected reward and expected loss for $2 \le c \le 11$ and associated optimal k_* together with two values below and two values above it

		cents						cents			
c	k	price	E[rew]	E[loss]	E[%loss]	c	k	price	E[rew]	E[loss]	E[%loss]
	2	77	48.10	28.90	37.53		5	105	79.88	25.12	23.92
	3	78	53.73	24.27	31.12		6	106	81.86	24.14	22.77
2	4	79	54.88	24.12	*30.53	3	7	107	82.79	24.21	*22.62
	5	80	55.26	24.74	30.92		8	108	83.26	24.74	22.91
	6	81	55.31	25.69	31.71		9	109	83.42	25.58	23.47
	8	133	109.37	23.63	17.77		12	162	140.13	21.87	13.50
	9	134	111.10	22.90	17.09		13	163	141.40	21.60	13.25
4	10	135	112.18	22.82	*16.90	5	14	164	142.34	21.66	*13.21
	11	136	112.79	23.21	17.07		15	165	142.99	22.01	13.34
	12	137	113.22	23.78	17.36		16	166	143.46	22.54	13.58
	16	191	170.29	20.71	10.84		20	220	199.89	20.11	9.14
	17	192	171.47	20.53	10.69		21	221	201.15	19.85	8.98
6	18	193	172.43	20.57	*10.66	7	22	222	202.10	19.90	*8.96
	19	194	173.19	20.81	10.73		23	223	202.96	20.04	8.98
	20	195	173.79	20.21	10.88		24	224	203.70	20.30	9.06
	25	250	229.90	20.10	8.04		30	280	258.78	21.22	7.58
	26	251	230.98	20.02	7.98		31	281	259.90	21.10	7.51
8	27	252	232.01	19.99	*7.93	†9	32	282	260.96	21.04	†*7.46
	28	253	232.90	20.10	7.94		33	283	261.85	21.15	7.47
	29	254	233.70	20.30	7.99		34	284	262.66	21.34	7.51
	37	312	288.02	23.98	7.69		46	346	314.63	31.37	9.07
	38	313	288.98	24.02	7.67		47	347	315.56	31.44	9.06
10	39	314	289.98	24.02	*7.65	11	48	348	316.51	31.49	*9.05
	40	315	290.82	24.18	7.67		49	349	317.40	31.60	9.05
	41	316	291.66	24.34	7.70		50	350	318.15	31.85	9.10

Based on the results of Table 7, we learn that Amber's conjecture for the optimal k_*, as a function of $c \ge 4$, was wrong. Moreover, we learn that the best choice game is $(c = 9, k = 32)$; and with this choice, a gambler faces a 7.47% expected loss (instead of a 9.55% loss as Amber had anticipated based on Table 6 where she misjudged k to be 46 when $c = 9$). Amazingly, Amber was right in choosing $c = 9$. But she was acting suboptimally by spending $46 - 32 = 14$ cents more to increase her expected reward by only $267.71 - 260.96 = 6.75$ cents. Having discovered the optimal choice, the gambler must stick to playing only the $(c = 9, k = 32)$ game, for any other game would cause her to lose a higher percentage of her wager.

Our increasingly wiser, inquisitive friend Amber cannot stop asking more questions. Here is a sample of questions she asked. (Readers will do well to ask more questions.)

(1) "How many nodes will I actually capture (when I set out to capture 9 nodes)?"

(2) "How many times will I actually toss (for more tosses mean more entertainment)? Equivalently, what is the distribution of the stopping time T defined in Eq. (1)?"

(3) "How much reward will I collect?"

The answers to Amber's questions are not numbers, rather they are random variables that can be described by their PMFs. These PMFs need not be exact; it suffices to estimate them based on simulation. These are reported in Table 8 and Figures 2 and 3 based on simulation involving 10^6 iterations.

Table 8: Frequencies of the number of nodes the gambler captured when playing the $(c = 9, k = 32)$ game 10^6 times

c	k	3	4	5	6	7	8	9	total
9	32	2	245	3590	18864	49448	91004	836847	10^6

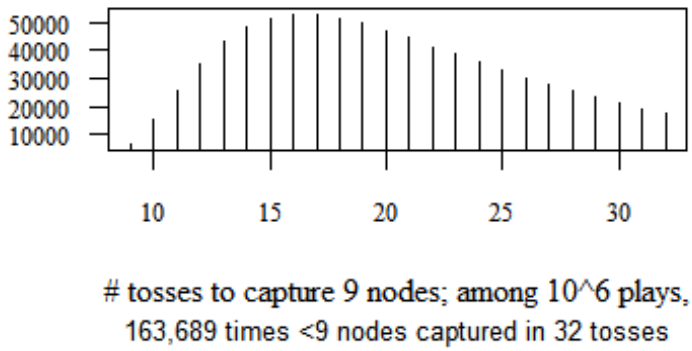


Figure 2: When playing the $(c = 9, k = 32)$ game 10^6 times, the stopping time T has a mean 21.98, SD 7.20, three quartiles 16, 21, 28. Moreover, about 16.4% of times fewer than 9 nodes are captured in all 32 tosses.

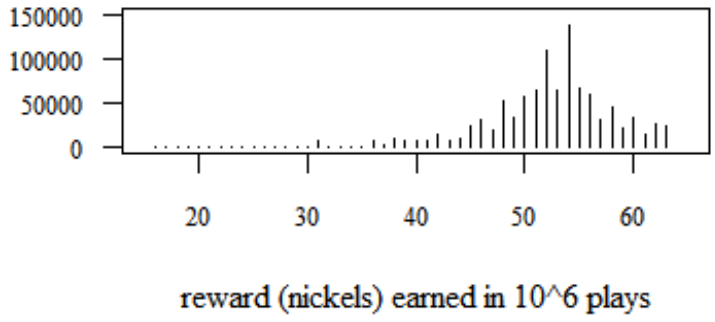


Figure 3: When playing the $(c = 9, k = 32)$ game 10^6 times, the number of nickels the gambler earned has a mean 52.19, SD 5.98 and three quartiles 50, 53, 56.

As we satisfy Amber's curiosity, additional features of the optimal game $(c = 9, k = 32)$ are revealed. For this game, the gambler has about 83.7% chance of capturing all 9 nodes she intended to capture and an additional 9.1% chance of capturing 8 nodes. Thus, there is a 7.2% chance of capturing 7 or fewer nodes when all 32 tosses are over. The mean number

of tosses until 9 nodes are captured (or 32 tosses are over) is 21.98, with a standard deviation of 7.20. About 82.0% of times the gambler will have to forgo one or more tosses (for she has already captured 9 nodes), and 16.3% of times she will wish she could toss more often (for she has not captured 9 nodes with 32 tosses). Also, for the same game, while she wagers 282 cents, she has a 19.7% chance of coming out a winner with 57 nickels or more (that is, earning more than her wager). She has a 15% chance of losing over half-a-dollar and a 2.2% chance of losing over a dollar.

5. Theoretical Results

For want of space we refrain from studying at length the manifold theoretical issues. We illustrate only a few theoretical queries to whet the readers' appetite; but we leave many other interesting questions for the readers to pursue on their own.

5.1 Theory when $c = 2$

In Subsection 3.2, we learned that for $c = 2$, the optimal number of tosses is $k = 4$, with a 30.51% expected percentage loss. In particular, this means that the additional expected reward earned by the fifth toss over and above that earned in the first four tosses must be less than 0.695 cents. For otherwise, paying for five tosses would be better than paying for four. In fact, this additional expected reward was shown in Table 2 to be $55.25 - 54.90 = 0.35$. The reason was explained to the gambler as a first message regarding the optimality of $k = 4$. Here we continue the argument by computing the additional expected reward earned in each successive toss, shown in Table 9. There we see that between the fourth and the fifth tosses the exact additional expected reward is $72/4^5 = 9/128$ nickels, or 0.3516 cents. Moreover, the additional expected reward earned by an infinite number of tosses after the first four tosses until two nodes are captured is

$$\left(\frac{72}{4^5} + \frac{58}{4^6}\right) + \left(\frac{72}{4^7} + \frac{58}{4^8}\right) + \ldots = \left(\frac{144}{2^{11}} + \frac{29}{2^{11}}\right)\ \left[1 + \frac{1}{16} + \frac{1}{16^2} + \ldots\right] = \frac{173}{2^{11}}\ \frac{16}{15} = 0.0901$$

nickels, or 0.4505 cents.

Here is another way to derive the expected reward until the gambler captures two nodes (using an unlimited number of tosses). We consider two mutually exclusive, exhaustive cases.

Case 1: With the first two tosses she captures two distinct nodes (with probability 3/4), and the number of nickels she earns on average is

$$\frac{1}{4}\left[\left(1 + \frac{16}{4}\right) + \left(2 + \frac{15}{4}\right) + \left(10 + \frac{14}{4}\right) + \left(11 + \frac{13}{4}\right)\right] = \frac{1}{4}\left[24 + \frac{58}{4}\right] = 9 + \frac{5}{8}. \tag{2}$$

Case 2: With the remaining probability 1/4, the gambler does not capture two distinct nodes with the first two tosses. In fact, with probability 1/16, she captures Node 1 and then returned to Node 0. Thereafter, how many nickels will she earn on average, if she is allowed an unlimited number of additional tosses until she captures a second node? Call this average number of additional nickels μ_1. Then by conditioning on the next two moves, we see that μ_1 satisfies the following recursive relation

$$\mu_1 = \frac{10 + 2 + 11}{4} + \frac{1}{4}\left[\frac{2 + 11 + 3 + \mu_1}{4}\right].$$

Table 9: Additional reward earned in each successive toss when $c = 2$

k	nickels					cents
1	10	2	1	11	$24/4 = 6$	30
	8	4	2	0		
	4	10	0	1		
	2	0	11	3		
	0	1	3	9		
2	14	15	16	13	$58/4^2 = 29/8$	145/8
	0	10	10	10		
	2	0	2	2		
	1	1	0	1		
	11	11	11	0		
3	14	22	23	13	$72/4^3 = 9/8$	45/8
	8	4	2	0		
	4	10	0	1		
	2	0	11	3		
	0	1	3	9		
4	14	15	16	13	$58/4^4$	145/128
	0	10	10	10		
	2	0	2	2		
	1	1	0	1		
	11	11	11	0		
5	14	22	23	13	$72/4^5$	45/128
6	14	15	16	13	$58/4^6$	145/1024
7	14	22	23	13	$72/4^7$	45/1024
⋮					⋮	⋮
sum					$6 + \frac{29+9}{8} * \frac{16}{15} = 11\frac{1}{15}$	$55\frac{1}{3}$

Solving the above recursive relation, we obtain $\mu_1 = 108/15$ nickels. Likewise, with probability 1/16, she captures Node 2 and then returns to Node 0; and thereafter, she will eventually earn on average $\mu_2 = 103/15$ nickels. With probability 1/16, she captures Node 10 and then returns to Node 0; thereafter, she will eventually earn on average $\mu_{10} = 70/15$ nickels. Finally, with probability 1/16, she captures Node 11 and then returns to Node 0; thereafter, she will eventually earn on average $\mu_{11} = 65/15$ nickels. Thus, in Case 2, the number of nickels the gambler earns on average is

$$\frac{1}{16}\left[\mu_1 + \mu_2 + \mu_{10} + \mu_{11}\right] = \frac{1}{16}\left[\frac{108 + 103 + 70 + 65}{15}\right] = \frac{173}{120}. \tag{3}$$

Adding Eq. (2) and Eq. (3) together, on average the gambler earns

$$9 + \frac{5}{8} + \frac{173}{120} = 9 + \frac{248}{120} = 11\frac{1}{15}$$

nickels, or $55\frac{1}{3}$ cents.

5.1.1 Formula for the PMF of N for the $(c = 2, k)$ game

In Table 1, we documented the distribution of N, the number of nickels earned, for $1 \le k \le 6$ using a computer software code. Here we shall discover a pattern among these frequencies and hence write down the formulas in general, so that we can construct the exact distribution of N for larger values of k without having to use the codes. First, from rows corresponding to odd $k = 2i - 1 \ge 3$, we subtract Row 1; and from rows corresponding to even $k = 2i \ge 4$, we subtract Row 2. See Table 10.

Table 10: To discover patterns in the frequencies of N for various $k \ge 1$, subtract the first row from all odd rows 3 or higher and the second row from all even rows 4 or higher.

	N nickels													
k	1	2	3	4	6	10	11	12	13	14	18	20	21	
1	1	1	0	0	0	1	1	0	0	0	0	0	0	
2	1	1	2	1	1	1	1	4	0	2	1	1	0	
3	0	0	10	4	4	0	2	20	2	8	4	4	2	
4	0	0	40	16	16	0	8	80	8	32	16	16	8	
5	0	0	170	68	68	0	34	340	34	134	68	68	34	
6	0	0	680	272	272	0	136	1360	136	544	272	272	136	
⋮							⋮							
⋮							⋮							
$2i-1$	(0	0	5	2	2	0	1	10	1	4	2	2	1	$)\times f_{2i-1}$
$2i$	(0	0	5	2	2	0	1	10	1	4	2	2	1	$)\times f_{2i}$

Then we divide each row 3 or higher by the entry f_i in the last column (under $N = 21$) to see that the quotient vector is a constant! It remains to find a formula for f_i, for then we can reverse the steps (multiply the constant quotient vector by f_i, and add Row 1 or Row 2 according as i is odd or even) to reconstruct all frequencies in each row.

The entry in the last column in any even position is four times the entry in the immediately previous odd position; that is, $f_{2i} = 4f_{2i-1}$. The entries in the odd positions are: $(0, 2, 34, 546, \ldots)$, which satisfy the recursive relation $f_{2i+1} = 16\, f_{2i-1} + 2$, and hence the formula

$$f_{2i-1} = \frac{2}{15}(16^{i-1} - 1). \tag{4}$$

Thereafter, using Eq. (4), we can obtain the limiting probabilities as

$$P\{N = 21 | k = 2i - 1\} = \frac{f_{2i-1}}{4^{2i-1}} = \frac{2}{15}\frac{16^{i-1} - 1}{4^{2i-1}} \to \frac{2}{15}16^{-1/2} = \frac{1}{30};$$

and

$$P\{N = 21 | k = 2i\} = \frac{f_{2i}}{4^{2i}} = \frac{4f_{2i-1}}{4\,4^{2i-1}} = \frac{f_{2i-1}}{4^{2i-1}} \ldots \to \frac{2}{15}16^{-1/2} = \frac{1}{30}.$$

In particular, as $k \to \infty$, either through even values or through odd values, in the limit the number N of nickels earned takes on values $(3, 4, 6, 11, 12, 13, 14, 18, 20, 21)$ with

associated probabilities $(5, 2, 2, 1, 10, 1, 4, 2, 2, 1)/30$. Hence, the limiting mean and SD of N are respectively $11 + 1/15$ and 5.41. Equivalently, the expected reward is $55 + 1/3$ cents with a SD of 27.05 cents.

5.2 Theory when $c = 3$

In Subsection 3.3, we learned that for $c = 3$, the optimal number of tosses is $k = 7$, with a 22.60% expected percentage loss. In particular, this means that the additional expected reward earned by the eighth toss over and above that earned in the first 7 tosses must be less than 0.774 cents. In fact, this additional expected reward was shown in Table 2 to be $83.23 - 82.81 = 0.42$. Likewise, the additional expected reward earned by an infinite number of tosses after the first 7 tosses until three nodes are captured is about $83.64 - 82.81 = 0.83 \pm 0.0263 > 0.774$, obtained by simulating the $(c = 3, k = 20)$ game 10^6 times. This means that to a player who agrees to play the $(c = 3, k = 7)$ game on payment of \$1.07, if the casino offers an unlimited number of tosses until three nodes are captured on payment of just one penny more, then the gambler should take it. But if the charge is two pennies or higher, the gambler should decline the offer.

The exact probability distribution of N, the number of nickels earned until three nodes are captured or k tosses are over, are documented in Table 11 using a modified code along the lines of that used to construct Table 1.

As we demonstrated in Subsection 5.1 for the case of $c = 2$, we now invite the reader to find a formula for the frequencies in Table 11. If the game must stop as soon as the player captures three nodes (using as many tosses as needed), then we conjecture that the PMF of the number of nickels the player will earn is as given in Table 12. If our conjecture holds, then the player will earn on average 16.79 nickels (SD 5.30), or 83.95 cents (SD 26.48 cents).

5.3 Theory when $c = 9$

To a player willing to play the overall optimal game $(c = 9, k = 32)$ with an admission fee of \$2.82 at a 7.46% expected loss, if the casino offers an unlimited number of tosses until nine nodes are captured on payment of 8 cents or less, then the gambler should take it. But if the charge is 9 cents or higher, then the gambler should decline the offer. How did we discover this threshold? We simply estimated the expected rewards of the $(c = 9, k = 100)$ game and the $(c = 9, k = 32)$ game via simulation based on 10^6 plays of each game, and then we computed their difference $269.19 - 260.96 = 8.23$ cents with a SD of .04 cents. It sufficed to consider $k = 100$ because in all but 108 cases of the 10^6 plays of the $(c = 9, k = 100)$ game, all 9 nodes were captured. Among the other 108 cases, 99 times 8 nodes are captured and the remaining 7 times 7 nodes are captured.

6. Binary Vs. Quaternary Random Walks

In this paper, we have studied a game that allows a symmetric quaternary random walk on the parity dial. How does this game compare with the original game of a symmetric binary random walk on the parity dial studied by Sarkar (2020 a)? In Table 13, we summarize the expected performance of the (c, k_*) game for $1 \le c \le 11$ and the associated optimal number of tosses k_* side by side for the two types of random walks.

Table 11: Frequencies of N for $c = 3$ and $1 \le k \le 9$

row	N\k-> 1	2	3	4	5	6	7	8	9
[1,]	1	1	1	1	1	1	1	1	1
[2,]	2	1	1	1	1	1	1	1	1
[3,]	3	0	2	6	14	30	62	126	254	510
[4,]	4	0	1	1	3	3	7	7	15	15
[5,]	6	0	1	2	10	37	156	627	2544	10223
[6,]	7	0	0	1	7	34	149	620	2529	10208
[7,]	9	0	0	1	4	18	72	292	1168	4680
[8,]	10	1	1	1	1	1	1	1	1	1
[9,]	11	1	1	3	3	7	7	15	15	31
[10,]	12	0	4	13	32	78	196	544	1676	5700
[11,]	13	0	0	5	24	116	492	2050	8340	33694
[12,]	14	0	2	5	29	121	523	2145	8743	35249
[13,]	15	0	0	4	23	109	465	1919	7781	31331
[14,]	16	0	0	4	23	109	465	1919	7781	31331
[15,]	17	0	0	0	1	5	23	95	387	1555
[16,]	18	0	1	1	3	3	7	7	15	15
[17,]	19	0	0	1	5	23	95	387	1555	6235
[18,]	20	0	1	3	14	55	228	919	3712	14903
[19,]	21	0	0	3	9	40	155	634	2543	10238
[20,]	22	0	0	2	15	73	317	1311	5321	21419
[21,]	23	0	0	2	14	68	294	1216	4934	19864
[22,]	24	0	0	1	5	23	95	387	1555	6235
[23,]	25	0	0	2	9	41	167	679	2723	10915
[24,]	27	0	0	1	4	18	72	292	1168	4680
[25,]	29	0	0	0	1	5	23	95	387	1555
[26,]	30	0	0	0	1	5	23	95	387	1555
[All]	sum	4	4^2	4^3	4^4	4^5	4^6	4^7	4^8	4^9

Table 12: A conjecture regarding the limiting distribution, mean and SD of N, the number of nickels earned, for the $(c = 3, k = \infty)$ game

value	6	7	9	13	14	15	16	17	19
freq	7	7	3	22	23	20	20	1	4
value	20	21	22	23	24	25	27	29	30
freq	10	7	14	13	4	7	3	1	1

mean(N)=16.79 SD(N)=5.30 Total freq=167

For symmetric quaternary walk games, the optimal choice is the $(9, 32)$ game with an admission fee of \$2.82 and a 21.04 cents (or 7.46%) expected loss. The same for symmetric

Table 13: Expected percentage loss for $1 \le c \le 11$ and associated optimal k_*

	symmetric binary walk					symmetric quaternary walk				
		cents			%		cents			
c	k_*	price	E[rew]	E[loss]	E[loss]	k_*	price	E[rew]	E[loss]	E[% loss]
1	1	26	7.50	18.50	71.15	1	51	30.00	21.00	41.17
2	6	56	44.31	11.69	20.88	4	79	54.88	24.12	30.53
3	10	85	70.13	14.87	17.50	7	107	82.79	24.21	22.62
4	16	116	102.27	13.73	11.84	10	135	112.18	22.82	16.90
5	22	147	131.53	15.47	10.53	14	164	142.34	21.66	13.21
6	28	178	160.76	17.24	†9.69	18	193	172.43	20.57	10.66
7	36	211	189.72	21.28	10.09	22	222	202.10	19.90	8.96
8	44	244	220.29	23.71	9.72	27	252	232.01	19.99	7.93
9	54	279	248.65	30.35	10.88	32	282	260.96	21.04	†7.46
10	64	314	279.96	34.04	10.84	39	314	289.98	24.02	7.65
11	72	347	304.72	42.28	12.18	48	348	316.51	31.49	9.05

binary walk games is the $(6, 28)$ game with an admission fee of \$1.78 and a 17.20 cents (or 9.66%) expected loss. However, in the optimum quaternary walk game, the gambler tosses the pair of coins on average 21.98 times (with an SD of 7.20), which can be calculated from Table 8, and in the optimum binary walk game, she tosses the coin on average 16.2 times (with an SD of 6.23). Thus, the gambler loses just under a penny per toss in the optimum quaternary walk game, and just over a penny per toss in the optimum binary walk game. The entertainment value (proportional to the number of tosses) of the optimum quaternary walk game is only marginally higher than that of the optimum binary walk game.

7. Modifications to the Game

The reward random walk (binary or quaternary) on the parity dial is designed to educate gamblers make optimal decisions when the casino offers a game. Recognizing that different gamblers may respond differently, the casino may offer modifications to the initial offer—creating new decision-making opportunities. Sarkar (2020 a) proposed four modifications to the binary walk game: Should the player

(a) interchange nodes within any of the pairs $(1, 2), (3, 4), (5, 6), (8, 9), (10, 11)$?
(b) permute nodes $(8, 9, 10, 11)$?
(c) permute nodes $(5, 6, 7)$?
(d) pay an extra fee of $\lceil k/10 \rceil$ cents for the option to sell back at any time the remaining tosses at half-a-penny each?

Here we pose those same modifications to the quaternary walk game. Let us also pose a couple of new modifications:

(1) For the ternary random walk (which goes from any node to its two neighboring nodes and the node diametrically opposite it with probability 1/3 each), how much admission fee (of the form $a_0 + a_1\, c + k$, where a_0, a_1 are constants) should the casino charge so that even after making the optimal choice of (c, k), the gambler will lose between 5% and 10% of her wager?

(2) Change the usual dial of a 24-hour clock into a parity dial by replacing 24 by 0, and interchanging the pairs $(2, 23), (4, 21), (6, 19), (8, 17), (10, 15), (12, 13)$. Advice the casino how much admission fee they should charge in order to construct a reasonably attractive game (which is still profitable to the casino) involving either a binary, a ternary or a quaternary random walk on this new parity dial.

We also encourage interested readers to construct new games out of other random walks, such as those in Sarkar (2020 a) and Barhoumi, *et al.* (2020), and study business and economics lessons drawn from them.

Acknowledgments

I thank my student Ms. D. Chatterjee and my colleague Dr. M. Rashid for some discussions and technical assistance. Sincere thanks are due to the Chair Editor for his guidance and encouragement.

References

Barhoumi, A., Cheung, C. C., Pilla, M. and Sarkar, J. (2019). *Symmetric Random Walks on Three Half-cubes.* Unpublished manuscript.

Dudewicz, E. J. and Mishra, S. N. (1988). *Modern Mathematical Statistics.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Maiti, S. I. and Sarkar, J. (2019). Random walks on paths and cycles. *Mathematics Magazine,* **92(4)**, 252–268. Mathematical Association of America, Washington DC.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ross, S. M. (1996). *Stochastic Processes.* Second editon. 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, OH.

Sarkar, J. (2020 a). Rules, Resolutions, Randomness and Rewards: Lessons from a Game. In: *Proceeding of International Conference on Frontiers of Operations Research & Business Studies (FORBS)*, accepted. Conference held at Calcutta Business School during 27-28 December 2019. Operations Research Society of India.

Sarkar, J. (2020 b). Wanted! A Mathematician. In: *Special Volume in Honor of Bikas-Bimal Sinha (Twin Statisticians) on Their 75th Birthday*, *Statistics and Applications*, **18(2)**, accepted.

WikipediA: The Free Encyclopedia. Central Limit Theorem. Retrieved on 10 March 2020 from https://en.wikipedia.org/wiki/Central_limit_theorem

WikipediA: The Free Encyclopedia. Law of Large Numbers. Retrieved on 10 March 2020 from https://en.wikipedia.org/wiki/Law_of_large_numbers

APPENDIX: R Codes

1. Compute exact distribution of N for the $(2, 4)$ game

```
### (c=2, k=4)
rf=c(1,11,3,9,5,7,6,8,4,10,2,0)
rew=rep(0,4^4); j=1; toss=c(-2,-1,1,2)
for (x1 in toss){
  for (x2 in toss){
    for (x3 in toss){
      for (x4 in toss){
        u=c(y1,y2,y3,y4)
        cu=cumsum(u); cu=cu-12*floor(cu/12); cu=cu[cu>0]
        f1=min(length(unique(cu)),2); cu=unique(cu)[1:f1]
        rew[j]=sum(rf[cu])
        j=j+1
        } } } }
table(rew)
mean(rew)
sd(rew)
```

2. Simulate random reward earned in any (c, k) game

```
### random reward earned when 4 neighbors are equally likely
rw4=function(f,k){ # f=vertices to capture, k=tosses allowed
  rf=c(1,11,3,9,5,7,6,8,4,10,2,0); toss=c(-2,-1,1,2)
  step=sample(toss,k,replace=T)
  cs=cumsum(step); cs=cs-12*floor(cs/12); cs=cs[cs>0]
  f1=min(length(unique(cs)),f)
  cs=unique(cs)[1:f1]
  sum(rf[cs]) }
data=replicate(10^6, rw4(6, 28)) # vary
mean(data)
sd(data)
table(data)

k=2 # initialize the number of tosses
for (c in 2:11){
  k=k+c; pay=25*(c+1)+k
  data=replicate(10^4,rw4(c,k))
  me=5*mean(data); se=5*sd(data)/10^2
  print( round(c(c, k, pay, me, se, 100*(1-me/pay)), 2) )
  }
```

3. Document and verify optimal k, for each c, via simulation

```
### optimal (c,k) in rw4
k0=c(1, 4, 7, 10, 14, 18, 22, 27, 32, 39, 48)
for (f in 2:11){
  k1=k0[f]-2; k2=k0[f]+2
  for (k in k1:k2){
    data=replicate(10^6,rw4(f,k))
    price=25+25*f+k
    reward=5*mean(data)
    se=5*sd(data)/10^3
    print( round(c(k, price, reward, se, 100*(1-reward/price)),2) ) }
}
```

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 85-125

Improvement Over the Bar-Lev, Bobovitch and Boukai and Tarray and Singh Randomized Response Models Through the Use of Two Variables Having Common Mean

Housila P. Singh and Swarangi M. Gorey
School of Studies in Statistics, Vikram University, Ujjain-456010, M.P., India

Received: 28 December 2019; Revised: 25 April 2020; Accepted: 06 May 2020

Abstract

Taking the clue from Odumade and Singh (2010), we have suggested a procedure to improve the randomized response model envisaged by Tarray and Singh (2014). If there exist two sensitive variables associated to the principal study sensitive variable then those variables could be used to develop ratio type adjustments to the conventional estimators of the population mean of a sensitive variable due to Tarray and Singh (2014). Conditions are obtained under which the suggested ratio-type estimators are better than estimators of Bar-Lev et al (2004) and Tarray and Singh (2014). Numerical illustrations are given in support of the present study.

Key words: Randomized response model; Study variable; Auxiliary variable; Bias; Mean Squared Error.

AMS Subject Classification: 62D05.

1. Introduction

Obtaining information pertaining to sensitive or stigmatizing characteristics has been a vexing problem that is encountered in sample surveys. The questions that make the respondent suffer embarrassment if he (or she) answers the question affirmatively prompt him (or her) to select the path that is least likely to jeopardize his (or her) reputation. This would then entail data that are mostly unreliable. Research in statistical methodology to devise schemes to elicit answers in the above context has been in the direction of finding methods that ensure anonymity to the respondent in as far as his answer is concerned. It is believed that if the interviewer does not know what the answer from the respondent to the sensitive question is, then the respondent feels safe in responding truthfully to the sensitive question. In this direction, an attempt has been made by Warner (1965) by introducing an innovative technique commonly referred to as randomized response (RR) technique for estimating the proportion of population possessing certain stigmatized character (say) by protecting the privacy of respondents and preventing the unacceptable rate of non-response.

Corresponding Author: Swarangi M. Gorey,
Email: swarangi.gorey@gmail.com

Since Warner's (1965) model, a rich growth of literature can be found by the researchers for collecting data on both the qualitative and the quantitative variables. For details, one can refer to Horvitz et al. (1967), Greenberg et al. (1969), Franklin (1989), Fox and Tracy (1986), Grewal et al. (2005-2006), Hong (2005-2006), Ryu et al. (2005-2006), Mahajan et al. (2007), Perri (2008), Singh and Chen (2009), Odumade and Singh (2009, 2010), Singh and Tarray (2012, 2013, 2014), Barabesi et al. (2014) and Singh and Gorey (2016), etc.

1.1 Eichhorn and Hayre's (1983) model

Eichhorn and Hayre (1983) introduced the following RRT model that is based on multiplicative scrambling to collect information on sensitive quantitative variables like income, tax evasion, amount of drug used *etc*. If Y is the true response and S is a scrambling variable (independent of Y) with mean θ and standard deviation γ, then the reported response is given by

$$Z = \frac{YS}{\theta}. \tag{1.1}$$

It is assumed that the distribution of the scrambling variable S is known. In other words, mean (θ) and variance (γ^2) are assumed to be known and positive. Obviously $E(Z) = E(Y)$, which leads to an estimator of the population mean $\bar{Y}$ under simple random sampling with replacement (SRSWR) scheme given by

$$\bar{y}_{(EH)} = \bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \tag{1.2}$$

where $\bar{Z}$ is the sample mean of the reported responses. The variance of $\bar{y}_{(EH)}$ is given by

$$V\left(\bar{y}_{(EH)}\right) = \frac{\bar{Y}^2}{n}\left[C_y^2 + C_\gamma^2\left(1 + C_y^2\right)\right], \tag{1.3}$$

where $C_\gamma = \gamma/\theta$ and $C_y = \sigma_y/\bar{Y}$ are the coefficients of variation of scrambling variable S and the study variable Y, and σ_y is the standard deviation of the study variable y. We shall now discuss a randomized response model studied by Bar-Lev et al (2004), which we call BBB model hereafter.

1.2 Bar-Lev, Bobovitch and Boukai's (2004) RR model

In the BBB model, each respondent is requested to rotate a spinner unobserved by the interviewer. If the spinner stops in the shaded area then the respondent is requested to report the real response on the sensitive variable, say Y_i. If the spinner stops in the non-shaded area then the respondent is requested to report the scrambled response, say $Y_i S$, where S is any scrambling variable and its distribution is assumed to be known. Let p be the proportion of the shaded area of the spinner and $(1-p)$ be the non-shaded area of the spinner as demonstrated in the Figure 1.1.

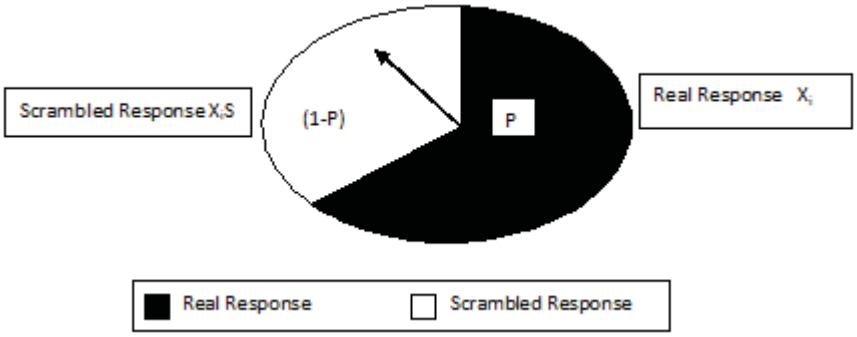


Fig. 1.1: BBB randomized response device

Let Z_i be the response from the ith respondent using BBB randomized response procedure. Then Z_i has the distribution:

$$Z_i = \begin{cases} Y_i S & with\ probability\ (1-p) \\ Y_i & with\ probability\ p \end{cases} \tag{1.4}$$

An unbiased estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_{(BBB)} = \frac{\sum_{i=1}^{n} Z_i}{n\{(1-p)\theta + p\}} . \tag{1.5}$$

The variance under SRSWR sampling is given by

$$V\left(\bar{y}_{(BBB)}\right) = \frac{\bar{Y}^2}{n}\left[C_y^2 + \left(1 + C_y^2\right)C_p^2\right], \tag{1.6}$$

where

$$C_p^2 = \left\{\frac{(1-p)\theta^2\left(1+C_\gamma^2\right)+p}{\{(1-p)\theta+p\}^2} - 1\right\} . \tag{1.7}$$

In Section 1.3, we have revisited the Tarray and Singh (2014) RR models and in section 1.4 description of optimal model of Tarray and Singh (2014) is given.

1.3 A revisit to Tarray and Singh (2014) RR model-I

Using the knowledge of mean θ of scrambling variable S and the design parameter p, Tarray and Singh's (2014) have suggested a randomized response procedure. In the Tarray and Singh's (2014) procedure the distribution of interviewee's response to the sensitive question is:

$$Z_i = \begin{cases} \dfrac{Y_i S}{(1-p)\theta} & with\ probaility\ (1-p) \\ \dfrac{Y_i}{p} & with\ probability\ p \end{cases} \tag{1.8}$$

The expected value of Z_i is given by

$$E(Z_i) = (1-p)\frac{E(Y_i)E(S)}{(1-p)\theta} + p\frac{E(Y_i)}{p}$$
$$= \frac{E(Y_i)(1-p)\theta}{(1-p)\theta} + p\frac{E(Y_i)}{p}$$
$$= 2E(Y_i) = 2\bar{Y}.$$

Thus an unbiased estimator of the population mean μ_y is given by

$$\hat{\mu}_{y(ST1)} = \frac{\bar{Z}}{2} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{2}. \tag{1.9}$$

The variance of $\bar{y}_{(ST1)}$ is given by

$$V(\bar{y}_{(ST1)}) = \frac{V(Z_i)}{4n}. \tag{1.10}$$

The variance of Z_i is obtained as follow

$$V(Z_i) = \left[E(Z_i^2) - (E(Z_i))^2\right]$$
$$= E(Z_i^2) - 4\bar{Y}^2$$
$$= (1-p)\frac{E(Y_i^2)E(S^2)}{\theta^2(1-p)^2} + \frac{pE(Y_i^2)}{p^2} - 4\bar{Y}^2$$
$$= E(Y_i^2)\left[\frac{\theta^2(1+C_\gamma^2)}{\theta^2(1-p)} + \frac{1}{p}\right] - 4\bar{Y}^2$$
$$= \frac{\bar{Y}^2(1+C_y^2)(1+pC_\gamma^2)}{p(1-p)} - 4\bar{Y}^2$$
$$= 4\bar{Y}^2\left[\frac{(1+C_y^2)(1+pC_\gamma^2)}{4p(1-p)} - 1\right]. \tag{1.11}$$

Thus, the variance of $\bar{y}_{(ST1)}$ is given by

$$V(\bar{y}_{(ST1)}) = \frac{V(Z_i)}{4n}$$
$$= \frac{\bar{Y}^2}{n}\left[\frac{(1+C_y^2)(1+pC_\gamma^2)}{4p(1-p)} - 1\right]$$
$$= \frac{\bar{Y}^2}{n}\left[C_y^2 + \frac{(1+C_y^2)(1+pC_\gamma^2)}{4p(1-p)} - 1 - C_y^2\right]$$
$$= \frac{\bar{Y}^2}{n}\left[C_y^2 + (1+C_y^2)\left\{\frac{(1+pC_\gamma^2)}{4p(1-p)} - 1\right\}\right]$$

$$= \frac{\bar{Y}^2}{n}\left[C_y^2 + \left(1 + C_y^2\right)C_{p0}^2\right], \tag{1.12}$$

where

$$C_{p0}^2 = \left[\frac{\left(1 + pC_\gamma^2\right)}{4p(1-p)} - 1\right].$$

We note that the variance of $\bar{y}_{(ST1)}$ obtained in (1.12) is correct while the variance of expression obtained by Tarray and Singh (2014, p.89, equation (2.5)) is incorrect therefore we have revisited the RR model (1.8) due to Tarray and Singh (2014).
From (1.3) and (1.12) we have

$$V\left(\bar{y}_{(EH)}\right) - V\left(\bar{y}_{(ST1)}\right) = \frac{\bar{Y}^2}{n}\left(1 + C_y^2\right)\left[1 + C_\gamma^2 - \frac{\left(1 + pC_\gamma^2\right)}{4p(1-p)}\right]$$

which is always positive if

$$\left\{1 + C_\gamma^2 - \frac{\left(1 + pC_\gamma^2\right)}{4p(1-p)}\right\} > 0. \tag{1.13}$$

Thus, the Tarray and Singh's (2014) estimator $\bar{y}_{(ST1)}$ is more efficient than Eichhorn and Hayre's (1983) estimator $\bar{y}_{(EH)}$ as long as the condition (1.13) is satisfied. Further from (1.6) and (1.12) we have

$$V\left(\bar{y}_{(BBB)}\right) - V\left(\bar{y}_{(ST1)}\right) = \frac{\bar{Y}^2\left(1 + C_y^2\right)}{n}\left(C_p^2 - C_{p0}^2\right)$$

which is positive if

$$\left(C_p^2 - C_{p0}^2\right) > 0$$

i.e. if

$$\frac{\left\{(1-p)\theta^2\left(1 + C_\gamma^2\right) + p\right\}}{\left\{(1-p)\theta + p\right\}^2} > \frac{\left(1 + pC_\gamma^2\right)}{4p(1-p)}. \tag{1.14}$$

Thus the estimator $\bar{y}_{(ST1)}$ due to Tarray and Singh (2014) is more efficient than the Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ if the condition (1.14) is satisfied.

To see the merits of the Tarray and Singh's (2014) unbiased estimator $\bar{y}_{(ST1)}$ we have computed the percent relative efficiency (PRE) of $\bar{y}_{(ST1)}$ with respect to $\bar{y}_{(EH)}$ and $\bar{y}_{(BBB)}$ by using the formulae:

$$PRE\left(\bar{y}_{(ST1)}, \bar{y}_{(EH)}\right) = \frac{\left[C_y^2 + \left(1 + C_y^2\right)C_{p0}^2\right]}{\left[C_y^2 + C_\gamma^2\left(1 + C_y^2\right)\right]} \times 100 \tag{1.15}$$

and

$$PRE\left(\bar{y}_{(ST1)},\bar{y}_{(BBB)}\right)=\frac{\left[C_y^2+\left(1+C_y^2\right)C_{p0}^2\right]}{\left[C_y^2+\left(1+C_y^2\right)C_p^2\right]}\times 100 \tag{1.16}$$

for different values of C_γ, C_x, θ *and* p. Findings are given in Tables 1.1 and 1.2. Tables1.1 and 1.2 show that the values of $PRE\left(\bar{y}_{(ST1)},\bar{y}_{(EH)}\right)$ and $PRE\left(\bar{y}_{(ST1)},\bar{y}_{(BBB)}\right)$ are greater than 100%. Thereby meaning is that the Tarray and Singh's (2014) estimator $\bar{y}_{(ST1)}$ is better than Eichhorn and Hayre's (1983) estimator $\bar{y}_{(EH)}$ and Bar-Lev et al's (2004) estimator $\bar{y}_{(BBB)}$ for the parametric values closed in Tables 1.1 and 1.2.

1.4 Tarray and Singh (2014) RR model- II

Tarray and Singh (2014) have suggested another RR model based on the knowledge of mean θ and square of the coefficient of variation $\left(i.e.\, C_\gamma^2\right)$ of the scrambling variable *S* and design parameter *p*. In this model, the distribution of the responses is given by

$$Z_{0i}=\begin{cases}\dfrac{Y_iS}{\theta\left(1+pC_\gamma^2\right)} & with\ probability\ (1-p)\\ \dfrac{Y_i\left(1+C_\gamma^2\right)}{\left(1+pC_\gamma^2\right)} & with\ probability\ p\end{cases} \tag{1.17}$$

An unbiased estimator of the population mean $\bar{Y}$ based on RR model (1.17) is given by

$$\bar{y}_{(ST2)}=\frac{1}{n}\sum_{i=1}^{n}Z_{0i} \tag{1.18}$$

and the variance is given by

$$V\left(\bar{y}_{(ST2)}\right)=\frac{\bar{Y}^2}{n}\left[C_y^2+\left(1+C_y^2\right)C_{p0}^{*2}\right] \tag{1.19}$$

where

$$C_{p0}^{*2}=\left[\frac{\left(1+C_\gamma^2\right)}{\left(1+pC_\gamma^2\right)}-1\right].$$

Tarray and Singh (2014) have shown that the estimator $\bar{y}_{(ST2)}$ is always better than Eichhorn and Hayre (1983) estimator $\bar{y}_{(EH)}$. They have further shown that the $\bar{y}_{(ST2)}$ is more efficient than the $\bar{y}_{(BBB)}$ due to Bar-Lev et al (2004) if the condition $C_p^2>C_{p0}^{*2}$.

2. Proposed Ratio-Type Estimator Based on Tarray and Singh (2014) Model-I

2.1. Notations

Following Tripathi and Chaubey (1992) let $\bar{X}_{1i}=\bar{X}_{2i}=\bar{X}$ that is these two auxiliary sensitive variables have common mean. Let Y_i be the sensitive variable under study whose mean is to be estimated. A simple random sample with replacement (SRSWR) of *n* respondents is

selected. Then each one of the respondents selected in the sample is requested to rotate three spinners.

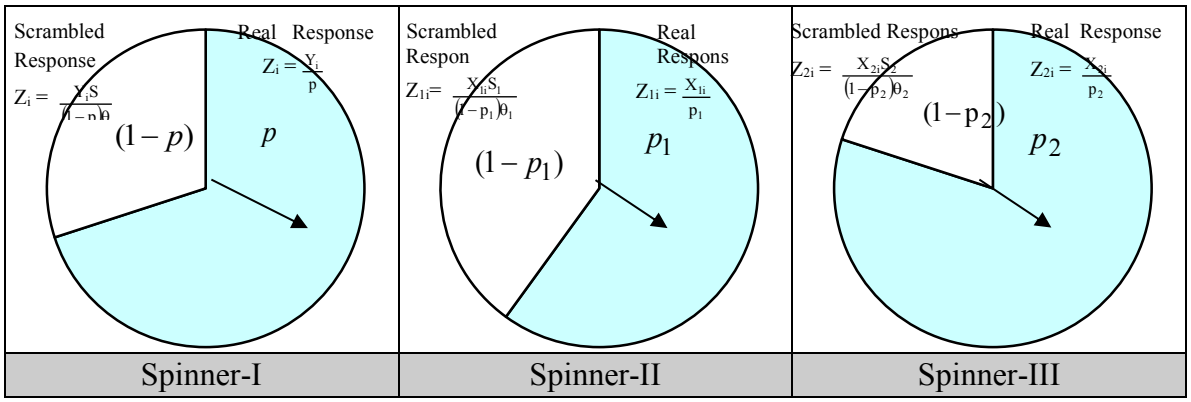


Fig. 2.1: Three spinners

The first spinner is used to gather scrambled response Z_i on the real study variable Y_i with the distribution of responses as:

$$Z_i = \begin{cases} \dfrac{Y_i}{p} & \text{with probability } p \\ \dfrac{Y_i S}{(1-p)\theta} & \text{with probaility } (1-p), \end{cases} \tag{2.1}$$

where the values of (p, θ) are known.

The second spinner is used to gather scrambled response Z_{1i} on the first auxiliary sensitive variable X_{1i} with the distribution of responses as:

$$Z_{1i} = \begin{cases} \dfrac{X_{1i}}{p_1} & \text{with probability } p_1 \\ \dfrac{X_{1i} S_1}{(1-p_1)\theta_1} & \text{with probaility } (1-p_1), \end{cases} \tag{2.2}$$

where the values of (p_1, θ) are known.

The third spinner is used to gather scrambled response Z_{2i} on the second auxiliary sensitive variable X_{2i} with the distribution of responses as:

$$Z_{2i} = \begin{cases} \dfrac{X_{2i}}{p_2} & \text{with probability } p_2 \\ \dfrac{X_{2i} S_2}{(1-p_2)\theta_2} & \text{with probaility } (1-p_2), \end{cases} \tag{2.3}$$

where the value of (p_2, θ) are known.

Assume that the sample mean of the scrambled responses obtained from the respondents in the sample as Z_i, Z_{1i} and Z_{2i} are given by:

$$\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{2},\ \bar{x}_1^* = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_{1i}}{2} \text{ and } \bar{x}_2^* = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_{2i}}{2}.$$

Let us define:

$$\in = \frac{\bar{y}^*}{\bar{Y}} - 1,\ \delta = \frac{\bar{x}_1^*}{\bar{X}} - 1,\ \eta = \frac{\bar{x}_2^*}{\bar{X}} - 1$$

such that

$$E(\in) = E(\delta) = E(\eta) = 0$$

and it can be shown that

$$E(\in^2) = \frac{1}{\bar{Y}^2}V(\bar{y}^*) = \frac{1}{n}\left[C_y^2 + (1 + C_y^2)C_{p0}^2\right],\ E(\delta^2) = \frac{1}{\bar{Y}^2}V(\bar{x}_1^*) = \frac{1}{n}\left[C_{x1}^2 + (1 + C_{x1}^2)C_{p1}^2\right],$$

$$E(\eta^2) = \frac{1}{\bar{Y}^2}V(\bar{x}_2^*) = \frac{1}{n}\left[C_{x2}^2 + (1 + C_{x2}^2)C_{p2}^2\right],\ E(\in\delta) = \frac{1}{n}\rho_{yx1}C_yC_{x1},$$

$$E(\in\eta) = \frac{1}{n}\rho_{yx2}C_yC_{x2}, \text{ and } E(\delta\eta) = \frac{1}{n}\rho_{x1x2}C_{.x1}C_{x2},$$

where $C_{x1} = \sigma_{x1}/\bar{X}$ and $C_{x2} = \sigma_{x2}/\bar{X}$ are the coefficients of variation of the auxiliary sensitive variables x_1 and x_2, $(\sigma_{x1}, \sigma_{x2})$ are the standard deviations of (X_1, X_2), ρ_{yx1} is the correlation coefficient between y and x_1, ρ_{yx2} is the correlation coefficient between y and x_2, ρ_{x1x2} is the correlation coefficient between x_1 and x_2,

$$C_{p0}^2 = \left[\frac{\left(1 + pC_\gamma^2\right)}{4p(1-p)} - 1\right],\ C_{p1}^2 = \left[\frac{\left(1 + p_1C_{\gamma1}^2\right)}{4p_1(1-p_1)} - 1\right],\ C_{p2}^2 = \left[\frac{\left(1 + p_2C_{\gamma2}^2\right)}{4p_2(1-p_2)} - 1\right].$$

We note that:

$$\left.\begin{aligned} &Cov(\bar{y}^*, \bar{x}_1^*) = \frac{\overline{YX}_1}{n}\rho_{yx1}C_yC_{x1}, \\ &Cov(\bar{y}^*, \bar{x}_2^*) = \frac{\overline{YX}_2}{n}\rho_{yx2}C_yC_{x2} \\ &and \\ &Cov(\bar{x}_1^*, \bar{x}_2^*) = \frac{\bar{X}_1\bar{X}_2}{n}\rho_{x1x2}C_{x1}C_{x2} \end{aligned}\right\} \quad (2.4)$$

Proof of the results in (2.4) is simple, so omitted.

2.2. Proposed ratio type estimator

Motivated by Odumade and Singh (2014) we define a ratio estimator for the population mean $\bar{Y}$ (based on the randomized response model-I due to Tarray and Singh (2014) as:

$$\bar{y}^*_{Ratio} = \bar{y}^*\left(\frac{\bar{x}_1^*}{\bar{x}_2^*}\right). \tag{2.5}$$

Note that

$$\bar{y}^* = \bar{Y}(1+\in),\ \bar{x}_1^* = \bar{X}(1+\delta)\ and\ \bar{x}_2^* = \bar{X}(1+\eta).$$

Thus the ratio estimator in (2.5) can be written in terms of $\in, \delta\ and\ \eta$ as:

$$\bar{y}^*_{Ratio} = \bar{Y}(1+\in)\frac{\bar{X}(1+\delta)}{\bar{X}(1+\eta)} = \bar{Y}(1+\in)(1+\delta)(1+\eta)^{-1}$$

We assume that $|\eta|<1$ so that $(1+\eta)^{-1}$ is expandable in terms of η.

Thus

$$\begin{aligned}\bar{y}^*_{Ratio} &= \bar{Y}(1+\in+\delta+\in\delta)\left[1-\eta+\eta^2+...\right]\\ &= \bar{Y}\left[1+\in+\delta-\eta+\eta^2+\in\delta-\in\eta-\delta\eta+...\right]\end{aligned}$$

or

$$\left(\bar{y}^*_{Ratio}-\bar{Y}\right) \cong \bar{Y}\left[\in+\delta-\eta+\eta^2+\in\delta-\in\eta-\delta\eta\right]. \tag{2.6}$$

Taking expectation of both sides of (2.5) we get the bias of the ratio estimator $\bar{y}^*_{Ratio}$ to the first degree of approximation as

$$B\left(\bar{y}^*_{Ratio}\right) = \left(\bar{Y}/n\right)\left[C_{x_2}^2 + \left(1+C_{x_2}^2\right)C_{p_2}^2 + \rho_{yx_1}C_yC_{x_1} - \rho_{yx_2}C_yC_{x_2} - \rho_{x_1x_2}C_{x_1}C_{x_2}\right]$$

Thus, we obtained the following theorem.

Theorem 2.1: The bias in the proposed ratio estimator $\bar{y}^*_{Ratio}$ to the first degree of approximation is given by:

$$B\left(\bar{y}^*_{Ratio}\right) = \left(\bar{Y}/n\right)\left[C_{x_2}^2 + \left(1+C_{x_2}^2\right)C_{p_2}^2 + \rho_{yx_1}C_yC_{x_1} - \rho_{yx_2}C_yC_{x_2} - \rho_{x_1x_2}C_{x_1}C_{x_2}\right] \tag{2.7}$$

Squaring both the sides of (2.6) and neglecting terms of $(\in, \delta, \eta)$ having power greater than two we have

$$\left(\bar{y}^*_{Ratio}-\bar{Y}\right)^2 = \bar{Y}^2\left[\in^2+\delta^2+\eta^2+2\in\delta-2\in\eta-2\eta\delta\right]. \tag{2.8}$$

Taking expectation of both sides of (2.8) we get the mean squared error (MSE) of the ratio estimator $\bar{y}^*_{Ratio}$ to the first degree of approximation, as

$$\begin{aligned}MSE\left(\bar{y}^*_{Ratio}\right) = \left(\bar{Y}^2/n\right)\Big[&C_y^2 + \left(1+C_y^2\right)C_{p_0}^2 + C_{x_1}^2 + \left(1+C_{x_1}^2\right)C_{p_1}^2 + C_{x_2}^2 + \left(1+C_{x_2}^2\right)C_{p_2}^2\\ &+ 2\rho_{yx_1}C_yC_{x_1} - 2\rho_{yx_2}C_yC_{x_2} - 2\rho_{x_1x_2}C_{x_1}C_{x_2}\Big]\end{aligned}$$

Thus, we established the following theorem.

Theorem 2.2: The mean square error of the proposed ratio estimator $\bar{y}^{*}_{Ratio}$ to the first degree of approximation is given by:

$$MSE\left(\bar{y}^{*}_{Ratio}\right)=\left(\bar{Y}^{2}/n\right)\left[C_y^2+\left(1+C_y^2\right)C_{p0}^2+C_{x1}^2+\left(1+C_{x1}^2\right)C_{p1}^2+C_{x2}^2+\left(1+C_{x2}^2\right)C_{p2}^2 + 2\rho_{yx1}C_yC_{x1}-2\rho_{yx2}C_yC_{x2}-2\rho_{x1x2}C_{x1}C_{x2}\right]. \quad (2.9)$$

2.3 Efficiency of the proposed ratio estimator

From (1.6), (1.12) and (2.9) it follows that the proposed ratio-type estimator $\bar{y}^{*}_{Ratio}$ is more efficient than:

(i) the Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ if

$$MSE\left(\bar{y}^{*}_{Ratio}\right)<MSE\left(\bar{y}_{(BBB)}\right)$$

i.e. if

$$\left[C_{x1}^2+C_{x2}^2+\left(1+C_{x1}^2\right)C_{p1}^2+\left(1+C_{x2}^2\right)C_{p2}^2\right] < \left[2\left\{C_y\left(\rho_{yx2}C_{x2}-\rho_{yx1}C_{x1}\right)+\rho_{x1x2}C_{x1}C_{x2}\right\}+\left(1+C_y^2\right)\left(C_p^2-C_{p0}^2\right)\right]. \quad (2.10)$$

(ii) the Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$ if

$$MSE\left(\bar{y}^{*}_{Ratio}\right)<MSE\left(\bar{y}_{(ST1)}\right)$$

i.e. if

$$\left[C_{x1}^2+C_{x2}^2+\left(1+C_{x1}^2\right)C_{p1}^2+\left(1+C_{x2}^2\right)C_{p2}^2\right] < 2\left\{C_y\left(\rho_{yx2}C_{x2}-\rho_{yx1}C_{x1}\right)+\rho_{x1x2}C_{x1}C_{x2}\right\}. \quad (2.11)$$

Thus, the proposed ratio estimator $\bar{y}^{*}_{Ratio}$ will be more efficient than Bar-Lev et al's (2004) estimator $\bar{y}_{(BBB)}$ and Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$ as long as the conditions (2.10) and (2.11) are satisfied respectively. In order to see the performance of the proposed ratio-type estimator $\bar{y}^{*}_{Ratio}$ relative to BBB model and Tarray and Singh's (2014) model, we have computed the percent relative efficiencies (PREs) using the following formulae:

(i) Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$

$$PRE\left(\bar{y}_{(BBB)},\bar{y}^{*}_{Ratio}\right)=\frac{V\left(\bar{y}_{(BBB)}\right)}{MSE\left(\bar{y}^{*}_{Ratio}\right)}\times 100. \quad (2.12)$$

We wrote the MATLAB code and retained those results where the percent relative efficiency (PRE) values are between 300 and 600 to discover the situations where the proposed model can perform better than the Bar-Lev et al (2004) model. In this study we have made a very reasonable choice of a few parameters such as $p, p_1, p_2, C_y, C_{x1}, C_{x2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}, \theta, \theta_1$ *and* θ_2

on which the percent relative efficiency of the ratio estimator depends. It is to be noted that the PRE is free from the sample size n and principal population parameter of interest $\bar{Y}$ the population mean of the study variable y.

We have also written the code to find the values of the parameters $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}, \rho_{x1x2}, \theta, \theta_1$ *and* θ_2 by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values ρ_{yx2} *and* ρ_{x1x2} were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between -0.9 to +0.9 with a step of 0.2. Findings are given in Table 2.1.

Table 2.3: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	525.31
Standard Error	11.37
Median	553.47
Standard Deviation	90.28
Sample Variance	8150.42
Kurtosis	1.75
Skewness	–1.78
Range	285.58
Minimum	313.99
Maximum	599.57
Count	63

It is observed from Table 2.3 that the average percent relative efficiency is 525.31% with the standard deviation 90.28 with median 553.47%, minimum of 313.99% and maximum of 599.57% (see Table 2.1). We also note that there are 63 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600.

(ii) Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$

$$PRE\left(\bar{y}_{(ST1)}, \bar{y}^{*}_{Ratio}\right) = \frac{V\left(\bar{y}_{(ST1)}\right)}{MSE\left(\bar{y}^{*}_{Ratio}\right)} \times 100. \tag{2.13}$$

We have also written the code to find the values of the parameters $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}$ *and* ρ_{x1x2} by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values ρ_{yx2} *and* ρ_{x1x2} were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between –0.9 to +0.9 with a step of 0.2. Findings are given in Table 2.2.

Table 2.4: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	429.27
Standard Error	15.89
Median	387.27
Standard Deviation	111.26
Sample Variance	12379.12
Kurtosis	−1.55
Skewness	0.39
Range	285.82
Minimum	302.12
Maximum	587.94
Count	49

Table 2.4 shows that the average percent relative efficiency is 429.27% with the standard deviation 111.26 with median 387.27%, minimum of 302.12% and maximum of 587.94% (see Table 2.2). It has been observed that there are 49 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600.

3. Proposed Power Transformation Ratio Type Estimator Based on Tarray and Singh (2014) Model –I

Using the repeated substitution method due to Srivastava (1967) and Garcia and Cebrian (1996), we consider a new power transformation ratio type estimator $\bar{y}^{*}_{Power}$ for the population mean $\bar{Y}$ as:

$$\bar{y}^{*}_{Power} = \bar{y}^{*}\left(\frac{\bar{x}_1^{*}}{\bar{x}_2^{*}}\right)^{\alpha}, \tag{3.1}$$

where α is a suitably chosen real constant. For example if $\alpha = 0$ then the proposed power transformation ratio type estimator $\bar{y}^{*}_{Power}$ reduces to the Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$. If $\alpha = 1$ then the proposed power transformation ratio type estimator $\bar{y}^{*}_{Power}$ reduces to the ratio estimator $\bar{y}^{*}_{Ratio}$. Expressing that the proposed transformation ratio-type estimator $\bar{y}^{*}_{Power}$ in terms of $\in$, δ and η, we have:

$$\bar{y}^{*}_{Power} = \bar{Y}(1+\in)\left[\frac{\bar{X}(1+\delta)}{\bar{X}(1+\eta)}\right]^{\alpha} = Y(1+\in)(1+\delta)^{\alpha}(1+\eta)^{-\alpha}. \tag{3.2}$$

We assume that $|\delta| < 1$ and $|\eta| < 1$ so that $(1+\delta)^{\alpha}$ and $(1+\eta)^{-\alpha}$ are expandable. Now expanding the right hand side of (3.2), multiplying out and neglecting terms of $(\in, \delta, \eta)$ having power greater than two we have

$$\bar{y}^*_{power} - \bar{Y} \cong \bar{Y}\left[\in +\alpha(\delta-\eta)+\alpha(\in\delta-\in\eta)-\alpha^2\delta\eta+\frac{\alpha(\alpha-1)}{2}\delta^2+\frac{\alpha(\alpha+1)}{2}\eta^2\right] \quad (3.3)$$

Taking expectation of both sides of (3.3) we get the bias of $\bar{y}^*_{Power}$ to the first degree of approximation as

$$B\left(\bar{y}^*_{Power}\right)=\left(\frac{\bar{Y}}{n}\right)\left[\alpha\left(\rho_{yx_1}C_yC_{x_1}-\rho_{yx_2}C_yC_x\right)-\alpha^2\rho_{x_1x_2}C_{x_1}C_{x_2}\right.$$
$$\left.+\frac{\alpha(\alpha-1)}{2}\left\{C^2_{x_1}+\left(1+C^2_{x_1}\right)C^2_{p_1}\right\}+\frac{\alpha(\alpha+1)}{2}\left\{C^2_{x_2}+\left(1+C^2_{x_2}\right)C^2_{p_2}\right\}\right]$$

Thus, we established the following theorem.

Theorem 3.1. The bias in the proposed power transformation ratio type estimator $\bar{y}^*_{Power}$ is given by:

$$B\left(\bar{y}^*_{Power}\right)=\left(\frac{\alpha\bar{Y}}{n}\right)\left[\rho_{yx_1}C_yC_{x_1}-\rho_{yx_2}C_yC_{x_2}-\alpha\rho_{x_1x_2}C_{x_1}C_{x_2}\right.$$
$$\left.+\frac{(\alpha-1)}{2}\left\{C^2_{x_1}+\left(1+C^2_{x_1}\right)C^2_{p_1}\right\}+\frac{(\alpha+1)}{2}\left\{C^2_{x_2}+\left(1+C^2_{x_2}\right)C^2_{p_2}\right\}\right]. \quad (3.4)$$

The mean squared error of the proposed estimator $\bar{y}^*_{Power}$ is obtained as follows. Squaring both sides of (3.3) and neglecting terms of $(\in,\delta,\eta)$ having power greater than two we have

$$\left(\bar{y}^*_{Power}-\bar{Y}\right)^2=\bar{Y}^2\left[\in^2+\alpha^2(\delta-\eta)^2+2\alpha(\delta\in-\eta\in)\right]. \quad (3.5)$$

Taking expectation of both sides of (3.5) we get the mean squared error of the estimator $\bar{y}^*_{Power}$ as

$$MSE\left(\bar{y}^*_{Power}\right)=\frac{\bar{Y}^2}{n}\left[C^2_y+\left(1+C^2_y\right)C^2_{p0}+\alpha^2\left\{C^2_{x_1}+\left(1+C^2_{x_1}\right)C^2_{p_1}+C^2_{x_2}\right.\right.$$
$$\left.\left.+\left(1+C^2_{x_2}\right)C^2_{p_2}-2\rho_{x_1x_2}C_{x_1}C_{x_2}\right\}-2\alpha\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)\right]$$

which is minimum when

$$\alpha=\frac{\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)}{\left[C^2_{x_1}+C^2_{x_2}-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C^2_{x_1}\right)C^2_{p_1}+\left(1+C^2_{x_2}\right)C^2_{p_2}\right]}$$
$$=\alpha_0(\text{say}).$$

Thus the resulting minimum MSE of $\bar{y}^*_{Power}$ is given by

$$Min.MSE\left(\bar{y}^{*}_{Power}\right)=\frac{\bar{Y}^{2}}{n}\left[C_y^2+\left(1+C_y^2\right)C_{p0}^2-\frac{C_y^2\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)^2}{\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p1}^2+\left(1+C_{x_2}^2\right)C_{p2}^2\right\}}\right]$$

Thus, we arrived at the following theorems.

Theorem 3.2. The mean squared error of the estimator $\bar{y}^{*}_{Power}$ to the first degree of approximation is given by

$$MSE\left(\bar{y}^{*}_{Power}\right)=\left(\frac{\bar{Y}^{2}}{n}\right)\left[C_y^2+\left(1+C_y^2\right)C_{p0}^2+\alpha^2\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p1}^2+\left(1+C_{x_2}^2\right)C_{p2}^2\right\}-2\alpha\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)\right] \tag{3.6}$$

Theorem 3.3. The optimum value of α (for which the MSE $\left(\bar{y}^{*}_{Power}\right)$ in (3.6) is minimum) and the minimum MSE of the estimator $\bar{y}^{*}_{Poewr}$ are respectively given by

$$\alpha=\frac{C_y\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)}{\left[C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p1}^2+\left(1+C_{x_2}^2\right)C_{p2}^2\right]}=\alpha_0 \text{ (say)} \tag{3.7}$$

and

$$Min.MSE\left(\bar{y}^{*}_{Power}\right)=\frac{\bar{Y}^{2}}{n}\left[C_y^2+\left(1+C_y^2\right)C_{p0}^2-\frac{C_y^2\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)^2}{\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p1}^2+\left(1+C_{x_2}^2\right)C_{p2}^2\right\}}\right] \tag{3.8}$$

3.1. Efficiency comparison

3.1.1. When the scalar α does not coincide exactly with its optimum value α_0

From (1.6) and (3.6)

$$MSE\left(\bar{y}^{*}_{Power}\right)-V\left(\bar{y}_{BBB}\right)=\left(\frac{\bar{Y}^{2}}{n}\right)\left[C_y^2+\left(1+C_y^2\right)\left(C_{p0}^2-C_p^2\right)+\alpha^2\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p1}^2+\left(1+C_{x_2}^2\right)C_{p2}^2\right\}-2\alpha\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)\right]$$

which is less than zero if

$$\alpha^2A-2\alpha B+C<0$$

i.e. if

$$\frac{B-\sqrt{\left(B^2-AC\right)}}{A}<\alpha<\frac{B+\sqrt{\left(B^2-AC\right)}}{A} \tag{3.9}$$

where

$$A=\left[C_{x1}^2+C_{x2}^2-2\rho_{x1x2}C_{x1}C_{x2}+\left(1+C_{x1}^2\right)C_{p1}^2+\left(1+C_{x2}^2\right)C_{p2}^2\right],$$
$$B=\left[\rho_{yx2}C_yC_{x2}-\rho_{yx1}C_yC_{x1}\right] \text{ and}$$
$$C=\left[C_y^2+\left(1+C_y^2\right)\left(C_{p0}^2-C_p^2\right)\right].$$

Thus, we state the following theorem.

Theorem 3.4. The proposed power transformation ratio-type estimator $\bar{y}_{Power}^{*}$ is more efficient than the Bar-Lev et al's (2004) estimator $\bar{y}_{BBB}$ as long as the condition (3.9) is satisfied.

Further from (1.12) and (3.6)

$$V\left(\bar{y}_{(ST1)}\right)-MSE\left(\bar{y}_{Power}^{*}\right)=\frac{\bar{Y}^2}{n}\left[2\alpha C_y\left(\rho_{yx2}C_{x2}-\rho_{yx1}C_{x1}\right)\right.$$
$$\left.-\alpha^2\left\{C_{x1}^2+C_{x2}^2-2\rho_{x1x2}C_{x1}C_{x2}+\left(1+C_{x1}^2\right)C_{p1}^2+\left(1+C_{x2}^2\right)C_{p2}^2\right\}\right]$$

which is non-negative if

$$\alpha\left(2\alpha_0-\alpha\right)>0$$

i.e. if $\alpha\left(\alpha-2\alpha_0\right)<0$

i.e. if $\left|\left(\alpha-\alpha_0\right)\right|<\left|\alpha_0\right|,$ (3.10)

where α_0 is given by (3.7) .

Thus. we state the following theorem.

Theorem 3.5. The proposed power transformation ratio-type estimator $\bar{y}_{Poewr}^{*}$ is more efficient than the Tarray and Singh's (2014) estimator $\bar{y}_{(ST1)}$ as long as the condition (3.10) is satisfied.

Further from (2.8) and (3.6)

$$MSE\left(\bar{y}_{Ratio}^{*}\right)-MSE\left(\bar{y}_{Power}^{*}\right)=\left(\frac{\bar{Y}^2}{n}\right)\left[\left(1-\alpha^2\right)A-2\left(1-\alpha\right)B\right]$$
$$=\left(\frac{\bar{Y}^2A}{n}\right)\left[\left(1-\alpha^2\right)-2\left(1-\alpha\right)\frac{B}{A}\right]$$
$$=\left(\frac{\bar{Y}^2A}{n}\right)\left[\left(1-\alpha^2\right)-2\left(1-\alpha\right)\alpha_0\right]$$

which is positive if

$$\left(1-\alpha^2-2\alpha_0+2\alpha\alpha_0\right)>0$$

i.e. if $\left(\alpha^2 - 2\alpha\alpha_0 - 1 + 2\alpha_0\right) < 0$

i.e. if $\left|\left(\alpha - \alpha_0\right)\right| < \left|\alpha_0\right|$ (3.11)

where $A = \left[C_{x1}^2 + C_{x2}^2 - 2\rho_{x1x2}C_{x1}C_{x2} + \left(1 + C_{x1}^2\right)C_{p1}^2 + \left(1 + C_{x2}^2\right)C_{p2}^2\right]$

$B = \left[\rho_{yx2}C_yC_{x2} - \rho_{yx1}C_yC_{x2}\right]$

and $\alpha_0 = \dfrac{B}{A}$ is same as given by (3.7).

Thus, we established the following theorem.

Theorem 3.6. The proposed power transformation ratio-type estimator $\bar{y}_{Power}^*$ is more efficient than the proposed ratio type estimator $\bar{y}_{Ratio}^*$ as long as the condition (3.11) is satisfied.

3.1.2. When the Optimum Value α_0 of the Scalar α is Exactly Known

$$V\left(\bar{y}_{(BBB)}\right) - Min.MSE\left(\bar{y}_{Power}^*\right) = \bar{Y}^2\left[\left(1 + C_y^2\right)\left(C_p^2 - C_{p0}^2\right) + \frac{C_y^2\left[\rho_{yx2}C_{x2} - \rho_{yx1}C_{x2}\right]^2}{\left[C_{x1}^2 + C_{x2}^2 - 2\rho_{x1x2}C_{x1}C_{x2} + \left(1 + C_{x1}^2\right)C_{p1}^2 + \left(1 + C_{x2}^2\right)C_{p2}^2\right]}\right]$$

which is non-negative if

$$C_p^2 > C_{p0}^2 \tag{3.12}$$

Thus, we state the following theorem.

Theorem 3.7. The proposed power transformation ratio-type estimator $\bar{y}_{power}^*$ (at its optimum condition i.e. when $\alpha = \alpha_0$) is better than Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ if $C_p^2 > C_{p0}^2$.

Further from (1.12) and (3.8) we have

$$V\left(\bar{y}_{(ST1)}\right) - Min.MSE\left(\bar{y}_{Power}^*\right) = \frac{S_Y^2}{n}\frac{\left(\rho_{yx2}C_{x2} - \rho_{yx1}C_{x1}\right)}{\left[C_{x1}^2 + C_{x2}^2 - 2\rho_{x1x2}C_{x1}C_{x2} + \left(1 + C_{x1}^2\right)C_{p1}^2 + \left(1 + C_{x2}^2\right)C_{p2}^2\right]} > 0 \text{ provided } \rho_{yx1}C_{x1} \neq \rho_{yx2}C_{x2} \tag{3.13}$$

Thus, we state the following theorem.

Theorem 3.8. The proposed power transformation ratio-type estimator $\bar{y}_{power}^*$ (at its optimum condition i.e. when $\alpha = \alpha_0$) is better than Tarray and Singh's (2014) estimator $\bar{y}_{(ST1)}$ unless $\rho_{yx1}C_{x1} \neq \rho_{yx2}C_{x2}$, the case where both the estimator $\bar{y}_{(ST1)}$ and $\bar{y}_{power}^*$ are equally efficient.

Next from (2.9) and (3.8) we have

$$MSE\left(\bar{y}^{*}_{Ratio}\right) - Min.MSE\left(\bar{y}^{*}_{Power}\right) = \frac{\bar{Y}^{2}}{n}\frac{(A-B)^{2}}{A} > 0 \text{ provided } A \neq B. \quad (3.14)$$

Thus, we state the following theorem

Theorem 3.9. The proposed power transformation ratio-type estimator $\bar{y}^{*}_{power}$ (at its optimum condition i.e. when $\alpha = \alpha_0$) is more efficient than the proposed ratio type estimator $\bar{y}^{*}_{Ratio}$ unless *A=B*, the case where both estimators $\bar{y}^{*}_{Ratio}$ and $\bar{y}^{*}_{power}$ are equally efficient.

3.2. Relative efficiency of the power transformation ratio type estimator

In order to see the magnitude, we computed the percent relative efficiency of the proposed power transformation ratio-type estimator $\bar{y}^{*}_{power}$ with respect to:

(i) Bar-Lev et al (2004) estimator $\bar{y}_{BBB}$

$$PRE\left(\bar{y}_{(BBB)}, \bar{y}^{*}_{Power}\right) = \frac{V\left(\bar{y}_{(BBB)}\right)}{MSE\left(\bar{y}^{*}_{Power}\right)} \times 100 \quad (3.15)$$

We have also written the code to find the values of the parameter $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}, \rho_{x1x2}, \theta, \theta_1$ *and* θ_2 by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x1}, C_{x2}, C_{\gamma}, C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values ρ_{yx_1}, ρ_{yx_2} and $\rho_{x_1x_2}$ were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between –0.9 to +0.9 with a step of 0.2.

Findings are shown in Table 3.1.

Table 3.3: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	507.30
Standard Error	13.64
Median	551.71
Standard Deviation	95.50
Sample Variance	9121.35
Kurtosis	0.38
Skewness	–1.47
Range	292.18
Minimum	300.17
Maximum	592.35
Count	49

Table 3.3 depicts that the average percent relative efficiency is 507.30% with the standard deviation 95.50 with median 551.71%, minimum of 300.17% and maximum of 592.35% (see Table 3.1). It has been observed that there are 49 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600. It has been observed that a choice of larger values of θ, θ_1 and θ_2 may lead to inefficient results, thus the choice of these values is must while using the proposed ratio method in actual practice.

(ii) Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$

$$PRE\left(\bar{y}_{(ST1)}, \bar{y}^{*}_{Power}\right) = \frac{V\left(\bar{y}_{(ST1)}\right)}{MSE\left(\bar{y}^{*}_{Power}\right)} \times 100 \tag{3.16}$$

We have also written the code to find the values of the parameters $C_y, C_{x1}, C_{x2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}$ *and* ρ_{x1x2} by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x1}, C_{x2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values $,\rho_{yx2}$ *and* ρ_{x1x2} were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between -0.9 to +0.9 with a step of 0.2. Findings are shown in Table 3.2.

Table 3.4: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	441.89
Standard Error	13.04
Median	379.03
Standard Deviation	92.18
Sample Variance	8497.77
Kurtosis	–1.83
Skewness	0.09
Range	253.31
Minimum	300.90
Maximum	554.21
Count	50

It is observed from Table 3.4 that the average percent relative efficiency is 441.89% with the standard deviation 92.18 with median 379.03%, minimum of 300.90% and maximum of 554.21% (see Table 3.2). It has been observed that there are 50 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600.

4. Proposed Ratio-Type Estimator Based on Tarray And Singh (2014) Model-II

4.1. Notations

By Tripathi and Chaubey (1992) let $\overline{X}_{1i} = \overline{X}_{2i} = \overline{X}$ that is these two auxiliary sensitive variables have common mean. Let Y_i be the sensitive variable under study whose mean is to be estimated. Consider we selected a simple random sample with replacement (SRSWR) of n respondents. Then each one of the respondents selected in the sample is requested to rotate three spinners.

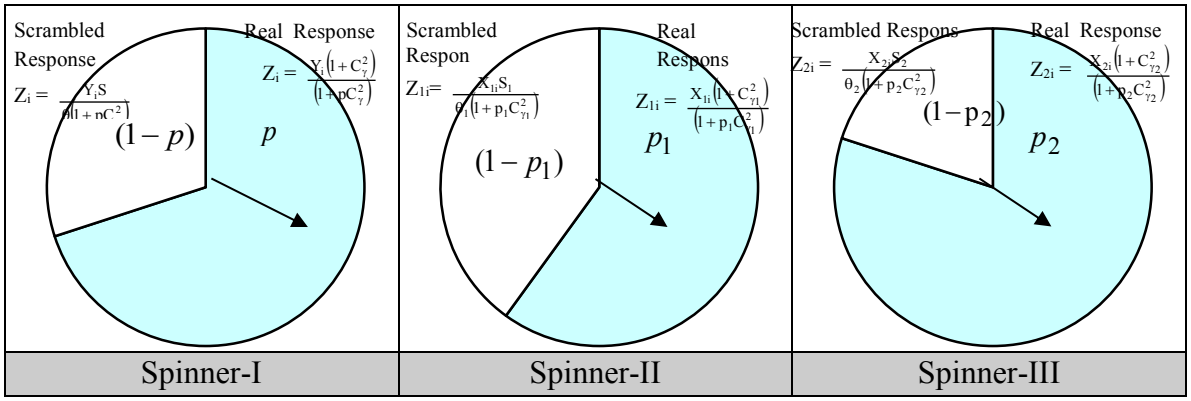


Fig. 4.1: Three spinners

The first spinner is used to collect scrambled response Z_i on the real study variable Y_i with the distribution of responses as:

$$Z_{0i} = \begin{cases} \dfrac{Y_i\left(1+C_\gamma^2\right)}{\left(1+pC_\gamma^2\right)} & \text{with probability } p \\ \dfrac{Y_i S}{\theta\left(1+pC_\gamma^2\right)} & \text{with probaility } (1-p), \end{cases} \tag{4.1}$$

where the value of p is assumed to be known.

The second spinner is used to collect scrambled response Z_{01i} on the first auxiliary sensitive variable X_{1i} with the distribution of responses as:

$$Z_{01i} = \begin{cases} \dfrac{X_{1i}\left(1+C_{\gamma 1}^2\right)}{\left(1+p_1C_{\gamma 1}^2\right)} & \text{with probability } p_1 \\ \dfrac{X_{1i}S_1}{\theta_1\left(1+p_1C_{\gamma 1}^2\right)} & \text{with probaility } (1-p_1), \end{cases} \tag{4.2}$$

where the value of p_1 is assumed to be known.

The third spinner is used to collect scrambled response Z_{02i} on the second auxiliary sensitive variable X_{2i} with the distribution of responses as:

$$Z_{2i} = \begin{cases} \dfrac{X_{2i}\left(1+C_{\gamma_2}^2\right)}{\left(1+p_2 C_{\gamma 2}^2\right)} & \text{with probability } p_2 \\ \dfrac{X_{2i} S_2}{\theta_2\left(1+p_2 C_{\gamma 2}^2\right)} & \text{with probaility } (1-p_2), \end{cases} \tag{4.3}$$

where the value of p_2 is assumed to be known.

Assume that the sample mean of the scrambled responses obtained from the respondents in the sample as Z_{0i}, Z_{01i} and Z_{02i} are given by:

$$\bar{y}^{**} = \frac{1}{n}\sum_{i=1}^{n} Z_{0i}\,,\ \bar{x}_1^{**} = \frac{1}{n}\sum_{i=1}^{n} Z_{01i} \text{ and } \bar{x}_2^{**} = \frac{1}{n}\sum_{i=1}^{n} Z_{02i}\,.$$

Let us define:

$$\in_1 = \frac{\bar{y}^{**}}{\bar{Y}} - 1,\ \delta_1 = \frac{\bar{x}_1^{**}}{\bar{X}} - 1,\ \eta_1 = \frac{\bar{x}_2^{**}}{\bar{X}} - 1$$

such that

$$E(\in_1) = E(\delta_1) = E(\eta_1) = 0$$

and it can be shown that

$$E\left(\in_1^2\right) = \frac{1}{\bar{Y}^2} V\left(\bar{y}^{**}\right) = \frac{1}{n}\left[C_y^2 + \left(1+C_y^2\right) C_{p0}^{*2}\right],$$

$$E\left(\delta_1^2\right) = \frac{1}{\bar{Y}^2} V\left(\bar{x}_1^{**}\right) = \frac{1}{n}\left[C_{x_1}^2 + \left(1+C_{x_1}^2\right) C_{p1}^{*2}\right],$$

$$E\left(\eta_1^2\right) = \frac{1}{\bar{Y}^2} V\left(\bar{x}_2^{**}\right) = \frac{1}{n}\left[C_{x2}^2 + \left(1+C_{x2}^2\right) C_{p2}^{*2}\right],$$

$$E(\in_1\ \delta_1) = \frac{1}{n}\rho_{yx_1} C_y C_{x_1},$$

$$E(\in_1\ \eta_1) = \frac{1}{n}\rho_{yx_2} C_y C_{x2},$$

and

$$E(\delta_1 \eta_1) = \frac{1}{n}\rho_{x_1 x_2} C_{.x_1} C_{x2},$$

where $C_{x_1} = \sigma_{x_1} / \bar{X}$ and $C_{x_2} = \sigma_{x_2} / \bar{X}$ are the coefficients of variation of the auxiliary sensitive variables x_1 and x_2 respectively,

ρ_{yx_1} : is the correlation coefficient between y *and* x_1,

ρ_{yx_2} : is the correlation coefficient between y *and* x_2,

$\rho_{x_1 x_2}$: is the correlation coefficient between x_1 *and* x_2,

$$C_{p0}^{*2}=\left[\frac{\left(1+C_{\gamma}^{2}\right)}{\left(1+pC_{\gamma}^{2}\right)}-1\right],\ C_{p1}^{*2}=\left[\frac{\left(1+C_{\gamma 1}^{2}\right)}{\left(1+p_1C_{\gamma 1}^{2}\right)}-1\right] \text{ and } C_{p2}^{2}=\left[\frac{\left(1+C_{\gamma 2}^{2}\right)}{\left(1+p_2C_{\gamma 2}^{2}\right)}-1\right].$$

4.2. Proposed ratio type estimator

We define a ratio estimator for the population mean $\bar{Y}$ (based on the randomized response model-II due to Tarray and Singh (2014) as:

$$\bar{y}_{Ratio}^{**}=\bar{y}^{**}\left(\frac{\bar{x}_1^{**}}{\bar{x}_2^{**}}\right) \tag{4.4}$$

Note that

$$\bar{y}^{**}=\bar{Y}\left(1+\in_1\right),\ \bar{x}_1^{**}=\bar{X}\left(1+\delta_1\right) \textit{and}\ \bar{x}_2^{**}=\bar{X}\left(1+\eta_1\right).$$

Thus the ratio estimator in (4.4) can be written in terms of $\in_1, \delta_1$ *and* η_1 as:

$$\bar{y}_{Ratio}^{**}=\bar{Y}\left(1+\in_1\right)\frac{\bar{X}\left(1+\delta_1\right)}{\bar{X}\left(1+\eta_1\right)}=\bar{Y}\left(1+\in_1\right)\left(1+\delta_1\right)\left(1+\eta_1\right)^{-1}$$

We assume that $\left|\eta_1\right|<1$ so that $\left(1+\eta_1\right)^{-1}$ is expandable in terms of η_1.

Thus

$$\begin{aligned}\bar{y}_{Ratio}^{**}&=\bar{Y}\left(1+\in_1+\delta_1+\in_1\delta_1\right)\left[1-\eta_1+\eta_1^2+\ldots\right]\\&=\bar{Y}\left[1+\in_1+\delta_1-\eta_1+\eta_1^2+\in_1\delta_1-\in_1\eta_1-\delta_1\eta_1+\ldots\right]\end{aligned}$$

or

$$\left(\bar{y}_{Ratio}^{**}-\bar{Y}\right)\cong\bar{Y}\left[\in_1+\delta_1-\eta_1+\eta_1^2+\in_1\delta_1-\in_1\eta_1-\delta_1\eta_1\right] \tag{4.5}$$

Theorem 4.1: The bias in the proposed ratio estimator $\bar{y}_{\text{Ratio}}^{**}$ to the first degree of approximation is given by:

$$B\left(\bar{y}_{Ratio}^{**}\right)=\left(\bar{Y}/n\right)\left[C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p2}^{*2}+\rho_{yx_1}C_yC_{x_1}-\rho_{yx_2}C_yC_{x_2}-\rho_{x_1x_2}C_{x_1}C_{x_2}\right] \tag{4.6}$$

Proof- Taking expectation of both sides of (4.5) we get the bias of the ratio estimator $\bar{y}_{Ratio}^{**}$ to the first degree of approximation as

$$\begin{aligned}B\left(\bar{y}_{Ratio}^{**}\right)&=E\left(\bar{y}_{Ratio}^{**}\right)-\bar{Y}\\&=\bar{Y}\left[E\left(\eta_1^2\right)+E\left(\in_1\delta_1\right)-E\left(\in_1\eta_1\right)-E\left(\delta_1\eta_1\right)\right]\\&=\left(\bar{Y}/n\right)\left[C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p2}^2+\rho_{yx_1}C_yC_{x_1}-\rho_{yx_2}C_yC_{x_2}-\rho_{x_1x_2}C_{x_1}C_{x_2}\right]\end{aligned}$$

which proves the theorem.

Theorem 4.2: The mean square error of the proposed ratio estimator $\bar{y}^{**}_{Ratio}$ to the first degree of approximation is given by:

$$MSE\left(\bar{y}^{**}_{Ratio}\right)=\left(\bar{Y}^2/n\right)\left[C_y^2+\left(1+C_y^2\right)C_{p0}^{*2}+C_{x1}^2+\left(1+C_{x1}^2\right)C_{p1}^{*2}+C_{x2}^2+\left(1+C_{x2}^2\right)C_{p2}^{*2}\right.$$
$$\left.+2\rho_{yx1}C_yC_{x1}-2\rho_{yx2}C_yC_{x2}-2\rho_{x1x2}C_{x1}C_{x2}\right] \tag{4.7}$$

Proof. Squaring both the sides of (4.5) and neglecting terms of $\left(\in_1,\delta_1,\eta_1\right)$ having power greater than two we have

$$\left(\bar{y}^{**}_{Ratio}-\bar{Y}\right)^2=\bar{Y}^2\left[\in_1^2+\delta_1^2+\eta_1^2+2\in_1\delta_1-2\in_1\eta_1-2\eta_1\delta_1\right] \tag{4.8}$$

Taking expectation of both sides of (4.8) we get the mean squared error (MSE) of the ratio estimator $\bar{y}^{**}_{Ratio}$ as

$$MSE\left(\bar{y}^{**}_{Ratio}\right)=\left(\bar{Y}^2/n\right)\left[C_y^2+\left(1+C_y^2\right)C_{p0}^{2}+C_{x1}^2+\left(1+C_{x1}^2\right)C_{p1}^{2}+C_{x2}^2+\left(1+C_{x2}^2\right)C_{p2}^{2}\right.$$
$$\left.+2\rho_{yx1}C_yC_{x1}-2\rho_{yx2}C_yC_{x2}-2\rho_{x1x2}C_{x1}C_{x2}\right]$$

which proves the theorem.

4.3. Efficiency of the proposed ratio estimator

From (1.6), (1.19) and (4.7) it follows that the proposed ratio-type estimator $\bar{y}^{**}_{Ratio}$ is more efficient than:

(i) the Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ if

$$MSE\left(\bar{y}^{**}_{Ratio}\right)<MSE\left(\bar{y}_{(BBB)}\right)$$

i.e. if

$$\left[C_{x1}^2+C_{x2}^2+\left(1+C_{x1}^2\right)C_{p1}^{*2}+\left(1+C_{x2}^2\right)C_{p2}^{*2}\right]<$$
$$<\left[2\left\{C_y\left(\rho_{yx2}C_{x2}-\rho_{yx1}C_{x1}\right)+\rho_{x1x2}C_{x1}C_{x2}\right\}+\left(1+C_y^2\right)\left(C_p^2-C_{p0}^{*2}\right)\right] \tag{4.9}$$

(ii) the Tarray and Singh (2014) estimator $\bar{y}_{(ST2)}$ if

$$MSE\left(\bar{y}^{**}_{Ratio}\right)<MSE\left(\bar{y}_{(ST2)}\right)$$

i.e. if

$$\left[C_{x1}^2+C_{x2}^2+\left(1+C_{x1}^2\right)C_{p1}^{*2}+\left(1+C_{x2}^2\right)C_{p2}^{*2}\right]$$
$$<2\left\{C_y\left(\rho_{yx2}C_{x2}-\rho_{yx1}C_{x1}\right)+\rho_{x1x2}C_{x1}C_{x2}\right\} \tag{4.10}$$

Thus, the proposed ratio estimator $\bar{y}^{**}_{Ratio}$ will be more efficient than Bar-Lev et al's (2004) estimator $\bar{y}_{(BBB)}$ and Tarray and Singh (2014) estimator $\bar{y}_{(ST2)}$ as long as the conditions (4.9) and (4.10) are satisfied respectively.

We have computed the percent relative efficiencies (PREs) in order to see the performance of the proposed ratio-type estimator $\bar{y}^{**}_{Ratio}$ with respect to Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ by using the formula:

$$PRE\left(\bar{y}_{(BBB)}, \bar{y}^{**}_{Ratio}\right) = \frac{V\left(\bar{y}_{(BBB)}\right)}{MSE\left(\bar{y}^{*}_{Ratio}\right)} \times 100 \tag{4.11}$$

We have also written the code to find the values of the parameter $C_y, C_{x1}, C_{x2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}, \rho_{x1x2}, \theta, \theta_1$ *and* θ_2 by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x1}, C_{x2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values $,\rho_{yx2}$ *and* ρ_{x1x2} were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between -0.9 to +0.9 with a step of 0.2. Findings are given in Table 4.1.

Table 4.3: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	542.31
Standard Error	1.41
Median	539.91
Standard Deviation	10.00
Sample Variance	99.92
Kurtosis	–1.47
Skewness	0.44
Range	24.00
Minimum	531.91
Maximum	555.91
Count	50

It is observed from Table 4.3 that the average percent relative efficiency is 542.31% with the standard deviation 10.00with median 539.91%, minimum of 531.91% and maximum of 555.91% (see Table 4.1). It has been observed that there are 50 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600. It has been observed that a choice of larger values of θ, θ_1 and θ_2 may lead to inefficient results, thus the choice of these values is must while using the proposed ratio method in actual practice.

We have also computed the percent relative efficiencies (PREs) of the proposed ratio-type estimator $\bar{y}^{**}_{Ratio}$ with respect to Tarray and Singh (2014) estimator $\bar{y}_{ST2}$ by using the formula:

$$PRE\left(\bar{y}_{(ST2)}, \bar{y}^{**}_{Ratio}\right) = \frac{V\left(\bar{y}_{(ST2)}\right)}{MSE\left(\bar{y}^{**}_{Ratio}\right)} \times 100 \tag{4.12}$$

We have also written the code to find the values of the parameter $C_y, C_{x_1}, C_{x_2}, C_{\gamma,} C_{\gamma_1}, C_{\gamma_2}, \rho_{yx_1}, \rho_{yx_2}$ *and* $\rho_{x_1x_2}$ by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x_1}, C_{x_2}, C_{\gamma,} C_{\gamma_1}, C_{\gamma_2}$ between 0.1 to 0.5 with a step of 0.2. The values ρ_{yx_2} *and* $\rho_{x_1x_2}$ were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx_1} was changed between –0.9 to +0.9 with a step of 0.2. Findings are given in Table 4.2.

Table 4.4: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	542.31
Standard Error	1.41
Median	539.91
Standard Deviation	10.00
Sample Variance	99.92
Kurtosis	–1.47
Skewness	0.44
Range	24.00
Minimum	531.91
Maximum	555.91
Count	50

Table 4.4 demonstrates that the average percent relative efficiency is 369.77% with the standard deviation 17.03 with median 368.97%, minimum of 348.97% and maximum of 396.97% (see Table 8). It has been observed that there are 50 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600.

5. Proposed Power Transformation Ratio Type Estimator Based on Tarray and Singh (2014) Model –II

A generalized version of the ratio-type estimator $\bar{y}^{**}_{Ratio}$ is given by:

$$\bar{y}^{**}_{Power} = \bar{y}^{**}\left(\frac{\bar{x}_1^{**}}{\bar{x}_2^{**}}\right)^{\alpha_1}, \tag{5.1}$$

where α_1 is a suitably chosen real constant. For example if $\alpha_1 = 0$ then the proposed power transformation ratio type estimator $\bar{y}^{**}_{Power}$ reduces to the Tarray and Singh (2014) estimator $\bar{y}_{ST2}$. If $\alpha_1 = 1$ then the proposed estimator $\bar{y}^{**}_{Power}$ reduces to the ratio estimator $\bar{y}^{**}_{Ratio}$.

Proceeding as earlier the bias and MSE of the estimator $\bar{y}^{**}_{Power}$ can be easily obtained and given in the following theorems.

Theorem 5.1. The bias in the proposed estimator $\bar{y}^{**}_{Power}$ is given by:

$$B\left(\bar{y}^{**}_{Power}\right)=\left(\frac{\alpha_1\bar{Y}}{n}\right)\left[\rho_{yx_1}C_yC_{x_1}-\rho_{yx_2}C_yC_{x_2}-\alpha_1\rho_{x_1x_2}C_{x_1}C_{x_2}\right.$$
$$\left.+\frac{(\alpha_1-1)}{2}\left\{C^2_{x_1}+\left(1+C^2_{x_1}\right)C^{*2}_{p_1}\right\}+\frac{(\alpha_1+1)}{2}\left\{C^2_{x_2}+\left(1+C^2_{x_2}\right)C^{*2}_{p_2}\right\}\right]. \quad (5.2)$$

Theorem 5.2. The mean squared error of the estimator $\bar{y}^{**}_{Power}$ to the first degree of approximation is given by

$$MSE\left(\bar{y}^{**}_{Power}\right)=\left(\frac{\bar{Y}^2}{n}\right)\left[C^2_y+\left(1+C^2_y\right)C^{*2}_{p0}+\alpha_1^2\left\{C^2_{x_1}+C^2_{x_2}-2\rho_{x_1x_2}C_{x_1}C_{x_2}\right.\right.$$
$$\left.\left.+\left(1+C^2_{x_1}\right)C^{*2}_{p_1}+\left(1+C^2_{x_2}\right)C^{*2}_{p_2}\right\}-2\alpha_1\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)\right] \quad (5.3)$$

The optimum value of α_1 and the resulting minimum MSE of the estimator $\bar{y}^{**}_{Power}$ are given in the following theorem.

Theorem 5.3. The optimum value of α (for which the MSE $\left(\bar{y}^{**}_{Power}\right)$ in (5.3) is minimum) and the minimum MSE of the estimator $\bar{y}^{**}_{Poewr}$ are respectively given by

$$\alpha_1=\frac{C_y\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)}{\left[C^2_{x_1}+C^2_{x_2}-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C^2_{x_1}\right)C^{*2}_{p_1}+\left(1+C^2_{x_2}\right)C^{*2}_{p_2}\right]}$$
$$=\alpha_{10}\ \text{(say)} \quad (5.4)$$

and

$$Min.MSE\left(\bar{y}^{**}_{Power}\right)=\frac{\bar{Y}^2}{n}\left[C^2_y+\left(1+C^2_y\right)C^{*2}_{p0}-\right.$$
$$\left.\frac{C^2_y\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)^2}{\left\{C^2_{x_1}+C^2_{x_2}-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C^2_{x_1}\right)C^{*2}_{p_1}+\left(1+C^2_{x_2}\right)C^{*2}_{p_2}\right\}}\right] \quad (5.5)$$

Proof is simple so omitted.

5.1 Efficiency comparison

5.1.1. When the scalar α_1 does not coincide exactly with its optimum value α_{10}

From (1.6) and (5.6)

$$MSE\left(\bar{y}^{**}_{Power}\right)-V\left(\bar{y}_{BBB}\right)=\left(\frac{\bar{Y}^2}{n}\right)\left[C^2_y+\left(1+C^2_y\right)\left(C^{*2}_{p0}-C^2_p\right)\right.$$

$$+\alpha_1^2\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p_1}^{*2}+\left(1+C_{x_2}^2\right)C_{p_2}^{*2}\right\}$$
$$-2\alpha_1\left(\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right)\Big]$$

which is less than zero if

$$\alpha_1^2A-2\alpha_1B+C<0$$

i.e. if

$$\frac{B_1-\sqrt{B_1^2-A_1C_1}}{A_1}<\alpha_1<\frac{B_1+\sqrt{B_1^2-A_1C_1}}{A_1} \tag{5.6}$$

where

$$A_1=\left[C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p_1}^{*2}+\left(1+C_{x_2}^2\right)C_{p_2}^{*2}\right],$$
$$B_1=\left[\rho_{yx_2}C_yC_{x_2}-\rho_{yx_1}C_yC_{x_1}\right],$$
$$C_1=\left[C_y^2+\left(1+C_y^2\right)\left(C_{p0}^{*2}-C_p^2\right)\right].$$

Thus, we state the following theorem.

Theorem 5.4. The proposed estimator $\bar{y}_{Power}^{**}$ is more efficient than the Bar-Lev et al's (2004) estimator $\bar{y}_{(BBB)}$ as long as the condition (5.6) is satisfied.

Further from (1.19) and (5.3) we have

$$V\left(\bar{y}_{ST2}\right)-MSE\left(\bar{y}_{Power}^{**}\right)=\frac{\bar{Y}^2}{n}\left[2\alpha_1C_y\left(\rho_{yx_2}C_{x_2}-\rho_{yx_1}C_{x_1}\right)\right.$$
$$\left.-\alpha_1^2\left\{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}+\left(1+C_{x_1}^2\right)C_{p_1}^{*2}+\left(1+C_{x_2}^2\right)C_{p_2}^{*2}\right\}\right]$$

which is non negative if

$$\alpha_1\left(2\alpha_{10}-\alpha_1\right)>0$$

i.e. if $\alpha_1\left(\alpha_1-2\alpha_{10}\right)<0$

i.e. if $\left|\left(\alpha_1-\alpha_{10}\right)\right|<\left|\alpha_{10}\right|$ (5.7)

where α_{10} is given by (5.4) .

Thus, we state the following theorem.

Theorem 5.5. The proposed estimator $\bar{y}_{Poewr}^{**}$ is more efficient than the Tarray and Singh's (2014) estimator $\bar{y}_{(ST1)}$ as long as the condition (5.7) is satisfied.

Further from (4.8) and (5.3) we have

$$MSE\left(\bar{y}_{Ratio}^{**}\right)-MSE\left(\bar{y}_{Power}^{**}\right)=\left(\frac{\bar{Y}^2}{n}\right)\left[\left(1-\alpha_1^2\right)A_1-2\left(1-\alpha_1\right)B_1\right]$$
$$=\left(\frac{\bar{Y}^2A_1}{n}\right)\left[\left(1-\alpha_1^2\right)-2\left(1-\alpha_1\right)\frac{B_1}{A_1}\right]$$

$$= \left(\frac{\bar{Y}^2 A_1}{n} \right) \left[\left(1 - \alpha_1^2\right) - 2\left(1 - \alpha_1\right)\alpha_{10} \right]$$

which is positive if

$$\left(1 - \alpha_1^2 - 2\alpha_{10} + 2\alpha_1\alpha_{10}\right) > 0$$

i.e. if $\left(\alpha_1^2 - 2\alpha_1\alpha_{10} - 1 + 2\alpha_{10}\right) < 0$

i.e. if $\left(\alpha_1 - \alpha_{10}\right)^2 < \left(1 - \alpha_{10}\right)^2$

i.e. if $\left|\left(\alpha_1 - \alpha_{10}\right)\right| < \left|\alpha_{10}\right|$ (5.8)

and $\alpha_{10} = \frac{B_1}{A_1}$ is same as given by (5.4).

Thus, we established the following theorem.

Theorem 5.6. The proposed estimator $\bar{y}^{**}_{Power}$ is more efficient than the proposed ratio type estimator $\bar{y}^{**}_{Ratio}$ as long as the condition (5.8) is satisfied.

5.1.2 When the optimum value α_{10} of the scalar α_1 is exactly known

$$V\left(\bar{y}_{(BBB)}\right) - Min.MSE\left(\bar{y}^{**}_{Power}\right) = \bar{Y}^2 \left[\left(1 + C_y^2\right)\left(C_p^2 - C_{p0}^{*2}\right) + \frac{C_y^2 \left[\rho_{yx_2} C_{x_2} - \rho_{yx_1} C_{x_2}\right]^2}{\left[C_{x_1}^2 + C_{x_2}^2 - 2\rho_{x_1x_2} C_{x_1} C_{x_2} + \left(1 + C_{x_1}^2\right) C_{p1}^{*2} + \left(1 + C_{x_2}^2\right) C_{p2}^{*2}\right]} \right]$$

which is non negative if

$$C_p^2 > C_{p0}^{*2} \tag{5.9}$$

Thus, we state the following theorem.

Theorem 5.7. The proposed power transformation ratio-type estimator $\bar{y}^{**}_{power}$ (at its optimum condition *i.e.* when $\alpha_1 = \alpha_{10}$) is better than Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ if $C_p^2 > C_{p0}^{*2}$.

Further from (1.19) and (5.5) we have

$$V\left(\bar{y}_{(ST2)}\right) - Min.MSE\left(\bar{y}^{**}_{Power}\right) = \frac{S_Y^2}{n} \frac{\left(\rho_{yx_2} C_{x_2} - \rho_{yx_1} C_{x_1}\right)}{\left[C_{x_1}^2 + C_{x_2}^2 - 2\rho_{x_1x_2} C_{x_1} C_{x_2} + \left(1 + C_{x_1}^2\right) C_{p1}^{*2} + \left(1 + C_{x_2}^2\right) C_{p2}^{*2}\right]}$$

$$> 0 \text{ provided } \rho_{yx_1} C_{x_1} \neq \rho_{yx_2} C_{x_2} \tag{5.10}$$

Thus, we state the following theorem.

Theorem 5.8. The proposed power transformation ratio-type estimator $\bar{y}^{**}_{power}$ (at its optimum condition *i.e.* when $\alpha_1 = \alpha_{10}$) is better than Tarray and Singh's (2014) estimator $\bar{y}_{(ST2)}$ unless $\rho_{yx_1} C_{x_1} \neq \rho_{yx_2} C_{x_2}$, the case where both the estimator $\bar{y}_{(ST2)}$ and $\bar{y}^{**}_{power}$ are equally efficient.

Next from (4.8) and (5.5) we have

$$MSE\left(\bar{y}^{**}_{Ratio}\right) - Min.MSE\left(\bar{y}^{**}_{Power}\right) = \frac{\bar{Y}^2}{n}\frac{\left(A_1 - B_1\right)^2}{A_1} > 0 \text{ provided } A_1 \neq B_1. \tag{5.11}$$

Thus, we state the following theorem

Theorem 5.9. The proposed estimator $\bar{y}^{**}_{power}$ (at its optimum condition i.e. when $\alpha_1 = \alpha_{10}$) is more efficient than the proposed ratio type estimator $\bar{y}^{**}_{Ratio}$ unless $A_1 = B_1$, the case where both estimators $\bar{y}^{**}_{Ratio}$ and $\bar{y}^{**}_{power}$ are equally efficient.

5.2 Relative efficiency of the power transformation ratio type estimator

To see the performance of the proposed estimator $\bar{y}^{**}_{power}$ we computed the percent relative efficiency of the proposed estimator $\bar{y}^{**}_{power}$ with respect to Bar-Lev et al (2004) estimator $\bar{y}_{(BBB)}$ by using the formula

$$PRE\left(\bar{y}_{(BBB)}, \bar{y}^{**}_{Power}\right) = \frac{V\left(\bar{y}_{(BBB)}\right)}{MSE\left(\bar{y}^{**}_{Power}\right)} \times 100 \tag{5.12}$$

We have also written the code to find the values of the parameter $C_y, C_{x_1}, C_{x_2}, C_{\gamma}, C_{\gamma_1}, C_{\gamma_2}, \rho_{yx_1}, \rho_{yx_2}, \rho_{x_1x_2}, \theta, \theta_1$ *and* θ_2 by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x_1}, C_{x_2}, C_{\gamma}, C_{\gamma_1}, C_{\gamma_2}$ between 0.1 to 0.5 with a step of 0.2. The values of θ, θ_1 *and* θ_2 were changed between 0 and 1 with a step of 0.5. The values ρ_{yx_2} *and* $\rho_{x_1x_2}$ were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx_1} was changed between -0.9 to +0.9 with a step of 0.2. Findings are displayed in Table 5.1

Table 5.3: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	496.26
Standard Error	14.66
Median	555.07
Standard	103.68

Deviation	
Sample Variance	10749.68
Kurtosis	–0.98
Skewness	–0.89
Range	296.27
Minimum	300.66
Maximum	596.93
Count	50

It is observed from the Table 5.3that the average percent relative efficiency is 496.26% with the standard deviation 103.68 with median 555.07%, minimum of 300.66% and maximum of 596.93% (see Table 5.1). It has been observed that there are 50 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600. It has been observed that a choice of larger values of θ, θ_1 and θ_2 may lead to inefficient results, thus the choice of these values is must while using the proposed ratio method in actual practice.

We have further computed the percent relative efficiency of the proposed estimator $\bar{y}^{**}_{power}$ with respect to Tarray and Singh (2014) estimator $\bar{y}_{(ST2)}$ by using the formula:

$$PRE\left(\bar{y}_{(ST2)}, \bar{y}^{**}_{Power}\right) = \frac{V\left(\bar{y}_{(ST2)}\right)}{MSE\left(\bar{y}^{**}_{Power}\right)} \times 100 \tag{5.12}$$

We have also written the code to find the values of the parameter $C_y, C_{x_1}, C_{x_2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}, \rho_{yx1}, \rho_{yx2}$ *and* $\rho_{x_1x_2}$ by keeping p, p_1 *and* p_2 each equal to 0.7. We changed the value of $C_y, C_{x_1}, C_{x_2}, C_{\gamma,} C_{\gamma 1}, C_{\gamma 2}$ between 0.1 to 0.5 with a step of 0.2. The values ρ_{yx2} *and* $\rho_{x_1x_2}$ were changed between 0.1 to 0.9 with a step of 0.2 and that of ρ_{yx1} was changed between -0.9 to +0.9 with a step of 0.2.

Findings are presented in Table 5.2.

Table 5.4: Descriptive statistics of the percent relative efficiency

Relative Efficiency	
Mean	408.09
Standard Error	13.94
Median	379.79
Standard Deviation	98.62
Sample Variance	9727.28
Kurtosis	–1.32

Skewness	0.43
Range	294.50
Minimum	300.37
Maximum	594.87
Count	50

Table 5.4 exhibits that the average percent relative efficiency is 408.09% with the standard deviation 98.62 with median 379.79%, minimum of 300.37% and maximum of 594.87% (see Table 5.2). It has been observed that there are 50 cases where the percent relative efficiency of the proposed ratio estimator remains between 300 to 600.

6. Conclusion

In this paper, taking clue from Odumade and Singh (2010), two new ratio-type and power transformation ratio-type estimators have been proposed and compared to BBB model and Tarray and Singh (2014) randomized response model. In the case of scrambled response unlike the repeated substitution method due to Srivastava (1967) and Garcia and Cebrian (1996) it has been observed in general enormity of percent relative efficiency of ratio estimator remains better than the power transformation ratio-type estimator.

Acknowledgement

Authors are indeed thankful to the learned referee and Dr. Rajender Parsad, Executive Editor for their valuable suggestions that helped in preparing nice presentation of the paper.

References

Bar-Lev, S. K., Bobovitch, E. and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, **60**, 255-260.

Barabesi, L., Diana, G. and Perri, P. F. (2014). Horvitz-Thompson estimation with randomized response and non-response. *Model Assisted Statistics and Applications*, **9(1)**, 3-10.

Eichhorn, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, **7**, 307-316.

Fox, J. A. and Tracy, P. E. (1986). *Randomized Response: A Method of Sensitive Surveys.* Newbury Park, CA: SAGE Publications.

Garcia, M. R. and Cebrian, A. A. (1996). Repeated substitution method: The ratio estimator for the population variance. *Metrika*, **43**, 101-105.

Greenberg, B., Abul-Ela, A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response: theoretical framework. *Journal of the American Statistical Association*, **64 (5)**, 529-539.

Grewal, I. S., Bansal, M. L. and Sidhu. S. S. (2005-2006). Population mean corresponding to Horvitz-Thompson's estimator for multi-characteristics using randomized response technique. *Model Assisted Statistics and Applications*, **1**, 215-220.

Hong, Z. (2005-2006). Estimation of mean in randomized response surveys when answers are incompletely truthful. *Model Assisted Statistics and Applications*, **1,** 221-230.

Horvitz, D. G., Shah, B. V. and Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistical Section, American Statistical Association*, 65-72.

Mahajan, P. K., Sharma, P. and Gupta, R. K. (2007). Optimum stratification for allocation proportional to strata totals for scrambled response. *Model Assisted Statistics and Applications*, **2(2)**, 81-88.

Odumade, O. and Singh, S. (2009). Improved Bar-Lev, Bobovitch, and Boukai randomized response models. *Communication in Statistics - Simulation and Computation*, **38(3)**, 473-502.

Odumade, O. and Singh, S. (2010). An alternative to the Bar-Lev, Bobovitch, and Boukai randomized response model. *Sociological Method and Research*, **39(2)**, 206-221.

Perri, P. F. (2008). Modified randomized devices for Simmons' model. *Model Assisted Statistics and Applications*, **3(3)**, 233-239.

Ryu, J. B., Kim, J. M., Heo, T. Y. and Park, C. G. (2005-2006). On stratified randomized response sampling. *Model Assisted Statistics and Applications*, **1**, 31–36.

Singh, H. P. and Tarray, T. A. (2012). A stratified unknown repeated trial in randomized response sampling. *Communication of the Korean Statistical Society*, **19(6)**, 751-759.

Singh, H. P. and Tarray, T. A. (2014). A dexterous randomized response model for estimating a rare sensitive attribute using Poisson distribution. *Statistics and Probability Letters*, **90**, 42-45.

Singh, H. P. and Tarray, T. A. (2013). A modified survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *International Journal of Advanced Scientific and Technical Research*, **3(6)**, 459-472.

Singh, H. P. and Gorey, S. M. (2016). An improved version of Gjestvang and Singh's randomized response model. *Model Assisted Statistics and Applications*, **11 (3)**, 185-190.

Singh, S. and Cheng, S. C. (2009). Utilization of higher order moments of scrambling variables in randomized response sampling. *Journal of Statistical Planning and Inference*, **139**, 3377-3380.

Srivastava, S. K. (1967). An Estimator Using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin*, **16**, 121-132.

Tarray, T. A. and Singh, H. P. (2014). A proficient randomized response model. ISTAT˙IST˙IK: *Journal of the Turkish Statistical Association*, **7(3)**, 987-98.

Tripathi, T. P. and Chaubey, Y. P. (1992). Improved estimation of a finite population mean based on paired observations. *Communication in Statistics-Theory and Methods*, **21**, 3327-3333.

Warner, S. L. (1965). Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, **60**, 63-69.

Annexure

Table 1.1: The Percent Relative Efficiency of the Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$, with respect to Eichhorn and Hayre's (1983) estimator $\bar{y}_{(EH)}$

θ	P	C_γ	C_x	PRE
1.00	0.1	0.10	0.10	8996.82
1.50	0.1	0.10	0.15	5632.14
2.00	0.1	0.10	0.20	3753.53
2.50	0.1	0.10	0.40	1296.88
1.00	0.1	0.25	0.10	2493.11
1.50	0.1	0.25	0.15	2150.34
2.00	0.1	0.25	0.20	1816.14
2.50	0.1	0.25	0.40	964.46
1.00	0.1	0.50	0.10	714.55
1.50	0.1	0.50	0.15	687.20
2.00	0.1	0.50	0.20	653.70
2.50	0.1	0.50	0.40	511.73
1.00	0.1	0.75	0.10	339.61
1.50	0.1	0.75	0.15	334.65
2.00	0.1	0.75	0.20	328.22
2.50	0.1	0.75	0.40	295.81
1.00	0.1	1.00	0.10	204.52
1.50	0.1	1.00	0.15	203.28
2.00	0.1	1.00	0.20	201.65
2.50	0.1	1.00	0.40	192.76
1.00	0.1	1.25	0.10	141.29
1.50	0.1	1.25	0.15	140.98
2.00	0.1	1.25	0.20	140.56
2.50	0.1	1.25	0.40	138.18
1.00	0.1	1.50	0.10	106.76
1.50	0.1	1.50	0.15	106.72
2.00	0.1	1.50	0.20	106.68
2.50	0.1	1.50	0.40	106.40

Table 1.2: The Percent Relative Efficiency of the Tarray and Singh (2014) estimator $\bar{y}_{(ST1)}$ with respect to Bar-Lev et al's (2004) estimator $\bar{y}_{(BBB)}$

θ	P	C_γ	C_x	PRE
1.00	0.1	0.10	0.10	9472.82
1.50	0.1	0.10	0.15	4257.56
2.00	0.1	0.10	0.20	2479.42
2.50	0.1	0.10	0.40	1038.23
1.00	0.1	0.25	0.10	2728.67
1.50	0.1	0.25	0.15	1955.97
2.00	0.1	0.25	0.20	1458.49
2.50	0.1	0.25	0.40	811.33
1.00	0.1	0.50	0.10	790.60
1.50	0.1	0.50	0.15	683.47
2.00	0.1	0.50	0.20	603.03
2.50	0.1	0.50	0.40	462.48
1.00	0.1	0.75	0.10	376.62
1.50	0.1	0.75	0.15	340.49
2.00	0.1	0.75	0.20	315.93
2.50	0.1	0.75	0.40	277.17
1.00	0.1	1.00	0.10	227.00
1.50	0.1	1.00	0.15	208.62
2.00	0.1	1.00	0.20	197.43
2.50	0.1	1.00	0.40	183.84
1.00	0.1	1.25	0.10	156.88
1.50	0.1	1.25	0.15	145.28
2.00	0.1	1.25	0.20	138.77
2.50	0.1	1.25	0.40	133.05
1.00	0.1	1.50	0.10	118.56
1.50	0.1	1.50	0.15	110.23
2.00	0.1	1.50	0.20	105.81
2.50	0.1	1.50	0.40	103.01

Table 2.1: The Percent Relative Efficiency of the proposed estimator $\bar{y}^{*}_{Ratio}$ with respect to Bar-Lev et al's (2004) estimator $\bar{y}_{BBB}$ for the different choice of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	PRE
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.1	0.1	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.1	-0.9	0.3	0.1	0	0	0	524.70
0.1	0.1	0.1	0.1	0.3	0.3	-0.9	0.1	0.1	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.1	0.1	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.5	-0.9	0.1	0.1	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.1	0.1	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.1	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	-0.7	0.1	0.1	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	-0.7	0.1	0.1	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.3	-0.7	0.1	0.1	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.1	0.1	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.5	-0.7	0.1	0.1	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.1	0.1	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.1	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	-0.5	0.1	0.1	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	-0.5	0.1	0.1	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.3	-0.5	0.1	0.1	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.1	0.1	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.5	-0.5	0.1	0.1	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	-0.5	0.9	0.1	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.1	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	-0.3	0.1	0.1	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	-0.3	0.1	0.1	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.3	-0.3	0.3	0.1	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	-0.3	0.3	0.1	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.5	-0.3	0.9	0.9	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	-0.3	0.9	0.9	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	-0.3	0.9	0.9	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	-0.1	0.9	0.9	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	-0.1	0.9	0.9	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.3	-0.1	0.7	0.9	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	-0.1	0.7	0.9	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.5	-0.1	0.3	0.3	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	-0.1	0.3	0.3	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	-0.1	0.3	0.3	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.3	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.3	0.5	0.5	0.5	570.80
0.3	0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.3	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.3	0.1	0.7	0.3	0	0	0	534.29
0.1	0.1	0.1	0.1	0.3	0.5	0.1	0.5	0.1	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.1	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.3	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.1	0.3	0.5	0.5	0.5	570.80
0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.3	0.3	0.5	0.1	0	0	0	534.29
0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.9	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.5	0.3	0.1	0.3	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.1	0.3	0.5	0.5	0.5	599.57
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	1	1	1	313.99
0.1	0.1	0.1	0.1	0.3	0.5	0.5	0.3	0.1	0	0	0	553.47
0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.3	0.1	0.5	0.5	0.5	599.57
0.1	0.1	0.1	0.1	0.3	0.3	0.5	0.9	0.7	0	0	0	534.29
0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.9	0.7	0.5	0.5	0.5	580.39
0.1	0.1	0.1	0.1	0.3	0.1	0.5	0.9	0.7	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.9	0.7	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.1	0.7	0.9	0.1	0	0	0	524.70
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.9	0.1	0.5	0.5	0.5	570.80
0.1	0.1	0.1	0.1	0.3	0.3	0.7	0.3	0.3	0	0	0	534.29

Table 2.2: The Percent Relative Efficiency of the proposed ratio estimator $\bar{y}^{*}_{Ratio}$ with respect to Tarray and Singh's (2014) $\bar{y}_{ST1}$ estimator for different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	PRE
0.1	0.1	0.1	0.1	0.3	0.1	-0.9	0.1	0.1	355.67
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.7	0.5	315.54
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.1	0.1	450.47
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.5	0.5	302.12
0.1	0.1	0.1	0.1	0.3	0.1	-0.9	0.3	0.1	533.30
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.3	0.1	576.75
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.1	0.5	307.58
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.3	0.1	587.94
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.1	315.54
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.9	0.7	321.30
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.5	0.1	343.87
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.7	0.5	302.12
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.3	0.5	307.58
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.9	315.54
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.1	0.9	321.30
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.1	0.1	587.94
0.3	0.3	0.3	0.3	0.3	0.3	-0.3	0.3	0.1	587.94
0.3	0.3	0.3	0.3	0.3	0.1	0.9	0.9	0.9	576.75
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.7	0.9	576.75
0.3	0.3	0.3	0.3	0.5	0.5	0.9	0.5	0.1	307.58
0.3	0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.3	587.94
0.5	0.5	0.5	0.5	0.5	0.5	0.9	0.5	0.1	349.11
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.7	0.1	321.30
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.7	0.1	372.65
0.1	0.1	0.1	0.1	0.3	0.5	-0.9	0.7	0.1	386.04
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.7	0.1	417.64
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.7	0.1	480.84
0.3	0.3	0.3	0.3	0.3	0.3	-0.1	0.1	0.3	587.94
0.1	0.1	0.1	0.1	0.3	0.5	-0.9	0.7	0.1	566.87
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.7	0.3	302.12
0.5	0.5	0.5	0.5	0.3	0.3	-0.5	0.7	0.3	353.46
0.1	0.1	0.1	0.1	0.3	0.3	-0.5	0.7	0.3	365.80
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.7	0.3	397.39
0.5	0.5	0.5	0.5	0.3	0.3	-0.3	0.7	0.3	460.59
0.1	0.1	0.1	0.1	0.3	0.3	-0.3	0.7	0.3	544.49
0.3	0.3	0.3	0.3	0.3	0.3	-0.3	0.7	0.3	587.94
0.5	0.5	0.5	0.5	0.5	0.3	-0.3	0.7	0.3	326.73
0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.9	315.54
0.5	0.5	0.5	0.5	0.3	0.1	0.7	0.3	0.9	343.87
0.1	0.1	0.1	0.1	0.3	0.1	0.7	0.3	0.9	355.67
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.5	0.5	387.27
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.1	0.1	587.94
0.5	0.5	0.5	0.5	0.3	0.1	0.7	0.5	0.5	450.47
0.1	0.1	0.1	0.1	0.3	0.1	0.7	0.5	0.5	533.30
0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.3	587.94
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.5	0.5	576.75
0.1	0.1	0.1	0.1	0.3	0.1	-0.7	0.1	0.1	355.67
0.3	0.3	0.3	0.3	0.3	0.1	-0.7	0.1	0.1	387.27
0.5	0.5	0.5	0.5	0.3	0.1	-0.7	0.1	0.1	450.47

Table 3.1: The Percent Relative Efficiency of the proposed power transformation ratio-type estimator $\bar{y}^{*}_{Power}$ with respect to Bar-Lev et al's (2004) estimator $\bar{y}_{BBB}$ for different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	α_0	PRE
0.3	0.3	0.3	0.3	0.1	0.1	-0.5	0.1	0.1	0.5	0.5	0.5	1.80	553.96
0.5	0.5	0.5	0.3	0.3	0.1	-0.5	0.1	0.1	1	1	1	0.53	553.28
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.1	0.1	0.5	0.5	0.5	0.31	553.15
0.5	0.5	0.5	0.3	0.1	0.1	-0.3	0.1	0.1	1	1	1	1.20	552.16
0.3	0.3	0.3	0.3	0.3	0.1	-0.3	0.1	0.1	0.5	0.5	0.5	0.33	551.72
0.5	0.5	0.5	0.3	0.5	0.1	-0.3	0.1	0.1	1	1	1	0.19	551.64
0.3	0.3	0.3	0.3	0.1	0.1	-0.1	0.1	0.1	0.5	0.5	0.5	0.60	551.08
0.5	0.5	0.5	0.3	0.3	0.1	-0.1	0.1	0.1	1	1	1	0.13	550.88
0.3	0.3	0.3	0.3	0.5	0.1	-0.1	0.1	0.1	0.5	0.5	0.5	0.07	550.85
0.5	0.5	0.5	0.3	0.1	0.1	0.1	0.1	0.1	1	1	1	0.00	550.72
0.3	0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.1	0.5	0.5	0.5	-0.07	550.76
0.5	0.5	0.5	0.3	0.5	0.1	0.1	0.1	0.1	1	1	1	-0.05	550.78
0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.1	0.1	0.5	0.5	0.5	-0.27	551.36
0.5	0.5	0.5	0.3	0.5	0.1	0.3	0.1	0.1	1	1	1	-0.17	551.42
0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.1	0.1	0.5	0.5	0.5	-0.47	552.68
0.5	0.5	0.5	0.3	0.5	0.1	0.5	0.1	0.1	1	1	1	-0.29	552.79
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.1	0.1	0.5	0.5	0.5	-0.67	554.72
0.5	0.5	0.5	0.3	0.5	0.1	0.7	0.1	0.1	1	1	1	-0.41	554.88
0.3	0.3	0.3	0.3	0.3	0.1	0.9	0.1	0.1	0.5	0.5	0.5	-0.87	557.48
0.3	0.3	0.3	0.3	0.5	0.1	0.9	0.1	0.1	0.5	0.5	0.5	-0.53	557.69
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	3.60	563.68
0.3	0.3	0.3	0.5	0.1	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	6.00	301.14
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	0.58	559.01
0.3	0.3	0.3	0.3	0.5	0.1	-0.7	0.3	0.1	0.5	0.5	0.5	0.46	555.92
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.3	0.1	0.5	0.5	0.5	0.34	553.54
0.3	0.3	0.3	0.3	0.5	0.1	-0.3	0.3	0.1	0.5	0.5	0.5	0.22	551.88
0.3	0.3	0.3	0.3	0.5	0.1	-0.1	0.3	0.1	0.5	0.5	0.5	0.10	550.95
0.3	0.3	0.3	0.3	0.5	0.1	0.1	0.3	0.1	0.5	0.5	0.5	-0.02	550.73
0.3	0.3	0.3	0.3	0.5	0.1	0.3	0.3	0.1	0.5	0.5	0.5	-0.14	551.24
0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.3	0.1	0.5	0.5	0.5	-0.26	552.46
0.3	0.3	0.3	0.3	0.5	0.1	0.7	0.3	0.1	0.5	0.5	0.5	-0.38	554.40
0.3	0.3	0.3	0.3	0.5	0.1	0.9	0.3	0.1	0.5	0.5	0.5	-0.50	557.07
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	4.20	568.36
0.3	0.3	0.3	0.5	0.1	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	7.00	314.14
0.5	0.5	0.5	0.3	0.3	0.1	-0.9	0.5	0.1	1	1	1	1.07	560.96
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.9	0.1	0.5	0.5	0.5	1.24	311.58
0.5	0.5	0.5	0.5	0.1	0.3	-0.5	0.9	0.1	1	1	1	16.00	529.14
0.3	0.3	0.3	0.5	0.1	0.3	-0.5	0.9	0.1	0.5	0.5	0.5	16.00	364.19
0.5	0.5	0.5	0.5	0.1	0.3	-0.5	0.9	0.1	1	1	1	16.00	377.35
0.3	0.3	0.3	0.3	0.5	0.3	-0.5	0.9	0.1	0.5	0.5	0.5	0.62	568.45
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.9	0.1	1	1	1	1.04	300.18
0.3	0.3	0.3	0.3	0.1	0.5	-0.9	0.1	0.1	0.5	0.5	0.5	4.20	592.36
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.1	0.1	1	1	1	7.00	338.14
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.1	0.1	0.5	0.5	0.5	1.07	584.96
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.1	0.1	1	1	1	1.78	317.58
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.1	0.1	0.5	0.5	0.5	0.60	583.72
0.5	0.5	0.5	0.5	0.1	0.3	-0.7	0.7	0.1	1	1	1	14.00	469.14
0.3	0.3	0.3	0.5	0.1	0.3	-0.7	0.7	0.1	0.5	0.5	0.5	14.00	304.19
0.5	0.5	0.5	0.5	0.1	0.3	-0.7	0.7	0.1	1	1	1	14.00	317.35
0.3	0.3	0.3	0.3	0.1	0.1	-0.5	0.1	0.1	0.5	0.5	0.5	1.80	553.96

Table 3.2: The Percent Relative Efficiency of the proposed power transformation ratio-type estimator $\bar{y}^{*}_{Power}$ with respect to Tarray and Singh (2014) estimator $\bar{y}_{ST1}$ for different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x1}	C_{x2}	ρ_{yx1}	ρ_{yx2}	ρ_{x1x2}	α_0	PRE
0.3	0.3	0.3	0.3	0.5	0.1	0.3	0.1	0.1	-0.17	537.18
0.3	0.3	0.3	0.3	0.1	0.1	0.5	0.1	0.1	-1.20	361.90
0.3	0.3	0.3	0.3	0.1	0.1	0.5	0.1	0.1	-1.20	537.91
0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.1	0.1	-0.47	362.42
0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.1	0.1	-0.47	538.43
0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.1	0.1	-0.29	362.53
0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.1	0.1	-0.29	538.55
0.3	0.3	0.3	0.3	0.1	0.1	0.7	0.1	0.1	-1.80	363.70
0.3	0.3	0.3	0.3	0.1	0.1	0.7	0.1	0.1	-1.80	539.71
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.1	0.1	-0.67	364.46
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.1	0.1	-0.67	540.47
0.3	0.3	0.3	0.3	0.5	0.1	0.7	0.1	0.1	-0.41	364.62
0.3	0.3	0.3	0.3	0.5	0.1	0.7	0.1	0.1	-0.41	540.63
0.3	0.3	0.3	0.3	0.1	0.1	0.9	0.1	0.1	-2.40	366.22
0.3	0.3	0.3	0.3	0.1	0.1	0.9	0.1	0.1	-2.40	542.23
0.3	0.3	0.3	0.3	0.3	0.1	0.9	0.1	0.1	-0.87	367.22
0.3	0.3	0.3	0.3	0.3	0.1	0.9	0.1	0.1	-0.87	543.23
0.3	0.3	0.3	0.3	0.5	0.1	0.9	0.1	0.1	-0.53	367.43
0.3	0.3	0.3	0.3	0.5	0.1	0.9	0.1	0.1	-0.53	543.44
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.3	0.1	3.60	373.42
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.3	0.1	3.60	549.43
0.5	0.5	0.5	0.5	0.1	0.1	-0.9	0.3	0.1	6.00	311.26
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.3	0.1	0.58	368.75
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.3	0.1	0.58	544.77
0.3	0.3	0.3	0.3	0.1	0.1	-0.7	0.3	0.1	3.00	369.46
0.3	0.3	0.3	0.3	0.1	0.1	-0.7	0.3	0.1	3.00	545.47
0.3	0.3	0.3	0.3	0.5	0.1	-0.7	0.3	0.1	0.46	365.66
0.3	0.3	0.3	0.3	0.5	0.1	-0.7	0.3	0.1	0.46	541.67
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.3	0.1	0.34	363.28
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.3	0.1	0.34	539.30
0.3	0.3	0.3	0.3	0.5	0.1	-0.3	0.3	0.1	0.22	361.62
0.3	0.3	0.3	0.3	0.5	0.1	-0.3	0.3	0.1	0.22	537.64
0.3	0.3	0.3	0.3	0.5	0.1	-0.1	0.3	0.1	0.10	360.69
0.5	0.5	0.5	0.5	0.1	0.3	0.5	0.5	0.1	5.00	308.26
0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.1	0.00	368.46
0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.1	0.00	544.47
0.3	0.3	0.3	0.3	0.1	0.3	0.7	0.5	0.1	2.40	374.22
0.3	0.3	0.3	0.3	0.1	0.3	0.7	0.5	0.1	2.40	550.23
0.3	0.3	0.3	0.3	0.3	0.3	0.9	0.5	0.1	-0.40	369.90
0.3	0.3	0.3	0.3	0.3	0.3	0.9	0.5	0.1	-0.40	545.91
0.3	0.3	0.3	0.3	0.1	0.3	-0.9	0.7	0.1	9.00	361.45
0.5	0.5	0.5	0.5	0.1	0.3	-0.9	0.7	0.1	15.00	383.26
0.3	0.3	0.3	0.3	0.1	0.3	-0.9	0.7	0.1	9.00	449.46
0.5	0.5	0.5	0.5	0.1	0.3	-0.9	0.7	0.1	15.00	424.92
0.5	0.5	0.5	0.5	0.1	0.3	-0.9	0.7	0.1	15.00	508.26
0.3	0.3	0.3	0.3	0.5	0.3	-0.5	0.9	0.1	0.62	554.21
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.9	0.1	1.04	310.30
0.3	0.3	0.3	0.3	0.5	0.3	-0.3	0.9	0.1	0.50	374.81
0.3	0.3	0.3	0.3	0.5	0.3	-0.3	0.9	0.1	0.50	550.82
0.5	0.5	0.5	0.5	0.5	0.3	-0.3	0.9	0.1	0.84	300.90

Table 4.1: The Percent Relative Efficiency of the proposed ratio-type estimator $\bar{y}^{**}_{Ratio}$ with respect to Bar-Lev et al (2004) estimator $\bar{y}_{BBB}$ for different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	α_0	PRE
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.5	0.1	0.5	531.97
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.7	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.7	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.7	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.5	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.5	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.3	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.3	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.3	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.1	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.1	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.1	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.1	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.1	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.3	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.3	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.5	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.5	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.7	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.7	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.7	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.9	0.5	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.9	0.5	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.9	0.5	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.9	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.7	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.7	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.7	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.5	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.5	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.5	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.3	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.3	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.3	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	-0.1	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	-0.1	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	-0.1	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.1	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.1	0.7	0.1	0.5	539.91
0.3	0.3	0.3	0.3	0.5	0.1	0.1	0.7	0.1	0.5	555.91
0.3	0.3	0.3	0.3	0.1	0.1	0.3	0.7	0.1	0.5	531.91
0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.7	0.1	0.5	539.91

Table 4.2: The Percent Relative Efficiency of the proposed ratio-type estimator $\bar{y}^{**}_{Ratio}$ with respect to Tarray and Singh (2014) estimator $\bar{y}_{ST2}$ for different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	$C_{\gamma 1}$	$C_{\gamma 2}$	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	PRE
0.5	0.5	0.5	0.5	0.1	0.1	-0.9	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	-0.7	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	-0.7	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	-0.7	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	-0.5	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	-0.5	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	-0.5	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	-0.3	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	-0.3	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	-0.3	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	-0.1	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	-0.1	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	-0.1	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	0.1	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	0.3	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	0.3	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	0.3	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	0.5	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	0.5	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	0.7	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	0.7	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	0.7	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.1	0.1	0.9	0.1	0.1	348.97
0.5	0.5	0.5	0.5	0.3	0.1	0.9	0.1	0.1	356.97
0.5	0.5	0.5	0.5	0.5	0.1	0.9	0.1	0.1	372.97
0.5	0.5	0.5	0.5	0.3	0.3	0.5	0.1	0.1	364.97
0.5	0.5	0.5	0.5	0.5	0.3	0.5	0.1	0.1	380.97
0.5	0.5	0.5	0.5	0.3	0.3	0.7	0.1	0.1	364.97
0.5	0.5	0.5	0.5	0.5	0.3	0.7	0.1	0.1	380.97
0.5	0.5	0.5	0.5	0.3	0.3	0.9	0.1	0.1	364.97
0.5	0.5	0.5	0.5	0.5	0.3	0.9	0.1	0.1	380.97
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	-0.3	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	-0.1	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	0.7	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.5	0.9	0.1	0.1	396.97
0.5	0.5	0.5	0.5	0.5	0.3	-0.3	0.1	0.1	380.97
0.5	0.5	0.5	0.5	0.3	0.3	-0.1	0.1	0.1	364.97
0.5	0.5	0.5	0.5	0.5	0.3	-0.1	0.1	0.1	380.97
0.5	0.5	0.5	0.5	0.3	0.3	0.1	0.1	0.1	364.97

Table 5.1: The Percent Relative Efficiency of the proposed estimator $\bar{y}^{**}_{Power}$ with respect to Bar-Lev-et al (2004) estimator $\bar{y}_{BBB}$ for different choices of parameter with $p = p_1 = p_2 = 0.7$

C_γ	$C_{\gamma 1}$	$C_{\gamma 2}$	C_y	C_{x1}	C_{x2}	ρ_{yx1}	ρ_{yx2}	ρ_{x1x2}	θ	θ_1	θ_2	α_0	PRE
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.5	0.1	1	1	1	17.00	553.66
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.5	0.1	1	1	1	2.89	339.77
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.5	0.1	1	1	1	16.00	520.66
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.5	0.1	0.5	0.5	0.5	1.53	574.53
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.5	0.1	1	1	1	2.56	323.44
0.5	0.5	0.5	0.5	0.1	0.5	-0.5	0.5	0.1	1	1	1	15.00	320.48
0.3	0.3	0.3	0.3	0.3	0.5	-0.5	0.5	0.1	0.5	0.5	0.5	1.33	569.37
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.5	0.1	1	1	1	2.22	309.10
0.5	0.5	0.5	0.5	0.1	0.5	-0.3	0.5	0.1	1	1	1	14.00	460.66
0.3	0.3	0.3	0.3	0.3	0.5	-0.3	0.5	0.1	0.5	0.5	0.5	1.13	564.93
0.5	0.5	0.5	0.5	0.1	0.5	-0.1	0.5	0.1	1	1	1	13.00	433.66
0.3	0.3	0.3	0.3	0.3	0.5	-0.1	0.5	0.1	0.5	0.5	0.5	0.93	561.21
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.5	0.1	1	1	1	12.00	408.66
0.3	0.3	0.3	0.3	0.3	0.5	0.1	0.5	0.1	0.5	0.5	0.5	0.73	558.21
0.3	0.3	0.3	0.3	0.1	0.5	0.3	0.5	0.1	0.5	0.5	0.5	6.60	596.93
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.5	0.1	1	1	1	11.00	385.66
0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.5	0.1	0.5	0.5	0.5	0.53	555.93
0.3	0.3	0.3	0.3	0.1	0.5	0.5	0.5	0.1	0.5	0.5	0.5	6.00	589.37
0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.5	0.1	1	1	1	10.00	364.66
0.3	0.3	0.3	0.3	0.1	0.5	0.7	0.5	0.1	0.5	0.5	0.5	5.40	582.53
0.5	0.5	0.5	0.5	0.1	0.5	0.7	0.5	0.1	1	1	1	9.00	345.66
0.3	0.3	0.3	0.3	0.3	0.5	0.7	0.5	0.1	0.5	0.5	0.5	0.13	553.53
0.3	0.3	0.3	0.3	0.1	0.5	0.9	0.5	0.1	0.5	0.5	0.5	4.80	576.41
0.5	0.5	0.5	0.5	0.1	0.5	0.9	0.5	0.1	1	1	1	8.00	328.66
0.3	0.3	0.3	0.3	0.3	0.5	0.9	0.5	0.1	0.5	0.5	0.5	-0.07	553.41
0.3	0.3	0.3	0.3	0.1	0.5	-0.9	0.7	0.1	0.5	0.5	0.5	13.20	310.33
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.1	1	1	1	22.00	578.18
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.1	1	1	1	22.00	579.48
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.7	0.1	0.5	0.5	0.5	2.07	591.81
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.7	0.1	1	1	1	3.44	371.44
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.7	0.1	1	1	1	21.00	535.18
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.7	0.1	0.5	0.5	0.5	1.87	584.73
0.5	0.5	0.5	0.5	0.1	0.5	-0.5	0.7	0.1	1	1	1	20.00	494.18
0.3	0.3	0.3	0.3	0.3	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	1.67	578.37
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.7	0.1	1	1	1	2.78	334.10
0.3	0.3	0.3	0.3	0.5	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	0.72	566.33
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.7	0.1	1	1	1	1.20	300.66
0.3	0.3	0.3	0.3	0.3	0.5	-0.3	0.7	0.1	0.5	0.5	0.5	1.47	572.73
0.5	0.5	0.5	0.5	0.3	0.5	-0.3	0.7	0.1	1	1	1	2.44	318.44
0.3	0.3	0.3	0.3	0.5	0.5	-0.3	0.7	0.1	0.5	0.5	0.5	0.60	562.37
0.5	0.5	0.5	0.5	0.1	0.5	-0.1	0.7	0.1	1	1	1	18.00	588.66
0.3	0.3	0.3	0.3	0.5	0.5	-0.1	0.7	0.1	0.5	0.5	0.5	0.48	559.13
0.3	0.3	0.3	0.3	0.3	0.5	0.1	0.7	0.1	0.5	0.5	0.5	1.07	563.61
0.3	0.3	0.3	0.3	0.5	0.5	0.1	0.7	0.1	0.5	0.5	0.5	0.36	556.61
0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.7	0.1	0.5	0.5	0.5	0.87	560.13
0.3	0.3	0.3	0.3	0.5	0.5	0.3	0.7	0.1	0.5	0.5	0.5	0.24	554.81
0.3	0.3	0.3	0.3	0.3	0.5	0.5	0.7	0.1	0.5	0.5	0.5	0.67	557.37
0.3	0.3	0.3	0.3	0.3	0.5	0.7	0.7	0.1	0.5	0.5	0.5	0.47	555.33
0.3	0.3	0.3	0.3	0.3	0.5	0.9	0.7	0.1	0.5	0.5	0.5	0.27	554.01
0.3	0.3	0.3	0.3	0.5	0.5	0.9	0.7	0.1	0.5	0.5	0.5	-0.12	553.73

Table 5.2: The Percent Relative Efficiency of the proposed power transformation ratio-type estimator $\bar{y}^{**}_{Power}$ with respect to Tarray and Singh's (2014) $\bar{y}_{ST2}$ estimator for the different choices of the parameters with $p = p_1 = p_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	C_{x_1}	C_{x_2}	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	α_0	PRE
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.7	0.1	17.00	354.87
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.7	0.1	17.00	378.89
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.7	0.1	16.00	307.86
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.7	0.1	16.00	321.87
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.7	0.1	16.00	345.89
0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.7	0.1	15.00	314.89
0.3	0.3	0.3	0.3	0.1	0.5	-0.9	0.9	0.1	16.20	300.37
0.1	0.1	0.1	0.1	0.1	0.5	-0.9	0.9	0.1	5.40	314.26
0.3	0.3	0.3	0.3	0.1	0.5	-0.9	0.9	0.1	16.20	329.97
0.3	0.3	0.3	0.3	0.1	0.5	-0.9	0.9	0.1	16.20	380.70
0.1	0.1	0.1	0.1	0.1	0.5	-0.7	0.9	0.1	5.20	312.14
0.3	0.3	0.3	0.3	0.1	0.5	-0.7	0.9	0.1	15.60	310.89
0.3	0.3	0.3	0.3	0.1	0.5	-0.7	0.9	0.1	15.60	361.62
0.1	0.1	0.1	0.1	0.1	0.5	-0.5	0.9	0.1	5.00	310.10
0.3	0.3	0.3	0.3	0.1	0.5	-0.5	0.9	0.1	15.00	343.26
0.1	0.1	0.1	0.1	0.1	0.5	-0.3	0.9	0.1	4.80	308.14
0.3	0.3	0.3	0.3	0.1	0.5	-0.3	0.9	0.1	14.40	325.62
0.5	0.5	0.5	0.5	0.1	0.5	-0.1	0.9	0.1	23.00	580.86
0.1	0.1	0.1	0.1	0.1	0.5	-0.1	0.9	0.1	4.60	306.26
0.5	0.5	0.5	0.5	0.1	0.5	-0.1	0.9	0.1	23.00	594.87
0.3	0.3	0.3	0.3	0.1	0.5	-0.1	0.9	0.1	13.80	308.70
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.9	0.1	22.00	535.86
0.1	0.1	0.1	0.1	0.1	0.5	0.1	0.9	0.1	4.40	304.46
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.9	0.1	22.00	549.87
0.5	0.5	0.5	0.5	0.1	0.5	0.1	0.9	0.1	22.00	573.89
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.9	0.1	21.00	492.86
0.1	0.1	0.1	0.1	0.1	0.5	0.3	0.9	0.1	4.20	302.74
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.9	0.1	21.00	506.87
0.5	0.5	0.5	0.5	0.1	0.5	0.3	0.9	0.1	21.00	530.89
0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.9	0.1	20.00	451.86
0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.9	0.1	4.00	301.10
0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.9	0.1	20.00	465.87
0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.9	0.1	20.00	489.89
0.5	0.5	0.5	0.5	0.1	0.5	0.7	0.9	0.1	19.00	412.86
0.5	0.5	0.5	0.5	0.1	0.5	0.7	0.9	0.1	19.00	426.87
0.5	0.5	0.5	0.5	0.1	0.5	0.7	0.9	0.1	19.00	450.89
0.5	0.5	0.5	0.5	0.1	0.5	0.9	0.9	0.1	18.00	375.86
0.5	0.5	0.5	0.5	0.1	0.5	0.9	0.9	0.1	18.00	389.87
0.5	0.5	0.5	0.5	0.1	0.5	0.9	0.9	0.1	18.00	413.89
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.1	22.00	535.86
0.1	0.1	0.1	0.1	0.1	0.5	-0.9	0.7	0.1	4.40	304.46
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.1	22.00	549.87
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.1	22.00	573.89
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.7	0.1	21.00	492.86
0.1	0.1	0.1	0.1	0.1	0.5	-0.7	0.7	0.1	4.20	302.74
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.7	0.1	21.00	506.87
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.7	0.1	21.00	530.89
0.5	0.5	0.5	0.5	0.1	0.5	-0.5	0.7	0.1	20.00	451.86
0.1	0.1	0.1	0.1	0.1	0.5	-0.5	0.7	0.1	4.00	301.10
0.5	0.5	0.5	0.5	0.1	0.5	-0.5	0.7	0.1	20.00	465.87

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 127-140

On Moments of Order Statistics and Some Aspects of Robustness Issues of Lindley Distribution in the Presence of Multiple Outliers

Mahesh Kumar Panda
Department of Statistics, Central University of Odisha, Koraput 763004

Received: 16 October 2019; Revised: 11 May 2020; Accepted: 23 May 2020

Abstract

This article obtains the exact expressions for the single and product moments of order statistics from one parameter Lindley distribution when multiple outliers are present in the data. Using the obtained moment relations, we compute the single and product moments (*e.g.* means, variances and covariances) of various order statistics. Next, we explore the impact of the presence of pronounced outliers on these variances and covariances, while the shape parameters have been shifted in value. We also investigate the robustness issues of the sample L-moments.

Key words: Order statistics; Outliers; Single and product moments; Covariances; Sample L-moments.

AMS Subject Classification: 62G30; 62F10

1. Introduction

In the fields of engineering, medical and biological science, the statistical analysis of lifetime data plays a significant role. In fact, the lifetime distributions are being used in different forms of investigations from the issue of survival time of manufactured items in engineering to the researches involving human diseases in biomedical sciences. In the literature, there are several statistical distributions available for modelling lifetime data. Among these distributions, the predominantly used one is exponential distribution (due to its closed form) for its survival function. The Lindley distribution belongs to the exponential family distribution and can be written as a mixture of exponential and gamma distributions. This distribution is better than exponential failure time distribution, wherein hazard rate is not unimodal or bathtub shaped [see Bakouch et al. (2012)]. The Lindley distribution, having an advantage over the exponential distribution, is due to the fact that the former possesses the increasing hazard rate and decreasing mean residual life time function (MRLF), whereas the latter one possesses constant hazard rate and MRLF. Maybe, owing to this nice property, recently many authors have paid their attention to Lindley distribution as a life time model in different perspectives [see Kumar and Jose (2018)]. Ghitany et al. (2008) showed through waiting time data that Lindley distribution provides a better model as compared to the well-known exponential distribution. This distribution also provides a better fit to competing risks lifetime data in contrast to exponential and Weibull distributions [see Mazuchelia and Achcarb (2011)]. A discrete Lindley model was introduced by Gomoz-Deniz and Calderin-Ojeda (2011) with its applications in collective risk modelling. Krishna and Kumar (2011) demonstrated that Lindley distribution might fit better than exponential, lognormal and gamma distributions in some real life problems under progressive Type-II censoring scheme. Mazucheli et al. (2019) introduced a transformed form of Lindley distribution *i.e.* unit-Lindley distribution and demonstrated that

Corresponding Author: Mahesh Kumar Panda
E-mail: mahesh2123ster@gmail.com

unit-Lindley regression could offer a better fit as compared to beta regression model by using the data of inadequate water supply and sewage in the cities of Brazil from the southeast and northeast region.

The specific area moments of order statistics has been consistently being used in other disciplines such as life testing, reliability theory, signal and image processing *etc*. In the early 70s, many researchers started working on studies of order statistics based on outlier model due to the robustness issue. An outlier in a dataset is an observation that appears to be inconsistent with the remaining observations [see Prasad et al. (2008)]. In any dataset, the presence of single or multiple outlier(s) may leads to a flawed conclusion drawn from the experiment; thus it is important to detect and handle the outlier(s) efficiently. In fact, the detection of multiple outliers in comparison to detection of a single outlier is much more difficult [see Bhar et al. (2013)].

Much of the work on order statistics in connection with robustness issue has been focused when there is one outlier present in the sample (single outlier model), but nothing much in case of multiple outliers model. Barnett and Lewis (1994) extensively discussed the topic of development on the single outlier model. Arnold and Balakrishnan (1989) obtained the density function of r^{th} order statistic as well as the joint density function of $X_{r:n}$ and $X_{s:n}$ ($1 \le r < s \le n$) when the sample of size is n and the sample contains an unknown single outlier. Balakrishnan (1994a) obtained the recurrence relations for the single and product moments of order statistics from right truncated exponential distribution under the multiple-outliers model. Balakrishnan ((1994b), 2007) provided many results on order statistics from multiple-outliers model and the robustness issues involved in those models. Sultan and Moshref (2014) obtained the exact expressions of order statistics for the single and product moments of order statistics from Weibull distribution under the multiple-outliers model (*i.e.* with slippage of observations).

This article derives the exact expressions of order statistics for the single and product moments of order statistics from Lindley distribution when multiple outliers are present in the data. The rest of this article is organized as follows. In Section 2, we give the preliminaries which will be used to derive the main result. In Section 3, we derive the exact expression of the single and product moments of order statistics from Lindley distribution under the multiple-outlier model. In Section 4, we obtain the L-moments of order statistics and also examine the robustness of the sample L-moments in the presence of outliers through some numerical illustrations. In Section 5, we establish some special cases. Finally, in Section 6, we sketch a conclusion of the article.

2. Preliminaries

Under the multiple outliers model set up, we assume that $x_1, x_2, \ldots, x_n$ are independent variables with $x_1, x_2, \ldots, x_{n-p}$ are $(n-p)$ independent random variables from one form of the Lindley distribution with probability density function (pdf) $f(x)$ given by

$$f(x) = \frac{\theta^2}{(1+\theta)}(1+x)e^{-\theta x},\ x \ge 0,\ \theta > 0, \tag{1}$$

while $x_{n-p+1}, x_{n-p+2}, \ldots, x_n$ are the p independent random variables (*i.e.* p outliers) from another form of the Lindley distribution with pdf $g(x)$ given by

$$g(x) = \frac{\tau^2}{(1+\tau)}(1+x)e^{-\tau x},\ x \geq 0,\ \tau > 0, \tag{2}$$

where θ, τ are the shape parameters of pdfs $f(x)$ and $g(x)$ respectively. We also suppose that these parameters are linked with each other by a relationship $\tau = \frac{\theta}{h}$, $h \in (0, 1)$. It can be shown that both the cumulative density functions (cdfs) are related with the corresponding pdfs by the following relationships:

$$f(x) = \frac{\theta^2}{1+\theta+\theta x}(1+x)\{1-F(x)\} \tag{3}$$

and

$$g(x) = \frac{\tau^2}{1+\tau+\tau x}(1+x)\{1-G(x)\}. \tag{4}$$

Let $x_{1:n} \leq x_{2:n} \leq \ldots \leq x_{n:n}$ denote the order statistics obtained from a sample of n observations. Then, the pdf of the r^{th} order statistic $x_{r:n}$, under the multiple-outliers model is as follows [see Childs (1996) and Sultan and Moshref (2014)]:

$$\begin{aligned} f_{r:n}[p](x) = & \sum_{s=\max(0,r-p-1)}^{\min(n-p-1,r-1)} C_1 f(x)\{F(x)\}^s \{G(x)\}^{r-s-1}\{1-F(x)\}^{n-p-s-1}\{1-G(x)\}^{p-r+s+1} \\ & + \sum_{s=\max(0,r-p)}^{\min(n-p,r-1)} C_2 g(x)\{F(x)\}^s \{G(x)\}^{r-s-1}\{1-F(x)\}^{n-p-s}\{1-G(x)\}^{p-r+s}, -\infty < x < \infty \end{aligned} \tag{5}$$

where

$$C_1 = \frac{(n-p)!\,p!}{s!(r-s-1)!(n-p-s-1)!(p-r+s+1)!},$$

and

$$C_2 = \frac{(n-p)!\,p!}{s!(r-s-1)!(n-p-s)!(p-r+s)!}.$$

Similarly, the joint density function of the r^{th} and s^{th} order statistics $x_{r:n}$ and $x_{s:n}$ ($1 \leq r < s \leq n$), under the multiple-outliers model is given by [see Childs (1994) and Sultan and Moshref (2014)]:

$$\begin{aligned} f_{r,s:n}[p](x,y) = & \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j-2)}^{\min(n-p-j-2,r-1)} A_1 f(x) f(y)\{F(x)\}^i \{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j \\ & \times \{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j-2}\{1-G(y)\}^{p-s+i+j+2} \\ & + \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)} A_2 f(x) g(y)\{F(x)\}^i \{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j \\ & \times \{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j-1}\{1-G(y)\}^{p-s+i+j+1} \end{aligned}$$

$$+\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)} A_2 g(x) f(y)\{F(x)\}^i \{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j$$
$$\times\{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j-1}\{1-G(y)\}^{p-s+i+j+1}$$
$$+\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j)}^{\min(n-p-j,r-1)} A_3 g(x) g(y)\{F(x)\}^i \{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j$$
$$\times\{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j}\{1-G(y)\}^{p-s+i+j}, -\infty < x < y < \infty \quad (6)$$

where

$$A_1 = \frac{(n-p)!\,p!}{i!(r-1-i)!\,j!(s-r-1-j)!(n-p-i-j-2)!(p-s+i+j+2)!},$$
$$A_2 = \frac{(n-p)!\,p!}{i!(r-1-i)!\,j!(s-r-1-j)!(n-p-i-j-1)!(p-s+i+j+1)!},$$

and
$$A_3 = \frac{(n-p)!\,p!}{i!(r-1-i)!\,j!(s-r-1-j)!(n-p-i-j)!(p-s+i+j)!}.$$

3. Moments of Order Statistics

In this section, we obtain the exact expressions for the single and product moments of order statistics from Lindley distribution based on multiple-outliers model (based on *p*-outliers observations).

3.1. Single moments

In this subsection, we derive the *k*th moment of the r^{th} order statistics $\mu_{r:n}^{(k)}[p]$, $1 \le r \le n$ under the multiple-outliers model.

Relation 1: For $1 \le r \le n$, and $k = 0, 1, 2, \ldots$ the k^{th} moment $\mu_{r:n}^{(k)}[p]$ is given by

$$\mu_{r:n}^{(k)}[p] = \theta^2 \sum_{s=\max(0,r-p-1)}^{\min(n-p-1,r-1)} c_1 \sum_{i=0}^{s}\binom{s}{i}\sum_{j=0}^{r-s-1}\binom{r-s-1}{j}\sum_{l=0}^{n-p-s+i-1}\binom{n-p-s+i-1}{l}$$
$$\times\sum_{m=0}^{p-r+s+j+1}\binom{p-r+s+j+1}{m}\sum_{q=0}^{l+m+1}\binom{l+m+1}{q}(-1)^{i+j}\frac{\theta^l\tau^m}{(1+\theta)^{n-p-s+i}(1+\tau)^{p-r+s+j+1}}$$
$$\times\frac{\Gamma(k+q+1)}{[\theta(n-p-s+i)+\tau(p-r+s+j+1)]^{k+q+1}}$$
$$+\tau^2\sum_{s=\max(0,r-p)}^{\min(n-p,r-1)} c_2 \sum_{i=0}^{s}\binom{s}{i}\sum_{j=0}^{r-s-1}\binom{r-s-1}{j}\sum_{l=0}^{n-p-s+i}\binom{n-p-s+i}{l}\sum_{m=0}^{p-r+s+j}\binom{p-r+s+j}{m}$$
$$\times\sum_{q=0}^{l+m+1}\binom{l+m+1}{q}(-1)^{i+j}\frac{\theta^l\tau^m}{(1+\theta)^{n-p-s+i}(1+\tau)^{p-r+s+j+1}}$$
$$\times\frac{\Gamma(k+q+1)}{[\theta(n-p-s+i)+\tau(p-r+s+j+1)]^{k+q+1}} \quad (7)$$

Proof: For $1 \le r \le n$, and $k = 0, 1, 2, \ldots$ and by substituting equations (3) and (4) in equation (5), we have

$$\mu_{r:n}^{(k)}[p]=\theta^2\sum_{s=\max(0,r-p-1)}^{\min(n-p-1,r-1)}c_1\int_0^\infty\frac{x^k(1+x)}{1+\theta+\theta x}\{F(x)\}^s\{G(x)\}^{r-s-1}\{1-F(x)\}^{n-p-s}\{1-G(x)\}^{p-r+s+1}dx$$
$$+\tau^2\sum_{s=\max(0,r-p)}^{\min(n-p,r-1)}c_2\int_0^\infty\frac{x^k(1+x)}{1+\tau+\tau x}\{F(x)\}^s\{G(x)\}^{r-s-1}\{1-F(x)\}^{n-p-s}\{1-G(x)\}^{p-r+s+1}dx \tag{8}$$

Using binomial theorem in equation (8) and subsequently expanding the same we get

$$\mu_{r:n}^{(k)}[p]=c_1\theta^2\sum_{s=\max(0,r-p-1)}^{\min(n-p-1,r-1)}\sum_{i=0}^{s}\sum_{j=0}^{r-s-1}(-1)^{i+j}\binom{s}{i}\binom{r-s-1}{j}\frac{1}{(1+\theta)^{n-p-s+i}(1+\tau)^{p-r+s+1+j}}$$
$$\times\int_0^\infty x^k(1+x)(1+\theta+\theta x)^{n-p-s+i-1}\left(1+\tau+\tau x\right)^{p-r+s+1+j}e^{-[(n-p-s+i)\theta+(p-r+s+1+j)\tau]x}dx$$
$$+c_2\tau^2\sum_{s=\max(0,r-p)}^{\min(n-p,r-1)}\sum_{i=0}^{s}\sum_{j=0}^{r-s-1}(-1)^{i+j}\binom{s}{i}\binom{r-s-1}{j}\frac{1}{(1+\theta)^{n-p-s+i}(1+\tau)^{p-r+s+1+j}}$$
$$\times\int_0^\infty x^k(1+x)(1+\theta+\theta x)^{n-p-s+i}\left(1+\tau+\tau x\right)^{p-r+s+j}e^{-[(n-p-s+i)\theta+(p-r+s+1+j)\tau]x}dx \tag{9}$$

Again, using binomial theorem in equation (9) and further simplifying we get the result in equation (7). □

The expression in equation (7) is used to calculate the mean and variance of the order statistics when $n = 6$, $p = 0, 1, 2$, $\theta = 1$ and $\tau = \frac{\theta}{h}$, $h = 0.1, 0.2, 0.3, 0.4, 0.5$ and are presented in Table 1 (Annexure). We can verify the results in Table 1 for the case $p = 0$, by using the well-known identity [see Arnold and Balakrishnan (1989), p. 6]

$$\sum_{i=1}^{6}\mu_{i:6}=6E(X)=6\frac{(\theta+2)}{\theta(\theta+1)}.$$

From Table 1, we see the following:

(1) The variance decreases as p increases.
(2) The variance is an increasing function of h for $r = 1$ and it is a decreasing function of h for $r = 5, 6$. For $r = 2, 3$ and 4 the behaviour is not consistent.
(3) For small r, the relative change in variance is more with the increase in the number of outliers from $p = 1$ to $p = 2$ for different values of h.

3.2. Product moments

In this subsection, we derive the $(k, l)^{\text{th}}$ moment of the r^{th} and s^{th} order statistics $\mu_{r,s:n}^{(k,l)}[p]$, under the multiple-outliers model.

Relation 2. For $1 \le r < s \le n$, and $k, l = 0, 1, 2, \ldots$ the product moments $\mu_{r,s:n}^{(k,l)}[p]$ is given by

$$\mu_{r,s:n}^{(k,l)}[p]=A_1\theta^4\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-2)}^{\min(n-p-j-2,r-1)}\sum_{b=0}^{i}\sum_{d=0}^{r-1-i}\sum_{t=0}^{j}\sum_{q=0}^{s-r-1-j}\sum_{p_1=0}^{b+j-t}\sum_{p_2=0}^{d+s-r-j-q-1}\sum_{p_3=0}^{n-p-i-j+t-2}\sum_{p_4=0}^{p-s+i+j+q+2}\sum_{l_1=0}^{p_1+p_2+1}$$

$$\times \sum_{l_2=0}^{p_3+p_4+1} \sum_{a=0}^{l+l_2} (-1)^{b+d+t+q} \binom{i}{b} \binom{r-1-i}{d} \binom{j}{t} \binom{s-r-1-j}{q} \binom{b+j-t}{p_1}$$
$$\times \binom{d+s-r-j-q-1}{p_2} \binom{n-p-i-j+t-2}{p_3} \binom{p-s+i+j+q+2}{p_4}$$
$$\times \binom{p_1+p_2+1}{l_1} \binom{p_3+p_4+1}{l_2} \frac{\theta^{p_1+p_3} \tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i} (1+\tau)^{d-r+p+i+1}} \frac{(l+l_2)!}{a!}$$
$$\times \frac{\Gamma(k+l_1+a+1)}{\left[\theta(n-s+j+1)\right]^{l+l_1+1} \left[\theta(b+n-r+1)\right]^{k+l_1+a+1}}$$
$$+ 2A_2 \theta^2 \tau^2 \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)} \sum_{b=0}^{i} \sum_{d=0}^{r-1-i} \sum_{t=0}^{j} \sum_{q=0}^{s-r-1-j} \sum_{p_1=0}^{b+j-t} \sum_{p_2=0}^{d+s-r-j-q-1} \sum_{p_3=0}^{n-p-i-j+t-1} \sum_{p_4=0}^{p-s+i+j+q+1}$$
$$\times \sum_{l_1=0}^{p_1+p_2+1} \sum_{l_2=0}^{p_3+p_4+1} \sum_{a=0}^{l+l_2} (-1)^{b+d+t+q} \binom{i}{b} \binom{r-1-i}{d} \binom{j}{t} \binom{s-r-1-j}{q} \binom{b+j-t}{p_1}$$
$$\times \binom{d+s-r-j-q-1}{p_2} \binom{n-p-i-j+t-1}{p_3} \binom{p-s+i+j+q+1}{p_4}$$
$$\times \binom{p_1+p_2+1}{l_1} \binom{p_3+p_4+1}{l_2} \frac{\theta^{p_1+p_3} \tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i} (1+\tau)^{d-r+p+i+1}} \frac{(l+l_2)!}{a!}$$
$$\times \frac{\Gamma(k+l_1+a+1)}{\left[\theta(n-s+j+1)\right]^{l+l_1+1} \left[\theta(b+n-r+1)\right]^{k+l_1+a+1}}$$
$$+ A_3 \tau^4 \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j)}^{\min(n-p-j,r-1)} \sum_{b=0}^{i} \sum_{d=0}^{r-1-i} \sum_{t=0}^{j} \sum_{q=0}^{s-r-1-j} \sum_{p_1=0}^{b+j-t} \sum_{p_2=0}^{d+s-r-j-q-1} \sum_{p_3=0}^{n-p-i-j+t} \sum_{p_4=0}^{p-s+i+j+q} \sum_{l_1=0}^{p_1+p_2+1}$$
$$\times \sum_{l_2=0}^{p_3+p_4+1} \sum_{a=0}^{l+l_2} (-1)^{b+d+t+q} \binom{i}{b} \binom{r-1-i}{d} \binom{j}{t} \binom{s-r-1-j}{q} \binom{b+j-t}{p_1}$$
$$\times \binom{d+s-r-j-q-1}{p_2} \binom{n-p-i-j+t}{p_3} \binom{p-s+i+j+q}{p_4} \binom{p_1+p_2+1}{l_1}$$
$$\times \binom{p_3+p_4+1}{l_2} \frac{\theta^{p_1+p_3} \tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i} (1+\tau)^{d-r+p+i+1}} \frac{(l+l_2)!}{a!}$$
$$\times \frac{\Gamma(k+l_1+a+1)}{\left[\theta(n-s+j+1)\right]^{l+l_1+1} \left[\theta(b+n-r+1)\right]^{k+l_1+a+1}} \tag{10}$$

Proof: For $1 \le r < s \le n$, and $k, l = 0, 1, 2, \ldots$ and by using equations (3) and (4) in equation (6), we have

$$\mu_{r,s:n}^{(k,l)}[p] = \theta^4 \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j-2)}^{\min(n-p-j-2,r-1)} A_1 \int_0^\infty \int_x^\infty \frac{x^k (1+x) y^l (1+y)}{(1+\theta+\theta x)(1+\theta+\theta y)} \{F(x)\}^i \{G(x)\}^{r-1-i} \{F(y)-F(x)\}^j$$
$$\times \{G(y)-G(x)\}^{s-r-1-j} \{1-F(y)\}^{n-p-i-j-1} \{1-G(y)\}^{p-s+i+j+2} \{1-F(x)\} dx dy$$
$$+ \theta^2 \tau^2 \sum_{j=0}^{s-r-1} \sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)} A_2 \int_0^\infty \int_x^\infty \frac{x^k (1+x) y^l (1+y)}{(1+\theta+\theta x)(1+\tau+\tau y)} \{F(x)\}^i \{G(x)\}^{r-1-i} \{F(y)-F(x)\}^j$$
$$\times \{G(y)-G(x)\}^{s-r-1-j} \{1-F(y)\}^{n-p-i-j-1} \{1-G(y)\}^{p-s+i+j+2} \{1-F(x)\} dx dy$$

$$+\theta^2\tau^2\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)}A_2\int_0^\infty\int_x^\infty\frac{x^k(1+x)y^l(1+y)}{(1+\tau+\varpi x)(1+\theta+\theta y)}\{F(x)\}^i\{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j$$
$$\times\{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j}\{1-G(y)\}^{p-s+i+j+1}\{1-G(x)\}dxdy$$
$$+\tau^4\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j)}^{\min(n-p-j,r-1)}A_3\int_0^\infty\int_x^\infty\frac{x^k(1+x)y^l(1+y)}{(1+\tau+\varpi x)(1+\tau+\varpi y)}\{F(x)\}^i\{G(x)\}^{r-1-i}\{F(y)-F(x)\}^j$$
$$\times\{G(y)-G(x)\}^{s-r-1-j}\{1-F(y)\}^{n-p-i-j}\{1-G(y)\}^{p-s+i+j+1}\{1-G(x)\}dxdy. \quad (11)$$

Using binomial theorem in equation (11) and subsequently expanding the same we get

$$\mu_{r,s:n}^{(k,l)}[p]=\theta^4\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-2)}^{\min(n-p-j-2,r-1)}A_1\sum_{b=0}^{i}\sum_{d=0}^{r-1-i}\sum_{t=0}^{j}\sum_{q=0}^{s-r-1-j}\sum_{p_1=0}^{b+j-t}\sum_{p_2=0}^{d+s-r-j-q-1}\sum_{p_3=0}^{n-p-i-j+t-2}\sum_{p_4=0}^{p-s+i+j+q+2}\sum_{l_1=0}^{p_1+p_2+1}\sum_{l_2=0}^{p_3+p_4+1}$$
$$\times(-1)^{b+d+t+q}\binom{i}{b}\binom{r-1-i}{d}\binom{j}{t}\binom{s-r-1-j}{q}\binom{b+j-t}{p_1}\binom{d+s-r-j-q-1}{p_2}$$
$$\times\binom{n-p-i-j+t-2}{p_3}\binom{p-s+i+j+q+2}{p_4}\binom{p_1+p_2+1}{l_1}\binom{p_3+p_4+1}{l_2}$$
$$\times\frac{\theta^{p_1+p_3}\tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i}(1+\tau)^{d-r+p+i+1}}$$
$$\times\int_0^\infty\int_x^\infty x^{k+l_1}y^{l+l_2}e^{-x\{\theta(b+j-t+1)+\tau(d+s-r-1-j+q)\}}e^{-y\{\theta(n-p-i-j+t-1)+\tau(p-s+i+j+q+2)\}}dxdy$$
$$+\theta^2\tau^2\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)}A_2\sum_{b=0}^{i}\sum_{d=0}^{r-1-i}\sum_{t=0}^{j}\sum_{q=0}^{s-r-1-j}\sum_{p_1=0}^{b+j-t}\sum_{p_2=0}^{d+s-r-j-q-1}\sum_{p_3=0}^{n-p-i-j+t-1}\sum_{p_4=0}^{p-s+i+j+q+1}\sum_{l_1=0}^{p_1+p_2+1}\sum_{l_2=0}^{p_3+p_4+1}$$
$$\times(-1)^{b+d+t+q}\binom{i}{b}\binom{r-1-i}{d}\binom{j}{t}\binom{s-r-1-j}{q}\binom{b+j-t}{p_1}\binom{d+s-r-j-q-1}{p_2}$$
$$\times\binom{n-p-i-j+t-1}{p_3}\binom{p-s+i+j+q+1}{p_4}\binom{p_1+p_2+1}{l_1}\binom{p_3+p_4+1}{l_2}$$
$$\times\frac{\theta^{p_1+p_3}\tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i}(1+\tau)^{d-r+p+i+1}}$$
$$\times\int_0^\infty\int_x^\infty x^{k+l_1}y^{l+l_2}e^{-x\{\theta(b+j-t+1)+\tau(d+s-r-1-j+q)\}}e^{-y\{\theta(n-p-i-j+t-1)+\tau(p-s+i+j+q+2)\}}dxdy$$
$$+\theta^2\tau^2\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j-1)}^{\min(n-p-j-1,r-1)}A_2\sum_{b=0}^{i}\sum_{d=0}^{r-1-i}\sum_{t=0}^{j}\sum_{q=0}^{s-r-1-j}\sum_{p_1=0}^{b+j-t}\sum_{p_2=0}^{d+s-r-j-q-1}\sum_{p_3=0}^{n-p-i-j+t-1}\sum_{p_4=0}^{p-s+i+j+q+1}$$
$$\times\sum_{l_1=0}^{p_1+p_2+1}\sum_{l_2=0}^{p_3+p_4+1}(-1)^{b+d+t+q}\binom{i}{b}\binom{r-1-i}{d}\binom{j}{t}\binom{s-r-1-j}{q}\binom{b+j-t}{p_1}$$
$$\times\binom{d+s-r-j-q-1}{p_2}\binom{n-p-i-j+t-1}{p_3}\binom{p-s+i+j+q+1}{p_4}$$
$$\times\binom{p_1+p_2+1}{l_1}\binom{p_3+p_4+1}{l_2}\frac{\theta^{p_1+p_3}\tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i}(1+\tau)^{d-r+p+i+1}}$$
$$\times\int_0^\infty\int_x^\infty x^{k+l_1}y^{l+l_2}e^{-x\{\theta(b+j-t)+\tau(d+s-r-j+q)\}}e^{-y\{\theta(n-p-i-j+t)+\tau(p-s+i+j+q+1)\}}dxdy$$
$$+\tau^4\sum_{j=0}^{s-r-1}\sum_{i=\max(0,s-p-j)}^{\min(n-p-j,r-1)}A_3\sum_{b=0}^{i}\sum_{d=0}^{r-1-i}\sum_{t=0}^{j}\sum_{q=0}^{s-r-1-j}\sum_{p_1=0}^{b+j-t}\sum_{p_2=0}^{d+s-r-j-q-1}\sum_{p_3=0}^{n-p-i-j+t}\sum_{p_4=0}^{p-s+i+j+q}\sum_{l_1=0}^{p_1+p_2+1}\sum_{l_2=0}^{p_3+p_4+1}$$

$$\times(-1)^{b+d+t+q}\binom{i}{b}\binom{r-1-i}{d}\binom{j}{t}\binom{s-r-1-j}{q}\binom{b+j-t}{p_1}\binom{d+s-r-j-q-1}{p_2}$$
$$\times\binom{n-p-i-j+t}{p_3}\binom{p-s+i+j+q}{p_4}\binom{p_1+p_2+1}{l_1}\binom{p_3+p_4+1}{l_2}$$
$$\times\frac{\theta^{p_1+p_3}\tau^{p_2+p_4}}{(1+\theta)^{b+n-p-i}(1+\tau)^{d-r+p+i+1}}$$
$$\times\int_0^\infty\int_x^\infty x^{k+l_1}y^{l+l_2}e^{-x\{\theta(b+j-t)+\tau(d+s-r-j+q)\}}e^{-y\{\theta(n-p-i-j+t)+\tau(p-s+i+j+q+1)\}}dxdy \qquad (12)$$

After simplifying equation (12) by evaluating the integrals using gamma function we get the required relation in equation (10).

□

Next, we have evaluated the covariance using the product moments in equation (10) with $n = 6$, $\theta = 1$ and $\tau = \frac{\theta}{h}$ with $p = 0, 1, 2$ and tabulated in Table 2 (Annexure). The results in Table 1 can be verified for the case $p = 0$ by using the well-known identity [see Arnold and Balakrishnan (1989), p.10]

$$\sum_{i=1}^{5}\sum_{j=i+1}^{6}\mu_{i,j:n} = \binom{6}{2}[E(X)]^2.$$

From Table 2, we see that the covariance increases as h increases while it decreases with the increase in p values. For small r and s, the relative change in covariances is more with the increase in number of outliers from $p = 1$ to $p = 2$ for all values of h and p.

4. Robustness of the L-Moments

In this section, we discuss the issue of robustness by estimating the bias and mean square error (MSE) of sample L-moments of the population L-moments for the distribution in equation (1) under various choices of n.

According to Hosking (1990), the L-moments are basically linear functions of the data and are more robust than the usual moments when outliers are present in the data. Also, sometimes these estimators produce efficient parameter estimators as compare to maximum likelihood estimates (MLEs).

Using the expression of the first four population L-moments $\lambda_1, \lambda_2, \lambda_3$ and λ_4 [see Hosking (1990), p. 107] for the distribution in equation (1) and using equation (7) we compute the values of all the first population L-moments for $p = 0$, 1 and 2 which is given in Table 3 (Annexure).

The expressions of the first four sample L-moments [see Hosking (1990), p. 113] are as follows:

$$l_1 = n^{-1}\sum_i x_i, \qquad (13)$$

$$l_2 = \frac{1}{2}\binom{n}{2}^{-1} \sum\sum_{i>j} (x_{i:n} - x_{j:n}), \tag{14}$$

$$l_3 = \frac{1}{3}\binom{n}{3}^{-1} \sum\sum_{i>j>k} \sum (x_{i:n} - 2x_{j:n} + x_{k:n}), \tag{15}$$

and $$l_4 = \frac{1}{4}\binom{n}{4}^{-1} \sum\sum\sum_{i>j>k>l} \sum (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}). \tag{16}$$

Using equations (14), (15) and (16) and the population L-moments from the Table 3 we have estimated the Bias and MSE of sample L-moments in Table 4, Table 5 and Table 6 (Annexure) for n = 10, 20 and 30 respectively. The random samples are simulated from Lindley distribution for $\theta = 1$ using the LindleyR package in R software. The bias and MSE are computed using R (based on 10,000 runs). The R code is not included but it is available upon request from the author.

From Table 4, Table 5 and Table 6 we see that

1) In general for most of the values of 'p' and 'h' the MSE decreases as 'n' increases.
2) When $p = 0$, l_4 has the smallest MSE among the three sample L-moments.
3) The values of bias and MSE gradually decrease with the increase in the order of the values of 'p' and 'h' *i.e.* the bias and MSE are having inverse relation with the order of the sample L-moments. Again, the relative change in bias and MSE gradually decreases with the increase in the order of the sample L-moment for different values of 'p'.

5. Special Cases

By substituting $p = 0$ in equation (7), it reduces to

$$\mu_{r:n}^{(k)} = \frac{\theta^2}{(1+\theta)^{n-r+1}} c_{r:n} \sum_{i=0}^{r-1} \sum_{l=0}^{n-r+i} \sum_{q=0}^{l+1} (-1)^i \frac{\theta^l}{(1+\theta)^i} \binom{r-1}{i} \binom{n-r+i}{l} \binom{l+1}{q} \frac{\Gamma(k+q+1)}{[\theta(n-r+i+1)]^{k+q+1}} \tag{17}$$

where

$$c_{r:n} = \frac{n!}{(r-1)!(n-r)!} \quad .$$

Again, replacing $p = 0$ in equation (10), we get

$$\mu_{r,s:n}^{(k,l)} = \frac{\theta^4}{(1+\theta)^{n-r+1}} c_{r,s:n} \sum_{b=0}^{r-1} \sum_{j=0}^{s-r-1} \sum_{p_1=0}^{b+s-r-1-j} \sum_{p_3=0}^{n+j-s} \sum_{l_1=0}^{p_1+1} \sum_{l_2=0}^{p_3+1} \sum_{a=0}^{l+l_2} (-1)^{b+j} \frac{\theta^{p_1+p_3}}{(1+\theta)^b} \binom{r-1}{b} \binom{s-r-1}{j}$$
$$\times \binom{b+s-r-j-1}{p_1} \binom{n+j-s}{p_3} \binom{p_1+1}{l_1} \binom{p_3+1}{l_2} \frac{(l+l_2)!}{a!}$$
$$\times \frac{\Gamma(k+l_1+q+1)}{[\theta(n-s+j+1)]^{l+l_2+1}[\theta(b+n-r+1)]^{k+l_1+a+1}} \tag{18}$$

where

$$c_{r,s:n} = \frac{n!}{(r-1)!(s-r-1)!\,(n-s)!}.$$

The results in equations (17) and (18) are the single and product moment of order statistics from one parameter Lindley distribution [see Sultan and AL-Thubyani (2016), p. 3 and p. 4] respectively.

6. Conclusion

In this article, we obtain the explicit form of single and product moments of order statistics from one parameter Lindley distribution when multiple outliers are present in the data. These moment relations are generalized form of the moments of order statistics of the Lindley distribution obtained by Sultan and AL-Thubyani (2016) under the multiple-outliers model set up.

The findings of the present study suggest that both the variances and covariances of order statistics of Lindley distribution decreases with the increase in the number of outliers present in the data. For small values of r (in case of single moment order statistics) and r and s (in case of product moment order statistics) the relative change in variances and covariances is comparatively more *i.e.* the smaller order statistics are more sensitive to the presence of outliers, as one would expect. While for higher values of r the variance is negatively correlated with h (scaling factor); the covariance remains positively correlated with h for all r and s, $r < s$. We also find that the bias and MSE of higher sample L-moments gradually reduced. The robustness feature of the sample L-moments is evident from the fact that the higher order sample L-moments provide more protection against the presence of pronounced outliers as the relative change in bias and MSE is reasonably less with the increase in number of outliers.

Acknowledgements

I am thankful to the anonymous reviewer for his constructive suggestions, which helped to improve the presentation of this article.

References

Arnold, B. C. and Balakrishnan, N. (1989). *Relations, Bounds and Approximations for Order Statistics. Lecture Notes in Statistics*, Vol. 53, Springer-Verlag, New York.

Bakouch, H. S., Al-Zahrani, B. M., Al-Shomarani, A. A., Marchi, V. A. A. and Louzada, F. (2012). An extended Lindley distribution. *Journal of the Korean Statistical Society*, **41**, 75-85.

Balakrishnan, N. (1994a). On order statistics from non-identical exponential random variables and some applications (with discussion). *Computational Statistics and Data Analysis*, **18**, 203-253.

Balakrishnan, N. (1994b). On order statistics from non-identical exponential random variables and some applications. *Communication in Statistics - Theory and Methods*, **23**, 3373-3393.

Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness. *Revista Mathematica Complutense*, **20**(**1**), 7-107.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. 3rd Edition, John Wiley & Sons, Chichester.

Bhar, L. M., Gupta, V. K. and Prasad, R. (2013). Detection of outliers in designed experiments in presence of masking. *Statistics and Applications*, **11**(**1&2**), 147-160.

Childs, A. (1996). *Advances in Statistical Inference and Outlier Related Issues*. Ph. D. Thesis, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario.

Ghitany, M. E., Atieh, B. and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**, 493-506.

Gomoz-Deniz E., and Calderin-Ojeda, E. (2011). The discrete Lindley distribution: properties and applications. *Journal of Statistical Computation and Simulation*, **81(11)**, 1405–1416.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society Series B (Methodological)*, **52**(**1**), 105-124.

Krishna, H. and Kumar, K. (2011). Reliability estimation in Lindley distribution with progressively type II right censored sample. *Mathematics and Computers in Simulation*, **82**(**2**), 281-294.

Kumar, C. S., and Jose, R. (2018). On double Lindley distribution and some of its properties. *American Journal of Mathematical and Management Sciences*, **38**(**1**), 23-43.

Mazucheli, J., Menezes, A. F. B. and Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, **46**, 700-714.

Mazuchelia, J. and Achcar, J. A. (2011). The Lindley distribution applied to competing risks lifetime data. *Computer Methods and Programs in Biomedicine*, **104**(**2**), 188-192.

Prasad, R., Nandi, P. K., Bhar, L. M. and Gupta, V. K. (2008). Outliers in multi-response experiments. *Statistics and Applications*, **6**(**1& 2**), 275-292.

Sultan, K. F. and Al-Thubyani, W. S. (2016). Higher order moments of order statistics from the Lindley distribution and associated inference. *Journal of Statistical Computation and Simulation*, **86**(**17**), 3432-3445.

Sultan, K. F. and Moshref, M. E. (2014). Moments of order statistics from Weibull distribution in the presence of multiple outliers. *Communication in Statistics -Theory and Methods*, **43**(**10-12**), 2214-2226.

ANNEXURE

Table 1: The means and variances in the presence of multiple outliers when $n = 6$

r	p	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
1	0	0.2997	0.0773								
2		0.6292	0.1637								
3		1.0116	0.2774								
4		1.4885	0.4568								
5		2.1597	0.8255								
6		3.4110	2.1743								
		$h = 0.1$		$h = 0.2$		$h = 0.3$		$h = 0.4$		$h = 0.5$	
1	1	0.0851	0.0070	0.1448	0.0198	0.1871	0.0322	0.2179	0.0428	0.2409	0.0516
2		0.3740	0.0993	0.4132	0.0962	0.4549	0.1011	0.4930	0.1103	0.5260	0.1207
3		0.7605	0.2303	0.7781	0.2209	0.8066	0.2155	0.8402	0.2169	0.8744	0.2234
4		1.2507	0.4228	1.2567	0.4158	1.2721	0.4055	1.2961	0.3984	1.3259	0.3973
5		1.9360	0.8025	1.9376	0.7992	1.9439	0.7900	1.9575	0.7775	1.9789	0.7670
6		3.2025	2.1669	3.2027	2.1660	3.2043	2.1608	3.2093	2.1482	3.2202	2.1289
1	2	0.0492	0.0023	0.0949	0.0087	0.1355	0.0173	0.1708	0.0270	0.2012	0.0367
2		0.1338	0.0096	0.2376	0.0286	0.3196	0.0491	0.3864	0.0688	0.4425	0.0873
3		0.4645	0.1431	0.5297	0.1320	0.6050	0.1363	0.6791	0.1507	0.7481	0.1699
4		0.9590	0.3618	0.9848	0.3410	1.0322	0.3230	1.0932	0.3178	1.1606	0.3254
5		1.6622	0.7626	1.6691	0.7500	1.6902	0.7249	1.7281	0.6990	1.7811	0.6832
6		2.9492	2.1505	2.9502	2.1465	2.9557	2.1296	2.9706	2.0958	2.9995	2.0508

** The results remain same for all values of h when $p = 0$.*

Table 2: The covariances in the presence of multiple outliers when $n = 6$

$p = 0$			$p = 1$					$p = 2$				
r	***s***	***h****	***h* = 0.1**	***h* = 0.2**	***h* = 0.3**	***h* = 0.4**	***h* = 0.5**	***h* = 0.1**	***h* =0.2**	***h* = 0.3**	***h* = 0.4**	***h* = 0.5**
1	2	0.0575	0.0199	0.0278	0.0331	0.0377	0.0418	0.0035	0.0101	0.0172	0.0240	0.0303
1	3	0.2005	0.0564	0.0877	0.1091	0.1266	0.1420	0.0194	0.0386	0.0590	0.0800	0.1012
2	3	0.1483	0.1006	0.1043	0.1074	0.1112	0.1162	0.0386	0.0540	0.0649	0.0756	0.0868
1	4	0.8749	0.1941	0.3036	0.3741	0.4308	0.4835	0.0465	0.0941	0.1473	0.2062	0.2695
2	4	0.6203	0.3654	0.3910	0.4159	0.4413	0.4685	0.1159	0.1887	0.2426	0.2916	0.3411
3	4	0.3949	0.3102	0.3120	0.3136	0.3165	0.3224	0.1996	0.2111	0.2208	0.2327	0.2490
1	5	4.7651	0.9196	1.4366	1.7330	1.9385	2.1201	0.1625	0.3160	0.4725	0.6448	0.8379
2	5	3.0990	1.6620	1.7596	1.8437	1.9184	1.9970	0.4198	0.6883	0.8702	1.0230	1.1769
3	5	1.9092	1.3646	1.3849	1.4162	1.4560	1.5054	0.7706	0.8494	0.9366	1.0291	1.1322
4	5	1.3947	1.1475	1.1493	1.1524	1.1577	1.1687	0.8573	0.8675	0.8815	0.9011	0.9323
1	6	34.584	6.1630	9.6334	11.450	12.436	13.080	0.8723	1.6678	2.3711	3.0249	3.6854
2	6	19.979	10.217	10.755	11.149	11.381	11.536	2.2758	3.7227	4.5735	5.1174	5.5537
3	6	11.168	7.5911	7.6829	7.8078	7.9318	8.0554	3.9772	4.3269	4.6847	5.0045	5.3116
4	6	7.7404	6.0517	6.0717	6.1271	6.2207	6.3521	4.2195	4.3019	4.4670	4.6992	4.9895
5	6	8.1472	6.9541	6.9568	6.9680	6.9931	7.0394	5.6128	5.6253	5.6636	5.7365	5.8549

** The results remain same for all values of h when $p = 0$.*

Table 3: First four Population L-Moments

		λ_1	λ_2	λ_3	λ_4
$p = 0$	h*	1.5000	0.6875	0.1921	0.0978
$p = 1$	$h = 0.1$	0.1090	0.7013	0.2133	0.0856
	$h = 0.2$	0.2333	0.6591	0.2243	0.0862
	$h = 0.3$	0.3692	0.6268	0.2240	0.0911
	$h = 0.4$	0.5142	0.6056	0.2172	0.0951
	$h = 0.5$	0.6666	0.5956	0.2078	0.0971
$p = 2$	$h = 0.1$		0.0543	0.4180	0.0213
	$h = 0.2$		0.1152	0.3448	0.0716
	$h = 0.3$		0.1806	0.2852	0.1741
	$h = 0.4$		0.2489	0.2401	0.1027
	$h = 0.5$		0.3194	0.2084	0.1018

Table 4: Bias and MSE of sample L-moments l_2, l_3 and l_4 for $n = 10$

		l_2		l_3		l_4	
p	h	Bias	MSE	Bias	MSE	Bias	MSE
0	h*	0.0027	0.0528	0.0018	0.0252	0.0018	0.0187
1	0.1	0.0099	0.0533	0.0135	0.0251	0.0069	0.0189
1	0.2	0.0242	0.0550	0.0219	0.0267	0.0086	0.0193
1	0.3	0.0503	0.0566	0.0209	0.0254	0.0043	0.0186
1	0.4	0.0652	0.0570	0.0197	0.0251	0.0025	0.0185
1	0.5	0.0738	0.0571	0.0101	0.0243	0.0001	0.0182
2	0.1	0.6129	0.4279	0.1982	0.0637	0.0674	0.0228
2	0.2	0.5510	0.3575	0.1272	0.0412	0.0157	0.0190
2	0.3	0.4696	0.2725	0.0724	0.0299	0.0799	0.0248
2	0.4	0.3982	0.2101	0.0349	0.0257	0.0081	0.0181
2	0.5	0.3288	0.1591	0.0068	0.0250	0.0035	0.0250

** The results remain same for all values of h when p = 0.*

Table 5: Bias and MSE of sample L-moments l_2, l_3 and l_4 for $n = 20$

		l_2		l_3		l_4	
p	*h*	Bias	MSE	Bias	MSE	Bias	MSE
0	*h**	0.0015	0.0254	0.0012	0.0107	0.0013	0.0040
1	0.1	0.0115	0.6898	0.0171	0.1965	0.0098	0.0041
1	0.2	0.0269	0.0266	0.0250	0.0115	0.0114	0.0042
1	0.3	0.0542	0.0282	0.0282	0.0113	0.0050	0.0041
1	0.4	0.0749	0.0309	0.0206	0.0114	0.0033	0.0041
1	0.5	0.0838	0.0325	0.0126	0.0109	0.0012	0.0040
2	0.1	0.6329	0.4264	0.2164	0.0580	0.0702	0.0089
2	0.2	0.5646	0.3443	0.1414	0.0310	0.0212	0.0044
2	0.3	0.4924	0.2679	0.0840	0.0178	0.0788	0.0101
2	0.4	0.4191	0.2011	0.0408	0.0122	0.0054	0.0041
2	0.5	0.3494	0.1461	0.0107	0.0106	0.0028	0.0041

***** *The results remain same for all values of h when p* = 0.

Table 6: Bias and MSE of sample L-moments l_2, l_3 and l_4 for $n = 30$

		l_2		l_3		l_4	
p	*h*	Bias	MSE	Bias	MSE	Bias	MSE
0	*h**	0.0006	0.0167	0.0009	0.0067	0.0005	0.0040
1	0.1	0.0147	0.0168	0.0197	0.0071	0.0106	0.0041
1	0.2	0.0258	0.0171	0.0288	0.0076	0.0107	0.0042
1	0.3	0.0552	0.0193	0.0294	0.0078	0.0054	0.0041
1	0.4	0.0766	0.0223	0.0230	0.0074	0.0024	0.0040
1	0.5	0.0851	0.0238	0.0139	0.0069	0.0011	0.0039
2	0.1	0.6345	0.4192	0.2202	0.0553	0.0723	0.0091
2	0.2	0.5647	0.3356	0.1467	0.0283	0.0227	0.0045
2	0.3	0.4954	0.2621	0.0881	0.0144	0.0789	0.0101
2	0.4	0.4261	0.1981	0.0425	0.0088	0.0947	0.0130
2	0.5	0.3558	0.1432	0.0127	0.0070	0.0031	0.0041

***** *The results remain same for all values of h when p* = 0.

Statistics and Application {ISSN 2452-7395(online)}
Vol. 18, No. 1, 2020 (New Series), pp 141-156

Statistical Properties of Sampling Distributions of Different Test Statistics for Different Measures of Change and a New Test: Simulation Study

Handan Ankarali [1], Sengul Cangur [2] and Seyit Ankarali [3]
[1]*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Istanbul Medeniyet University, 34700, Istanbul, Turkey*
[2]*Department of Biostatistics and Medical Informatics, Faculty of Medicine, Duzce University, 81620, Duzce, Turkey*
[3]*Department of Physiology, Faculty of Medicine, Istanbul Medeniyet University, 34700, Istanbul, Turkey*

Received: 04 November 2019; Revised: 18 May 2020; Accepted: 30 May 2020

Abstract

The aim of this study is to analyse the distribution characteristics of four different test statistics, namely the Mean/Standard Error of Mean (Mean/SEmean), Median/Interquartile Range (Median/IQR), Trimmed Mean/Standard Error of Mean (TrMean/SEmean), and Trimmed Mean/Interquartile Range (TrMean/IQR), which can be used to test two measures of change, namely percent change (PC) and modified symmetrised percent change (MSPC). To ensure the selection of suitable test statistics using the two measures of change, the observed type-I errors and powers of the test statistics have been computed. Results demonstrate that the sampling distributions of the four different test statistics by using PC values exhibit skewness. The Mean/SEmean statistic for the MSPC measure exhibits a two-peak value and platykurtic distribution, while the TrMean/SEmean statistic shows a leptokurtic distribution. The Median/IQR test yields robust and powerful results, especially for large sample sizes. This new statistical measure is referred to as the *HS* test.

Key words: Pre-post designs; Paired samples t-test; Change measures; Per cent change; Symmetrised per cent change; Trimmed mean.

1. Introduction

A one-sample paired design is commonly used to assess treatment effects in clinical research. In this design, the measure of the change in continuous data is based on simple difference, relative change or the ratio of post-treatment to baseline. The appropriate measure of change is selected according to the ease of interpretation and ease of data analysis generally. In this process, it is necessary to take into account both in terms of clinical significance and statistical features. The clinical perspective is based on the ease of interpretation, while the statistical perspective comes from the ease of conducting data analysis based on the normal distribution. (Tornqvist *et al.*, 1985; Zhang and Han, 2009; Yamabe *et al.*, 2012). In some cases, it is not easy to decide which measure of change to use, as simple difference may not always identify the change correctly. For example, when the difference observed in an individual with a large initial value is the same as the difference observed in an individual with a small initial value, this difference does not have the same

Corresponding Author: S. Cangur
e-mail: sengulcangur@duzce.edu.tr

meaning biologically. Suppose that a particular drug, which is being evaluated to determine its effectiveness at reducing facial acne, has lowered the number of acne sites from 10 to 0 in one person and from 20 to 10 in another person. If only the difference is taken into consideration, it could be concluded that the drug has the same effect on both persons; however, the clinician would explain that this numerical change does not reflect the actual biological change, and that a more clinically meaningful change has occurred in the first person. Moreover, Waleekhachonloet *et al*. (2007) stated in their study that diet effect is the difference in a person's weight before and after the diet. In the work of these authors, treatment effect is generally expressed as simple difference. This is because the statistical properties of the distribution of simple difference are well known and can be tested using a t-test, which produces strong results. For their part, in a dietary trial with similar content, Kim *et al.* (2009) expressed weight loss as a percent change. In similar clinical studies, researchers argue that percent change better reflects the treatment effect from a clinical perspective compared to simple difference and that its clinical meaning is more accurate. In addition, individuals having large initial values may have large differences, while individuals having very small initial values may have large post/pre ratios. In the literature, various measures of change have been proposed with the aim of eliminating the influence of these differences in individuals' initial values on pre-post change values. There is no in-depth research or consensus on these measures of change (Berry and Ayers, 2006; Wilkinson, 1999). Therefore, it is necessary to examine the characteristics and performances of different measures of change in more detail.

The percent change (PC) and ratio (R) measures are tested with non-parametric tests due to the fact that they often exhibit a highly skewed distribution (Vickers, 2001; Koti, 2001). The other measure of change is symmetrised percent change (SPC) with good statistical properties (Berry and Ayers, 2006; Yamabe *et al*., 2012). However, due to the difficulty associated with interpreting SPC, a very limited amount of research on the matter has been carried out (Yamabe *et al.,* 2012; Berry and Ayers, 2006; Koti, 2001). Another measure of change is modified symmetrised percent change (MSPC) with good statistical properties. MSPC is defined as the mean of two values for a numerator. When we regard the difference in denominator as an index of variation, the measure may be considered as a variation standardised by mean, such as a coefficient of variation (Yamabe *et al.,* 2012). This measure is calculated by eliminating the effect of pre-post mean in the difference between the pre-post measures. This measure, which is similar to the coefficient of variation, is easy to interpret (Yamabe *et al.,* 2012).

The aims of this study are as follows:

a) To test the significance of the two measures of change, PC and MSPC, with four test statistics;
b) To obtain the sampling distributions and 95% confidence limits (critical tail values of the distributions) of these statistics;
c) To calculate the observed type-I errors and power of four tests of each measure of change.

2. Methods

2.1. Measures of Change used in the study

The three most frequently used measures of change in medical research are investigated in this study. The formulas for these measures are provided below.

I. Simple Difference (D) = $(Pre - Post)$

- This measure is commonly used and understood in the literature. It is used in this study for the validation of simulation data, as well as to compare the type I error and power of the t-test calculated for D with other measures' results.

II. Percent Change (PC) = $\frac{(Pre-Post)}{Pre} \times 100$ or PC= $\frac{(Post-Pre)}{Pre} \times 100$

- Although pre- and post- data have the same units, PC values are often unitless or expressed as percentages. This measure means "the proportion of increase (or decrease) for pre-value", and is preferable from a clinical perspective due to its ease of interpretation. However, it exhibits a positive skew when post-data is much larger than pre-data, even if pre- and post-data are normal.

III. Modified Symmetrised Percent Change (MSPC) = $\frac{(Pre-Post)}{\frac{(Pre+Post)}{2}} \times 100$

- Direct interpretation of the SPC value is difficult, but the distribution of SPC maintains symmetry without regard to shape, scale and correlation parameters of the distribution for pre- and post-data (Yamabe *et al.,* 2012). The numerator of the MSPC measure is pre-post difference, while its denominator is a mean of the pre and post values such as a coefficient of variation. This measure is called variability (%) in the bioanalytical field and is used to evaluate the level of reproducibility of assay results using incurred samples (Mario *et al.*, 2007; Douglas *et al.*, 2009). When SPC or MSPC are used to analyse the data, the results can be interpreted after transforming to the robust percent change (RPC) proposed by Berry (1989).

2.2. Test statistics used in the study

Four statistics are investigated in the hypothesis test for the two different measures of change (PC, MSPC) used in the present study. These statistics are as follows:

- Mean/Standard Error of Mean (Mean/SEmean)
- Median/Interquartile Range (Median/IQR)
- Trimmed Mean/Standard Error of Mean (TrMean/SEmean)
- Trimmed Mean/Interquartile Range (TrMean/IQR)

The empirical sampling distributions of these statistics have been obtained and the statistical properties of these distributions determined.

Trimmed mean is average of the remaining values after removing the smallest p% and the largest p% of the values. We have accepted p as equal to 10% in our calculations.

2.3. Simulation study

The present research employs a quasi-Monte Carlo simulation study. The data are generated from a bivariate standard normal distribution. The random variables X_i (i =1,2) denote the response of pre- and post-data following a bivariate standard normal distribution; moreover, the variables satisfy $(X_1, X_2) \sim BN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, where μ_i denotes the location parameters, σ_i the scale parameters, and ρ the correlation parameter between two random variables of pre- and post-data. For the observed type-I error, we generate (X_1, X_2) from$\sim BN(0, 0, 1, 1, \rho)$, while for the power of the tests, we generate (X_1, X_2) from$\sim BN(0, \mu_2[0.5, 1.0, 1.5, 2.0], 1, 1, \rho)$. A total of 18 conditions are evaluated considering six different sample sizes (10, 30, 60, 120, 500, 10,000) and three different correlations between the variates named as pre- and post-measures (0.3, 0.6, 0.9). Each condition is repeated 60,000 times.

In the first step, the results of the Mean/SEmean statistic for D are examined. The purpose of this step is to validate the results produced by the simulation data. The probability of type-I error and power values of the Mean/SEmean statistic are used for this. These values are calculated considering the critical values of the t-distribution at the 5% level (see: Table 3). In the second step, empirical sampling distributions are obtained for the four statistics with regard to PC and MSCP, and the critical tail values of these distributions are determined at the 5% level (see: Figure 2-3 and Table 1-2). In the third step, the observed probability of the type-I error and the power of the Mean/SEmean statistic for D are calculated according to the critical values of the t-distribution in Table 3. In addition, the type-I error and the power values of the Median/IQR and TrMean/IQR statistics selected for MSPC are calculated using the critical values of the empirical sampling distributions in Table 4-5. Subsequently, the results of the selected tests are compared with each other (see: Table 6).

The histograms of the test values from 60,000 trials for each condition have plotted. The lower and upper limit values of the 95% confidence intervals of these distributions are presented in the tables below. In addition, the skewness and kurtosis coefficients are calculated. The dataset used in this phase is referred to as the learning sample. The observed probability of type-I error and the power values of the selected tests are calculated from 40,000 trials by resampling from the bivariate distribution. This dataset is referred to as the test sample. To calculate the power of the tests, the standardised differences between the populations named as pre- and post- are accepted as 0.5SD, 1SD, 2SD and 2.5SD respectively.

The FORTRAN programming language, along with Microsoft Power Station Developer Studio, IMSL Library and the Minitab program (version 16.0) are used in the simulations and to obtain the graphics.

3. Results

3.1. Validation of simulation

In order to carry out the validation of the simulation program, samples (10,000 observations) are taken from a population with a bivariate normal distribution. There is a correlation of 0.60 among them, and differences have been calculated for each sample. The Mean/SEmean test statistic has been used for the differences It has been determined that the distribution of this statistic exhibits a t-distribution; this has been done by calculating the

descriptive statistics and plotting the shape of the sampling distribution following 60,000 trials. The mean of the simple difference is 0.0003±0.9911 and its first, second and third quartiles are -0.6733, -0.0011 and 0.6526 respectively. The mean of the test statistic is close to zero, while the standard deviation is close to one (-0.0042±1.0000), and its first, second and third quartiles are -0.6761, -0.0019 and 0.6753 respectively. Furthermore, when n = 10,000, the 95% confidence limits of this distribution are observed to be ±1.96 (Figure 1). All these results show that the simulation program produces reliable and accurate results.

The PC and MSPC measures have been simultaneously calculated in the validated simulation. In the hypothesis test established for these measures, the sampling distributions at the end of 60,000 trials have been obtained for the four test statistics defined in the methods section.

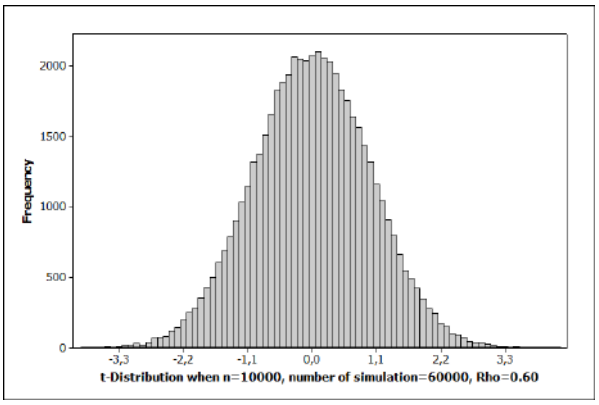


Figure 1: Empirical sampling distributions of Mean/SEmean statistics when n=10,000, number of simulation is 60,000 and Rho=0.60 (Close to standard normal distribution)

3.2. Results of percent change (PC)

The sampling distributions of the four statistics for PC in different conditions are presented in the Annexure A as Figure 2. Sampling distributions of Mean/SEmean statistics are skewed to the left and have two peak values (Figure 2a). The sampling distributions of the TrMean/SEmean statistics are skewed to the left but have only a single peak value (Figure 2b).

The sampling distributions of the Median/IQR and TrMean/IQR statistics have been found to have a single peak and to be approximately symmetrical for all correlation coefficients in large sample sizes. However, the sampling distributions obtained with small sample sizes exhibit a negative skew (Figures 2c and 2d).

The 95% confidence intervals of the sampling distributions of the four test statistics are considered as critical values; these values are presented in Annexure B, Table 1 for all conditions. It has been determined that the sampling distributions of Mean/SEmean and TrMean/SEmean statistics do not have symmetrical distributions under any conditions, while the sampling distributions of the Median/IQR and TrMean/IQR statistics are approximately symmetrical in population (n = 10,000).

3.3. Results of modified percent change (MSPC)

The difference between the pre-post measures has been expressed as MSPC. When the null hypothesis establishing that this difference is not significant has been tested with the

Mean/SEmean statistic, the lower and upper limit values of the test values calculated for the sampling distribution have found to be ±1.887 for n = 10 with 2.5% error probability for each tail (5% in total). These values have been determined to be ±1.710 when n = 10,000. The distribution of this statistic has been found to be symmetrical, but with two peaks. The distribution of the TrMean/SEmean statistic has been found to be symmetrical with a single peak value. The values of this distribution have been calculated as ±1.622 for n = 10 and ±0.081 for n = 10,000 with 2.5% type-I errors.

When compared with the first test statistic, it can be seen that the values of variance have a narrower range; in other words, the variance of the sampling distribution of this test is smaller. The shapes of the distributions of the Median/IQR and TrMean/IQR statistics have been also found to be perfectly symmetrical and single-peaked. It has been determined that as the sample size increases, the variances of the distributions decrease and the values of the ranges become narrower than is the case for the other two statistics. The upper and lower limits of the 95% confidence interval of the sampling distribution have been found to be ±0.430 for n = 10, ±0.015 for n = 10,000 for the Median/IQR statistic and ±0.470 for n = 10 and ±0.029 for n = 10,000 for the TrMean/IQR statistic. All these results demonstrate that the sampling distributions of the four test statistics are not t-distributions (Annexure B, Table 2).

The skewness values of the four sampling distributions in all conditions are close to zero. The kurtosis values of the Mean/SEmean statistic range from –0.5 to –1, depending on the sample size, and it has a platykurtic distribution (Figure 3a). The kurtosis values of the TrMean/SEmean statistic are large and it has positive values between 2 and 6.5 with increasing sample size (Figure 3b). However, the kurtosis values of the Median/IQR and TrMean/IQR statistics are very close to zero. These two distributions are also symmetrical (Figure 3c and 3d).

3.4. Suitable test statistics, observed type-I error and power

Since the Median/IQR and TrMean/IQR test statistics using MSPC exhibit a single peak and a perfectly symmetrical distribution, the critical values reported in Table 2 are used to calculate type-I errors and test powers.

After the critical values have been obtained, 40,000 new samples for each condition in the simulation have produced; these samples are named the test samples. Firstly, a t-test is used for D, and the type-I errors and powers of the test for each simulation condition are calculated as in Annexure B, Table 3. These calculations are made to validate the results of the simulated data. According to Table 3, the probabilities of the observed type-I error of the t-test are around 5%, and the deviation from 5% is not significant. When the sample size is around 10, it is found that the t-test has a low power, and it is also determined that this value increases slightly as the effect size increases. Furthermore, it is determined that the t-test yields fairly strong results in other sample sizes when the effect size is 0.5 standard deviations; moreover, when the effect size increases to 1 standard deviation and above, the power values become 100% or very close to that. Since all these results are known and expected results for the t-test, it can be determined that the simulation program produces valid data.

The majority of the probabilities of making a type-I error for the Median/IQR test statistic calculated for 18 different conditions are found to be very close to 5%, with the lowest being 4.80% and the highest being 5.63% (Annexure A, Table 4). The results of the

TrMean/IQR statistic are also found to be similar: most of the values are close to 5%, with the lowest being 4.63% and the highest being 5.24% (Annexure A, Table 5).

The Median/IQR and TrMean/IQR statistics are found to have high power only for very large sample sizes with an effect size of 0.5SD. While the Median/IQR statistic has sufficiently high power when the effect size is 1SD and $n > 60$, the TrMean/IQR statistic reaches an adequate level of power when the effect size is 1SD and $n \geq 500$. Moreover, the Median/IQR statistic reaches the desired level of power with an effect size of 1.5SD when $n \geq 30$, while the TrMean/IQR statistic reaches the desired level of power with the same effect size when $n \geq 120$. Finally, it is determined that the Median/IQR and TrMean/IQR statistics have the desired level of power with an effect size of 2SD when $n \geq 30$, and that the Median/IQR statistic reaches 100% power level when $n \geq 30$ (see Tables 4 and 5). In light of these results, the two new tests proposed using MSPC can be identified as robust tests like the well-known *t*-test.

In particular, it can be concluded that the Median/IQR test yields powerful results when evaluating differences that show deviations of 1SD or larger, while the TrMean/IQR test produces powerful results in cases where the differences are 1.5SD or larger. Since, in many conditions, the Median/IQR test is more powerful than the TrMean/IQR test, it can be concluded that the Median/IQR test could be used as an alternative test to the paired samples *t*-test in some conditions (Annexure B, Table 6). The Median/IQR test statistic can be referred to as the *HS* test, as it has been used for the first time in this study.

4. Discussion

Ankarali and Ankarali (2009) applied only the Mean/SEmean statistic for PC values in their study and found that the test power was very low when they tested the results with *t*-distribution. In light of these authors' simulation results, Mean/SEmean, Median/IQR, TrMean/SEmean and TrMean/IQR test statistics have been investigated for some measures of change.

It is known that the sampling distribution of Mean/SEmean statistic for D is the normal distribution. However, the sampling distributions of this statistic for PC and MSPC are not normal. In addition, it has been observed that the distribution of TrMean/SEmean statistics does not converge to a symmetrical or normal distribution under any conditions. Furthermore, the sampling distributions of the Median/IQR and TrMean/IQR statistics for PC have nearly symmetrical distributions at very large sample sizes, but exhibit a negatively skewed distribution under other conditions. For researchers, if it is more meaningful to use MSCP rather than D when defining a change, the *HS* test proposed in this study can be used when the sample size is large ($n > 100$) and the effect size is higher than 1.5. In other cases, however, Mean/SEmean statistics should be used for D (Table 6). In addition, no suitable statistics have been proposed for PC; this measure of change can be used to interpret the results descriptively.

In many clinical studies, percent change is used to assess treatment efficacy (Curran-Everett and Williams, 2015; Reuter *et al.,* 2012). According to Tornqvist *et al.* (1985), the fact that percent change is independent of the unit of measurement represents an advantage. Unlike simple difference, moreover, percent change is also an effect size that can be easily understood and interpreted by everyone. However, the authors provided no information as to which measure should be used to determine treatment effect. Some researchers would suggest

computing the percent change for each observation, then descriptively reporting the median while reporting no inferential statistics for percent change (Vickers, 2001).

The sampling distributions of the four statistics for the MSPC measure have been found to be symmetrical. However, the Mean/SEmean or TrMean/SEmean statistics are found to be unsuitable for testing the modified symmetrised percent change values; this is because the Mean/SEmean statistic has two peak values and a platykurtic distribution, as in Figure 3a, while the TrMean/SEmean statistic exhibits a leptokurtic distribution, as in Figure 3b. However, it has been concluded that the Median/IQR and Trmean/IQR statistics cannot be tested using a *t*-test, because while their distributions are symmetrical, they are not normal.

When the literature on this subject is examined, it can be seen that the majority of research focuses on how to evaluate the pre-post difference in parallel designs. The most widely preferred model is the ANCOVA model, in which the pre-values are covariate variables and post-values are response variables. However, it is also stated that the power of the test will be low if PC, R and SPC are used as the treatment effect. The aim of such studies is to compare several independent treatment groups in terms of treatment effect (Vickers, 2001; Curran-Everett and Williams, 2015). This study examined how the treatment effect should be expressed in the case of a comparison of dependent measure values obtained before and after treatment in an experimental design with only one group.

5. Conclusion

PC, R, SPC and MSPC measures should be preferred when the variables being studied, such as seizure frequency or laboratory tests, exhibit large intra- and inter-subject variability and a skewed distribution for pre- and post- data. In particular, PC and R can be used for easy interpretation in a clinical context. Statistical analysis based on the parametric tests is not recommended for PC, because PC (or the ratio of two values) will not be normal even if the pre- and post-data are normal (Yamabe *et al.* 2012). For some examples of suitable clinical evaluation, PC can be applied to the treatment evaluation of patients with high-density lipoprotein cholesterol (Adachi *et al.,* 2009), of patients with urge to urinate or urge incontinence based on the number of acraturesis (Homma *et al.* 2003), or of patients with climacteric disorder based on the number of hot flushes (Endrikat *et al.*, 2007).

Statistically, the properties of the SPC and MSPC measures are superior to PC and R. SPC and MSPC do not produce overly large values and show robustness to outliers on the same data analysis. Berry (1989) introduced SPC as the modified percent change with good statistical properties in the medical field. Brouwers and Mohr (1989) argued that the advantage of using SPC over PC is that the transformed variable does not depend on the denominator used in the transformation and the resultant distribution is symmetrical about its mean. However, the interpretation of SPC may not be intuitive for those accustomed to thinking in terms of PC. To improve the interpretability of analysis results, Berry (1989) suggested transforming SPC to the PC scale using the inverse transformation and then the obtained robust percent change (RPC) should be interpreted. SPC is applied to the treatment evaluation of patients with partial epilepsy based on the seizure frequency (Yamauchi *et al.,* 2006) and the evaluation of male patients with osteoporotic fracture based on physical activity (Janney *et al.,* 2010).

According to this study, the statistics TrMean/IQR and Median/IQR (*HS* test) perform well in terms of power for MSCP, especially for higher sample sizes. However, the choice between the change measures D and MSPC will be based on various factors, including the

measurement units, range of data, interpretation of the change measures, etc. The appropriate measure should be selected by striking a balance between both the clinical and statistical points of view. Accordingly, it is safe to conclude that the two test statistics work well using the MSPC measure. When MSPC serves as a more appropriate measure of changes to describe the effect, while the sample size (>120) and effect size (>1.5) are also high, the *HS* test for MSPC should be preffered over a *t*-test for D measure.

It is suggested that similar studies should be carried out for designs involving more than two repeated measures in future studies.

References

Adachi, H., Imaizumi, T., Murakami, M. and Abe, M. (2009). A phase III, randomized, parallel-group comparative study of Caduet (an amlodipine/atorvastatin combination drug) in patients with concurrent hypertension and hyper-LDL-cholesteremia. *Journal of New Remedies and Clinics*, **58**(**9**), 1496-1510.

Ankarali, H. C. and Ankarali, S. (2009). Which measure should be used for testing in a paired design, Simple difference, percent change, or symmetrized percent change? *Communications in Statistics - Simulation and Computation*, **38**(**1**), 402-415.

Berry, D. A. (1989). *Statistical Methodology in the Pharmaceutical Sciences*. Marcel Dekker, New York.

Berry, D. A. and Ayers, G. D. (2006). Symmetrized percent change for treatment comparisons. *The American Statistician*, **60**(**1**), 27-31.

Brouwers, P. and Mohr, E. (1989). A metric for the evaluation of change in clinical trials. *Clinical Neuropharmacology*, **12**,129-133

Curran-Everett, D. and Williams, C. L. (2015). Explorations in statistics, the analysis of change. *Advances in Physiology Education*, **39**(**2**), 49-54.

Endrikat, J., Graeser T., Mellinger U., Ertan, K. and Holz, C. (2007). A multicenter, prospective, randomized, double-blind, placebo-controlled study to investigate the efficacy of a continuous-combined hormone therapy preparation containing 1mg estradiol valerate/2 mg dienogest on hot flushes in postmenopausal women. *Maturitas*, **58**, 201-207.

Homma, Y., Paick, J. S., Lee J. G. and Kawabe K. On behalf of the Japanese and Korean Tolterodine Study Group (2003). Clinical efficacy and tolerability of extended-release tolterodine and immediate-release oxybutynin in Japanese and Korean patients with an overactive bladder: a randomized, placebo-controlled trial. *BJU International*, **92**(**7**), 741-747.

Janney, C. A., Cauley, J. A., Cawthon, P. M. and Kriska, A. M. (2010). Longitudinal physical activity changes in older men in the osteoporotic fractures in men study. *Journal of the American Geriatrics Society*, **58**(**6**), 1128-1133.

Kim, M. K., Tanaka, K., Kim, M. J., Matuso, T., Endo, T., *et al.* (2009). Comparison of epicardial, abdominal and regional fat compartments in response to weight loss. *Nutrition, Metabolism and Cardiovascular Diseases*, **19**(**11**), 760-766.

Koti, K. M. (2001). On a primary efficacy endpoint. *Drug Information Journal*, **35**(**1**), 157-162.

Rocci, M. L. Jr., Devanarayan, V., Haughey, D. B. and Jardieu, P. (2007). Confirmatory reanalysis of incurred bioanalytical samples. *The American Association of Pharmaceutical Scientists Journal*, **9**(**3**), E336–E343.

Reuter, M., Schmansky, N. J., Rosas, H. D. and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, **61**(**4**), 1402-1418.

Tornqvist, L., Vartia, P. and Vartia, Y. O. (1985). How should relative changes be measured? *The American Statistician*, **39**(**1**), 43-46.

Tsounis, S., Kimiskidis, V. K., Kazis, D., Gkiatas, K., Garganis, K., *et al.* (2011). An open-label, add-on study of pregabalin in patients with partial seizures, a multicenter trial in Greece. *Seizure*, **20**(**9**), 701-705.

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient, a simulation study. *BMC Medical Research Methodology*, **1**, 6.

Waleekhachonloet, O. A., Limwattananon, C., Limwattananon, S. and Gross, C. R. (2007). Group behavior therapy versus individual behavior therapy for healthy dieting and weight control management in overweight and obese women living in rural community. *Obesity Research and Clinical Practice*, **1**(**4**), 223-232.

Wilkinson, L. and Task Force on Statistical Inference, American Psychological Association, Science Directorate (1999). Statistical methods in psychology journals, Guidelines and explanations. *American Psychologist*, **54**(**8**), 594-604.

Yamabe, T., Maruo, K., Shirahata, S. and Goto, M. (2012). Statistical properties of two ratio measures based on pre- and post observed values which are assumed as bivariate power normal distribution. *Ouyou Toukeigaku*, **41**(**1**), 53-68.

Yamauchi, T., Kaneko, S., Yagi, K. and Sase, S. (2006). Treatment of partial seizures with gabapentin: Double-blind, placebo-controlled, parallel-group study. *Psychiatry and Clinical Neurosciences*, **60**, 507-515.

Zhang, L. and Han, K. (2009). How to analyze change from baseline, absolute or percentage change? D-level *Essay in Statistics 2009*. Dalarna University, Borlänge. http,//docplayer.net/20733008-How-to-analyze-change-from-baseline-absolute-or-percentage-change.html. Accessed March 2018.

ANNEXURE A

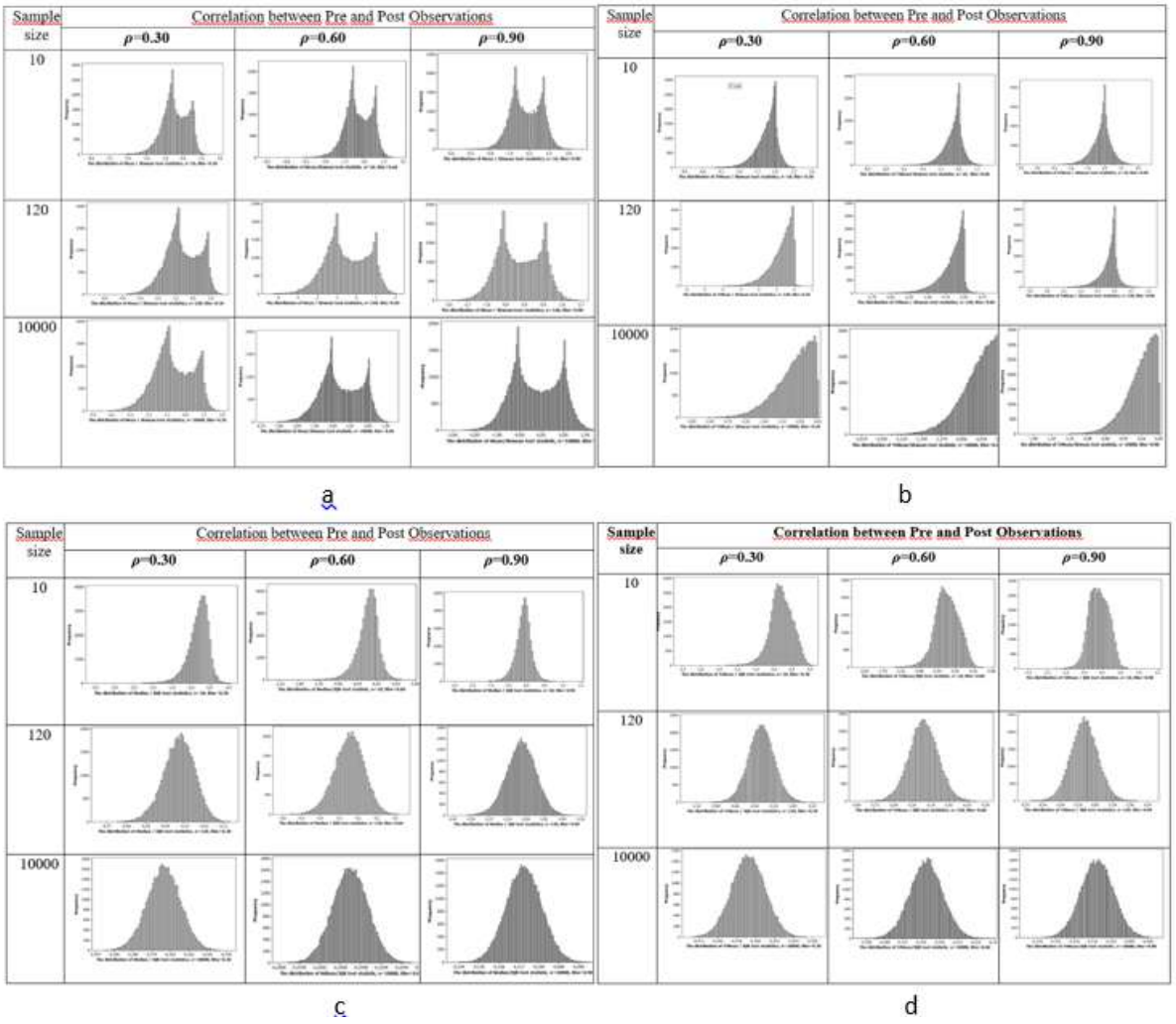


Figure 2: Empirical sampling distributions of a) Mean/SEmean statistics, b) TrMean/SEmean statistics, c) Median/IQR statistics and d) TrMean/IQR statistics for using PC

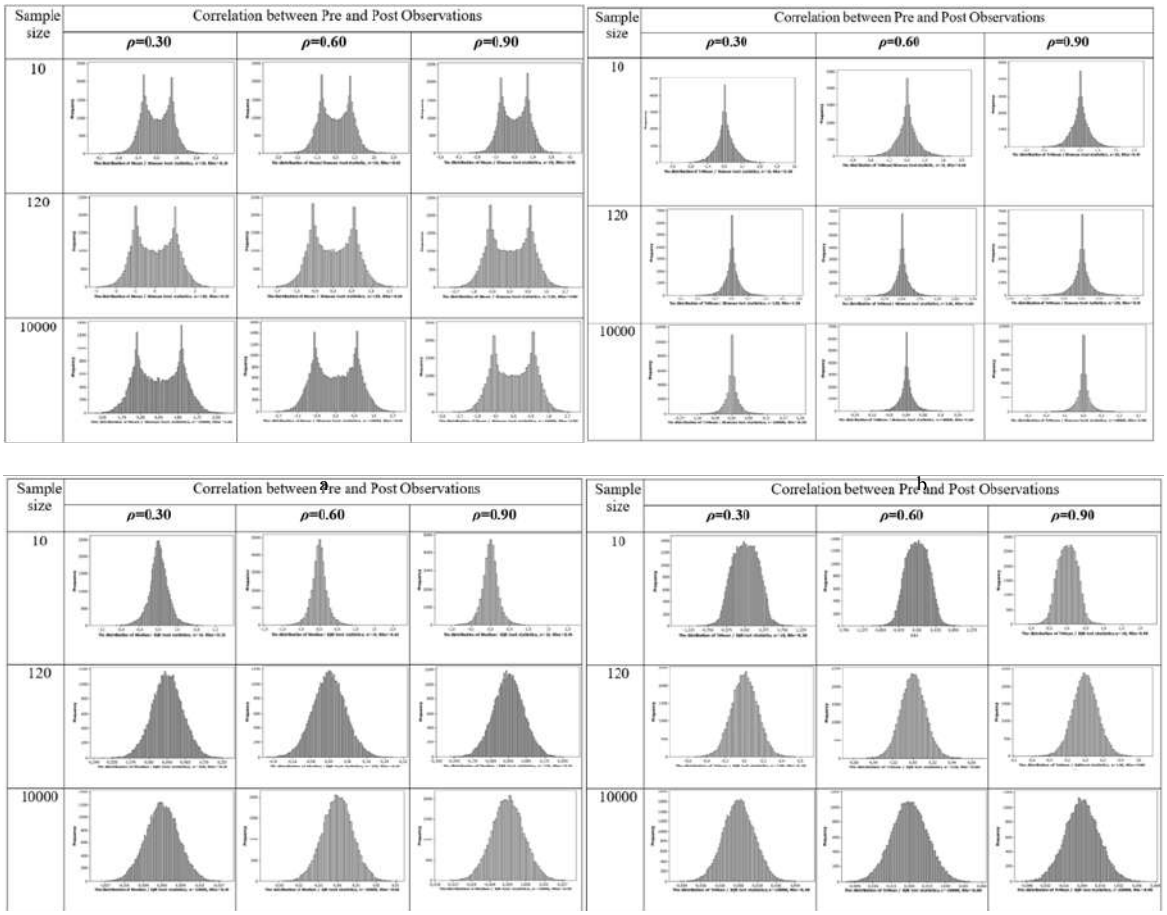


Figure 3: Empirical sampling distribution of (a) Mean/SEmean statistics, (b) TrMean/SEmean statistics, c) Median/IQR statistics and d) TrMean/IQR statistics for MSPC

ANNEXURE B

Table 1: Critical values for 95% confidence level of sampling distributions of four test statistics for PC

n	ρ	Percent Change ($PC = \left(\frac{Post-Pre}{Pre}\right) x100$)							
		Mean/SEmean		TrMean/SEmean		Median/IQR		TrMean/IQR	
		$S1L_\alpha$	$S1R_\alpha$	$S2L_\alpha$	$S2R_\alpha$	$S3L_\alpha$	$S3R_\alpha$	$S4L_\alpha$	$S4R_\alpha$
10	0.30	-3.376	1.207	-3.238	0.691	-0.952	0.112	-0.959	0.257
	0.60	-2.815	1.344	-2.657	0.911	-0.770	0.201	-0.784	0.309
	0.90	-2.232	1.593	-2.018	1.242	-0.568	0.317	-0.593	0.386
20	0.30	-3.166	1.181	-2.981	0.596	-0.820	0.002	-1.176	0.455
	0.60	-2.615	1.314	-2.376	0.788	-0.653	0.092	-1.042	0.547
	0.90	-2.131	1.515	-1.844	1.086	-0.469	0.208	-0.892	0.666
60	0.30	-3.011	1.168	-2.580	0.067	-0.622	-0.135	-0.788	0.063
	0.60	-2.517	1.287	-1.988	0.262	-0.479	-0.041	-0.653	0.167
	0.90	-2.050	1.480	-1.412	0.586	-0.320	0.081	-0.512	0.291
120	0.30	-2.981	1.147	-2.396	-0.014	-0.548	-0.201	-0.655	-0.071
	0.60	-2.527	1.279	-1.816	0.033	-0.410	-0.101	-0.532	0.031
	0.90	-2.031	1.484	-1.173	0.335	-0.260	0.025	-0.388	0.162
500	0.30	-2.960	1.146	-2.158	-0.028	-0.453	-0.284	-0.505	-0.226
	0.60	-2.495	1.289	-1.532	-0.017	-0.327	-0.174	-0.384	-0.113
	0.90	-2.034	1.489	-0.853	0.024	-0.185	-0.044	-0.246	0.019
10000	0.30	-2.952	1.145	-2.068	-0.030	-0.386	-0.348	-0.398	-0.336
	0.60	-2.488	1.272	-1.421	-0.0183	-0.267	-0.233	-0.280	-0.220
	0.90	-2.015	1.473	-0.659	-0.0107	-0.130	-0.099	-0.144	-0.086

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, $S1L_\alpha$ and $S1R_\alpha$: Left and Right tail critical values for Mean/SEmean; $S2L_\alpha$ and $S2R_\alpha$: Left and Right tail critical tables value for TrMean/SEmean; $S3L_\alpha$ and $S3R_\alpha$: Left and Right tail critical tables value for Median/IQR; $S4L_\alpha$ and $S4R_\alpha$: Left and Right tail critical table values for TrMean/IQR

Table 2: Critical values for 95% confidence level of sampling distributions of four test statistics for MSPC

n	*ρ*	Modified Percent Change ($MSPC = \left(\frac{Post-Pre}{\frac{Pre+Post}{2}}\right) x100$)							
		Mean/SEmean		TrMean/SEmean		Median/IQR		TrMean/IQR	
		$S1L_\alpha$	$S1R_\alpha$	$S2L_\alpha$	$S2R_\alpha$	$S3L_\alpha$	$S3R_\alpha$	$S4L_\alpha$	$S4R_\alpha$
10	0.30	–1.887	1.877	–1.622	1.622	–0.430	0.430	–0.470	0.470
	0.60	–1.886	1.884	–1.609	1.602	–0.434	0.434	–0.473	0.473
	0.90	–1.884	1.864	–1.600	1.600	–0.435	0.435	–0.476	0.476
30	0.30	–1.765	1.765	–1.180	1.180	–0.274	0.274	–0.470	0.470
	0.60	–1.760	1.760	–1.158	1.158	–0.273	0.273	–0.480	0.480
	0.90	–1.758	1.758	–1.170	1.170	–0.272	0.272	-0.482	0.482
60	0.30	–1.737	1.736	–0.968	0.954	–0.195	0.195	–0.395	0.395
	0.60	–1.740	1.726	–0.978	0.961	–0.196	0.196	–0.398	0.398
	0.90	–1.733	1.725	–0.969	0.973	–0.196	0.196	–0.400	0.400
120	0.30	–1.730	1.736	–0.698	0.693	–0.138	0.139	–0.271	0.271
	0.60	–1.736	1.728	–0.721	0.721	–0.139	0.139	–0.276	0.276
	0.90	–1.721	1.717	–0.715	0.717	–0.140	0.140	–0.277	0.277
500	0.30	–1.704	1.735	–0.358	0.363	–0.068	0.068	–0.130	0.130
	0.60	–1.724	1.710	–0.355	0.357	–0.068	0.068	–0.129	0.129
	0.90	–1.727	1.728	–0.351	0.354	–0.068	0.068	–0.131	0.131
10000	0.30	–1.710	1.711	–0.081	0.081	–0.015	0.015	–0.029	0.029
	0.60	–1.715	1.731	–0.081	0.081	–0.015	0.015	–0.029	0.029
	0.90	–1.718	1.701	–0.082	0.082	–0.015	0.015	–0.029	0.029

n: Sample size; *ρ*: Correlation coefficient between pre-values and post-values, PC: Percent Change, R: Ratio, MSPC: Modified Percent Change, SD: Standard Deviation, SEmean: Standard Error of Mean, TrMean: Trimmed Mean, IQR: Interquartile Range, $S1L_\alpha$ and $S1R_\alpha$: Left and Right tail critical values for Mean/SEmean; $S2L_\alpha$ and $S2R_\alpha$: Left and Right tail critical tables value for TrMean/SEmean; $S3L_\alpha$ and $S3R_\alpha$: Left and Right tail critical tables value for Median/IQR; $S4L_\alpha$ and $S4R_\alpha$: Left and Right tail critical table values for TrMean/IQR

Table 3: Observed type-I errors and power of *t*-test for difference in paired samples

n	*ρ*	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10 (t_α=±2.262)	0.30	5.13	22.23	66.43	94.48	99.74
	0.60	4.99	35.10	87.78	99.75	100
	0.90	4.81	87.86	100	100	100
30 (t_α=±2.045)	0.30	5.34	61.10	100	100	100
	0.60	4.88	84.00	100	100	100
	0.90	4.96	99.99	100	100	100
60 (t_α=±2.000)	0.30	4.99	89.73	100	100	100
	0.60	5.05	98.96	100	100	100
	0.90	5.02	100	100	100	100
120 (t_α=±1.980)	0.30	5.02	100	100	100	100
	0.60	5.13	100	100	100	100
	0.90	4.73	100	100	100	100
500 (t_α=±1.965)	0.30	5.12	100	100	100	100
	0.60	5.05	100	100	100	100
	0.90	5.23	100	100	100	100
10000 (t_α=±1.960)	0.30	5.06	100	100	100	100
	0.60	4.90	100	100	100	100
	0.90	4.81	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 4: Observed type-I errors and power of Median/IQR statistic (*HS* test) for MSPC

n	ρ	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10	0.30	5.01	5.71	12.93	32.66	58.88
	0.60	4.99	5.95	14.39	34.17	57.51
	0.90	4.95	7.19	19.68	37.84	56.82
30	0.30	5.05	7.12	31.10	81.31	100
	0.60	5.12	7.76	36.61	89.50	100
	0.90	4.82	12.5	47.86	80.53	100
60	0.30	4.91	9.56	56.77	97.99	100
	0.60	4.80	10.86	64.72	97.90	100
	0.90	4.95	2130	82.82	98.53	100
120	0.30	5.24	13.74	86.26	100	100
	0.60	4.97	17.36	92.06	100	100
	0.90	4.90	47.86	99.46	100	100
500	0.30	5.28	43.70	100	100	100
	0.60	5.26	55.67	100	100	100
	0.90	5.31	93.80	100	100	100
10000	0.30	5.50	100	100	100	100
	0.60	5.63	100	100	100	100
	0.90	5.55	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 5: Observed type-I errors and power of TrMean/IQR test statistic for MSPC

n	ρ	Type-I error (Observed alpha)	Standardized effects size			
			Δ=0.5	Δ=1.0	Δ=1.5	Δ=2.0
10	0.30	5.15	5.86	14.56	38.59	68.42
	0.60	5.01	6.08	16.29	39.40	64.55
	0.90	4.96	7.23	21.16	40.32	60.06
30	0.30	5.19	6.13	16.06	49.34	84.92
	0.60	4.99	5.78	16.85	49.10	82.10
	0.90	4.96	6.62	20.36	51.18	80.01
60	0.30	5.24	6.16	20.30	67.44	94.32
	0.60	4.82	6.26	21.91	63.72	92.10
	0.90	4.75	6.85	22.27	62.11	90.30
120	0.30	4.99	7.23	36.98	95.60	100
	0.60	4.70	7.64	37.30	96.81	100
	0.90	4.63	20.36	34.48	98.80	100
500	0.30	5.23	15.53	90.24	100	100
	0.60	5.12	18.23	90.50	100	100
	0.90	4.83	27.77	89.67	100	100
10000	0.30	5.09	100	100	100	100
	0.60	5.02	100	100	100	100
	0.90	4.92	100	100	100	100

n: Sample size; ρ: Correlation coefficient between pre-values and post-values, Δ: Standardized effect size

Table 6: Which test should be preferred?

<table>
<tr><th rowspan="2">n</th><th rowspan="2">ρ</th><th rowspan="2">Type-I error (Observed alpha)</th><th colspan="4">Power of Tests
Standardized effects size</th></tr>
<tr><th>Δ = 0.5</th><th>Δ = 1.0</th><th>Δ = 1.5</th><th>Δ = 2.0</th></tr>
<tr><td>10</td><td>0.30</td><td rowspan="19">Paired Samples t-test, Median/IQR statistic and TrMean/IQR statistic have same results in terms of observed type-I error</td><td colspan="4" rowspan="19">In this conditions, Paired Samples t-test should be used for simple difference

In these conditions, as the statistical properties of the t-test and HS test are similar, HS test (Median/IQR) used for MSPC is recommended.
This is because MSPC considers the differences between individuals in pre-treatment when computing the treatment effect. This makes it more accurate measure of change.</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>30</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>60</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>120</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>500</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
<tr><td>10000</td><td>0.30</td></tr>
<tr><td></td><td>0.60</td></tr>
<tr><td></td><td>0.90</td></tr>
</table>

n: Sample size; *ρ*: Correlation coefficient between pre-values and post-values

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 157-180

Forecasting Transmission Dynamics of COVID-19 in India Under Containment Measures- A Time-Dependent State-Space SIR Approach

Vishal Deo[1, 2], Anuradha R. Chetiya[2], Barnali Deka[2] and Gurprit Grover[1]

[1]*Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi, India*
[2]*Department of Statistics, Ramjas College, University of Delhi, Delhi, India*

Received: 12 May 2020; Revised: 30 May 2020; Accepted: 01 June 2020

Abstract

The primary objective of this paper is to predict the dynamics of COVID-19 epidemic in India while adjusting for the effects of various progressively implemented containment measures. Apart from forecasting the major turning points and parameters associated with the epidemic, we intend to provide an epidemiological assessment of the impact of these containment measures in India. This paper proposes a method based on time-series SIR (Susceptible, Infected, and Removed) model to estimate time-dependent modifiers for transmission rate of the infection. These modifiers are used in state-space SIR model to estimate the basic reproduction number R_0 and expected total incidence, and to forecast the daily prevalence till the end of the epidemic. We consider four different scenarios, two based on current developments and two based on hypothetical situations for the purpose of comparison.

Assuming gradual relaxation in lockdown post 17 May 2020, we expect the prevalence of infecteds to cross 9 million, with at least 1 million severe cases, around the end of October 2020. For the same case, estimates of R_0 for the phases no-intervention, partial-lockdown and lockdown are 4.46 (7.1), 1.47 (2.33), and 0.817 (1.29) respectively, assuming 14-day (24-day) infectious period.

Estimated modifiers give consistent estimates of unadjusted R_0 across different scenarios, demonstrating precision. Results corroborate the effectiveness of lockdown measures in substantially reducing R_0. Also, predictions are highly sensitive towards estimate of infectious period.

Key words: State-space SIR; Lockdown; Reproduction number; Time-dependent transmission; Infectious period; COVID-19; SARS-CoV-2; Total incidence.

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

1. Introduction

1.1. Context

In the absence of vaccines or effective antiviral therapies for COVID-19, governments all over the world are turning to classical non-pharmaceutical public health measures to contain the epidemic, such as isolation, quarantine, social distancing and community containment. Rigorous implementation of these four traditional counter-measures helped in halting the earlier epidemic of SARS-CoV in 2002-2003 [Kundapur *et al.* (2020); Wilder-Smith and Freedman (2020)]. As of May 8, 2020, infections across India have surged past 53,000 cases with 1783 deaths reported. The government of India has implemented these containment and mitigating interventions along with travel restrictions and lockdown of the entire country to slow down the spread of the virus. Epidemiological assessment using infectious disease modelling is the key to evaluate the impact of these measures on transmission dynamics of COVID-19 in India, and thus provide crucial information to the policy makers in government organizations to plan ahead for an effective and sustained public health response to manage the epidemic.

1.2. Review of epidemiological modeling of COVID-19

Recent studies of COVID-19 have attempted to predict number of case counts, rate of transmission, basic reproduction rate/ number (R_0), size of epidemic and end date of the epidemic. R_0 is an important factor for risk assessment of any epidemic and is defined as the expected number of secondary cases that arise from a typical infectious index-case in a completely susceptible host population. A large proportion of these studies have used different extended forms of the standard Susceptible-Infected-Removed (SIR) compartment model. In the context of COVID-19, the removed compartment includes both recovered and deceased cases. Wang *et al.* (2020) have developed a health informatics toolbox with an R package called eSIR to understand epidemiological trend of COVID -19 in Hubei province and other regions of China. Their model considers a time varying quarantine factor to forecast future trend of COVID -19 spread in these regions. An earlier study by Chinazzi *et al.* (2020) assessed the impact of such restrictions based on data of over 3200 sub-populations in roughly 200 different countries and territories across the world. They have used a meta-population network approach, in which each sub-population is modelled using a Susceptible-Latent-Infectious-Removed (SLIR) model.

COVID-19 in India

Using a compartmental Susceptible-Exposed-Infected-Removed (SEIR) model, Chatterjee *et al*. (2020a) have concluded that effective implementation of quarantine and other non-pharmacological interventions would bring down the epidemic spread of COVID-19 in India to a manageable level. Mandal *et al.* (2020) used a SEIR model with a quarantine component to predict an effective reduction in cumulative incidence in India. A SIR model is developed by Singh and Adhikari (2020) based on data up to the first phase of India's total lockdown to illustrate the need of sustained lockdowns with periodic relaxations. Some other recent work on COVID-19 in India include Tiwari (2020) and Gupta *et al.* (2020).

All these studies on COVID-19 infection dynamics in India are based on the assumption of constant disease transmission rate. However, phase-wise imposition of travel restrictions, lockdown and other non-pharmaceutical preventive measures, as well as increasing community level awareness with time, are expected to induce time varying effects in the transmission rate.

1.3. Our approach

To account for variations in transmission rate of the infection due to the implementation of various containment protocols, we propose to implement a time-dependent state-space SIR model to the observed data from India. Instead of taking a pre-specified step function modifier like Ray *et al.* (2020), we propose a time-series SIR based approach to estimate the phase-wise transmission modifiers. Modifier functions, both step and exponential, are estimated using the daily prevalence data reported in India from 2 March 2020 to 30 April 2020.

2. Methodology

2.1. The extended state-space SIR model with time varying transmission rate

The extended state-space SIR model proposed by Wang *et al.* (2020) is fitted to predict daily prevalence of susceptible, infected, and removed. This model is a time-dependent version of the state-space SIR model introduced by Osthus *et al.* (2017), and can be defined as follows.

Model description

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

$$Y_t^R|\boldsymbol{\theta_t}, \tau \sim Beta\left(\lambda^R\theta_t^R, \lambda^R(1-\theta_t^R)\right) \quad (2)$$

where,

Y_t^I - Time series of proportion of infected cases

Y_t^R - Time series of proportion of removed cases (Recovered + Dead)

θ_t^I - Prevalence of infection at time t in terms of probability (probability of a person being infected at time *t*)

θ_t^R - Prevalence of removal at time *t* (probability of a person being removed from the infected compartment)

Also, the constants λ^I and λ^R control the variances of the respective observed proportions.

$\boldsymbol{\theta_t} = (\theta_t^S, \theta_t^I, \theta_t^R)^T$ represents the latent population prevalence. It is a three–state Markov process where θ_t^S is the probability of a person being susceptible at time *t*. The Markov process (or the distribution of the transmissions of the Markov process) is defined as follows,

$$\boldsymbol{\theta_t}|\boldsymbol{\theta_{t-1}} \sim Dirichlet(\kappa f(\boldsymbol{\theta_{t-1}}, \beta_t, \gamma)) \quad (3)$$

Thus the complete model is a Dirichlet-Beta state-space model. Here, $\beta_t = \beta.\pi(t)$ is the modified/ effective transmission rate. The unadjusted transmission rate β is defined as contact rate multiplied by the probability of transmission given a contact between a susceptible and an infectious individual. That is, $\pi(t)$ is a time-dependent modifier function of the transmission rate. The function $f(.)$ in the argument of Dirichlet function is the SIR model given as follows.

$$\frac{d\theta_t^S}{dt} = -\beta_t \theta_t^I \theta_t^S \, , \frac{d\theta_t^I}{dt} = \beta_t \theta_t^S \theta_t^I - \gamma \theta_t^I \; and \; \frac{d\theta_t^R}{dt} = \gamma \theta_t^I \tag{4}$$

Solution of this set of differential equations is achieved using the Runge-Kutta approximation.

Defining time-dependent modifier function of transmission rate

Several containment measures have been implemented in India at different points in time creating phases of quarantine/ containment levels across the country. Such phases are expected to exhibit different rates of transmission of the disease, *i.e.*, transmission rates become time-dependent. If we assume that the change (or reduction) in the transmission rate is strictly because of macro level measures implemented by the authorities, we can define a specific transmission rate for each phase based on the level of containment. Wang *et al.* (2020) have proposed a step function approach to define such modifiers. However, it is also true that apart from the containment measures implemented by the government, rising awareness at micro community levels also contributes towards reducing the rate of transmission. To incorporate this idea, they have suggested defining the transmission modifier as a continuous function of time.

Suppose there are three different phases, with two points of major changes in quarantine/ lockdown protocols. Let P_i denotes i-th phase, such that P_1 represents the initial phase without any such protocol in place. Then, the step function for transmission rate modifier, $\pi(t)$, can be expressed as follows.

$$\pi(t) = \begin{cases} \pi_1 \; , if \; t \epsilon P_1 \\ \pi_2 \; , if \; t \epsilon P_2 \\ \pi_3 \; , if \; t \epsilon P_3 \end{cases} \tag{5}$$

where, $\pi_1 = 1$ if P_1 represents the phase without any intervention.

As an alternative technique, following exponential modifier functions can be used to account for continuous changes in modifier values with time.

$$\pi(t) = \exp(-\lambda_0 t) \; or \; \exp\{-(\lambda_0 t)^{\vartheta}\}, \lambda_0 > 0, \vartheta > 0 \tag{6}$$

However, in this case the modifier value is assumed to decrease at a constant rate over time, irrespective of the phases of containment measures. We have applied both approaches to define modifiers for the base transmission rate β. The effective rate of transmission at time t is given as, $\beta_t = \beta.\pi(t)$.

Overall success of this modelling structure depends heavily on the relevance of the modifier values specified for different phases. Using appropriate values of π_i*'s* in (5), and of the constants λ_0 and ϑ in (6) will be imperative towards achieving reliable predictions. To avoid misleading predictions resulting from speculative pre-specified values of the modifiers, we propose methods based on time-series SIR (TSIR) model to estimate these values. The proposed method is described in the following section.

2.2. Proposed method for estimating modifiers of β for different phases of quarantine/ lockdown measures

TSIR model [Bjørnstad *et al.* (2002); Finkenstadt, *et al.* (2002); Grenfell *et al.* (2002)] is used to estimate time-dependent modifier values. In the step function $\pi(t)$, the steps (or phases) are defined according to different levels of preventive measures implemented by the government over the observed period of time. In TSIR model, the response, being a count variable, is assumed to follow certain discrete count process distribution like Poisson distribution or Negative Binomial distribution; refer Bjørnstad (2018). The basic structure of TSIR model can be defined as follows:

$$S_{t+1} = S_t - I_t \tag{7}$$

$$\lambda_{t+1} = \beta_0 \frac{S_t}{N} I_t^{\alpha} \tag{8}$$

$$\text{Or, } log(\lambda_{t+1}) = log\beta_0 + \alpha log I_t + log\left(\frac{S_t}{N}\right) \tag{9}$$

where, S_t and I_t are the number of susceptibles and infecteds (or infectives) at time t, N is the population size, β_0 is the transmission rate and λ_{t+1} is the expected number of new infecteds at time t+1. New number of infecteds is assumed to follow Negative Binomial (or Poisson) distribution and a generalized Negative Binomial (or Poisson) linear model with log link is fitted with $logI_t$ as a covariate and $log\left(\frac{S_t}{N}\right)$ as an offset variable. The exponent α is expected to be just under 1 (*i.e.* close to 1) and is meant to account for discretizing the underlying continuous process. However, we can present an alternative interpretation of α based on the time-dependent SIR model given in (4). Using (4), the expression for expected number of new infecteds at time t+1 (taking $\alpha = 1$) with a time-varying transmission rate can be written as follows.

$$\lambda_{t+1} = \beta_t \frac{S_t}{N} I_t \tag{10}$$

Comparing equations (8) and (10), we can see that if $\alpha = 1$ (or close to 1), $\beta_t = \beta_0$ (constant over time). However, if the value of α deviates considerably from 1, it has impact on the effective value of transmission rate, thus making the effective rate of transmission time-dependent. That is, in such cases α assimilates the empirical changes in transmission rate over time. From equations (8) and (10), we can further write,

$$\widehat{\beta_t} = \beta_0 I_t^{\alpha-1} \tag{11}$$

Option 1: Defining step function for phase specific modifiers

Fitting TSIR models to the observed phase-specific data, we estimate β_0 and α separately for each phase. The effective transmission rate, β_t, is then estimated at each time t using equation (11). Average of these estimates over the time range of a phase is taken as an estimate of the effective transmission rate for that phase. Suppose we have three time phases in our study, say P_1, P_2, and P_3. Then, the estimate of phase specific transmission rate will be given as,

$$\hat{\beta}_{P_i} = \frac{1}{n_i}\sum_{t\epsilon P_i}\hat{\beta}_t, \quad i = 1,2,3 \tag{12}$$

And the estimated step function of modifiers will be,

$$\hat{\pi}(t) = \begin{cases} 1 & if\ t \in P_1 \\ \frac{\hat{\beta}_{P_2}}{\hat{\beta}_{P_1}} & if\ t\epsilon P_2 \\ \frac{\hat{\beta}_{P_3}}{\hat{\beta}_{P_1}} & if\ t\epsilon P_3 \end{cases} \tag{13}$$

Option 2: Defining continuous time-dependent exponential modifier function

Instead of fitting phase-wise models, we fit a generalized linear model on the entire observed data and obtain estimates of effective transmission rates, $\hat{\beta}_t$, using equation (11). We derive estimates of modifiers at each time point *t* for the entire observed period as,

$$\hat{\pi}(t) = \frac{\hat{\beta}_t}{\hat{\beta}_1} \tag{14}$$

where, $\hat{\beta}_1$ is the estimate at t =1. However, if the first phase P_1 is small, we can take $\hat{\beta}_1 \cong \hat{\beta}_{P_1}$ to avoid impact of extreme observation at t =1 (if present). As an alternative, we can take $\hat{\beta}_1$ as an average of first few values of $\hat{\beta}_t$. We can fit any of the two exponential functions given in equation (6) to the estimated modifiers using least squares estimation. We have used only the first form in our study, *i.e.*, $\pi(t) = \exp(-\lambda_0 t)$.

This continuous modifier function will not be phase specific and will describe steadily increasing awareness at community-level which encourages voluntary participation in quarantine and preventive measure. The steadily decreasing modifier function can also account for the learning curve of the organizational structure associated with implementation of proposed preventive measures like quarantine, travel ban, partial lockdown and complete lockdown.

3. Implementation

3.1. Data

Since some states in India have not reported any cases and some have reported only few, we have considered populations of states with at least 10 confirmed cases reported till 20 April 2020 for calculating total number of susceptibles. Baseline state-wise population data is obtained from the 2011 census of India (www.censusindia.gov.in). The estimated average growth rate based on the current total population of India and the total population of India in 2011 is estimated to be 1.23% per annum. This rate is used to estimate current total populations of the 25 states which have been included in the calculation of total number of susceptibles. Data on the timeline of implementation of travel restrictions, isolation, lockdown, quarantine and other preventive measures taken by the central and state governments is compiled from various

notifications issued by the Ministry of Home Affairs and the Ministry of External Affairs available on their official websites. Time-series data on daily prevalence of total confirmed, total recovered and total deaths is sourced from the github repository of the Centre for Systems Science and Engineering, Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19).

3.2. Defining longitudinal phases based on containment protocols

While analyzing the effect of containment measures on rate of transmission of infection, it is important that we take into account the average incubation period. The mean incubation period of COVID-19, defined as the time from exposure to the onset of illness, is reported to be around 5 days by many studies; refer Lauer *et al.* (2020); Chatterjee *et al.* (2020b) and Yuan *et al.* (2020) among others. This means that the impact of any intervention on the transmission rate can be expected to be visible only after 5 days, on an average. Given the fact that India has preferred focused group testing over random testing, it becomes important to address the expected lag in reporting of cases. So, for improving the analysis, cut-off dates for defining phases have been extended by 5 days to accommodate for the lag in effect induced by the incubation period. Complete lockdown in India came into effect on 25 March 2020. However, because of sudden loss of jobs and earnings of daily wagers, and the uncertainty looming over the extension of lockdown period, there were huge movements of migrant workers across India, with most of them trying to reach their homes. Overwhelming number of reports emerged about inter-state travels of large groups of people, with many even forced to travel hundreds of kilometers on foot. According to an article published in Business Standards, [Jha (2020)], on 31 March 2020 the central government reported in the Supreme Court that 500,000-600,000 migrants reached their villages on foot during the lockdown. However, as per news reports, most of the state governments, assisted by various NGOs, had come up with adequate relief shelters and food arrangements for the stranded migrant laborers by 30 March 2020. Also, affected states started compulsory quarantine facilities for people migrating from other states. These measures helped in containing any significant movement and ensuring implementation of complete lockdown. Citing these developments, we have assumed the effective date of implementation of complete lockdown as 31 March 2020. Adding incubation period of 5 days, the cut-off date for the third phase for our analysis is taken as 04 April 2020. The actual and effective dates of implementation of preventive measures in India considered in the study are listed in Table 1.

3.3. Modifier functions and hyper-parameters

Based on the phases defined in section 3.2, step-function modifier, $\pi(t)$, is estimated using equation (13). Negative Binomial TSIR models are chosen over Poisson TSIR models to find estimates of $\widehat{\beta_t}$ (equation (11)). Poisson models showed inflated residual deviance and proved unfit for the data. Estimated step-function is given below.

$$\hat{\pi}(t) = \begin{cases} 1 & if\ t\epsilon[02\ March\ 2020 - 17\ March\ 2020] \\ 0.33 & if\ t\epsilon[18\ March\ 2020 - 03\ April\ 2020] \\ 0.182 & if\ t\epsilon[04\ April\ 2020 - 30\ April\ 2020] \end{cases} \quad (15)$$

Hyper-parameters for Bayesian estimation

Using data on COVID-19 patients in China, Verity *et al.* (2020) have estimated mean duration from onset of symptoms to death to be 17.8 days (95% credible interval 16.9−19.2) and to hospital discharge to be 24.7 days (22.9−28.1). Mean infectious period is calculated as weighted average of these durations using observed proportions of deaths and recoveries among the total removed cases till 30 April 2020 in India as weights. The estimated mean infectious period is: 0.113 x 17.8 + 0.887 x 24.7 ≈ 24 days. Thus, the estimate for hyper-parameter for γ is, $\gamma_0 = 1/24 = 0.042$. However, since there is dearth of comprehensive reports confirming infectious period at this early stage of the epidemic, we have also performed analyses taking mean infectious period of 14 days (*i.e.* $\gamma_0 = 1/14 = 0.0714$), as reported by the World Health Organisation; see WHO (2020). So, at this juncture it is safe to assume that the reality may lie somewhere between the projections based on our two assumed cases for γ_0. The value for the hyper-parameter β is estimated as the average of effective transmission rates over the total observed period (02 March 2020- 30 April 2020). This is achieved by fitting the Negative Binomial TSIR model and using equation (12) for the entire observed period.

Continuous modifier function

The continuous modifier function obtained using the modifier values estimated from equation (14) is given below.

$$\hat{\pi}(t) = \exp(-0.0285t) \tag{16}$$

3.4. Forecasting assumptions

We have assumed four different scenarios for forecasting the trajectory of the COVID-19 epidemic. The four cases are summarized in Table 2. Case 1 and case 3 are realistic scenarios based on current developments, while case 2 and case 4 are hypothetical scenarios strictly for the purpose of comparison.

3.5. Data calibration

In India, till now, testing strategy has been focused primarily on high risk individuals. However, to understand the community spread in the country, large scale random testing should be conducted among those who have no travel history [Rao *et al*. (2020)]. As reported recently by the Indian Council of Medical Research, around 80% of the total infected (confirmed) cases in India are asymptomatic; refer www.indiatoday.in (2020). In the absence of rigorous testing, it is but natural that a large number of true cases are going undetected and hence unreported in India. This subsequently leads to concerns about the actual number of deaths due to COVID -19 also going unreported [Shaikh (2020); Biswas (2020)].

We have used a simple intuitive technique for data calibration to account for possible under-reporting. We divide the observed data on total confirmed, recovered and deaths by a constant ρ (where $0 < \rho \leq 1$). Proportion of under-reporting is 1- ρ, i.e., $\rho = 1$ implies zero under-reporting. We have considered two levels of under-reporting, 75% (ρ =0.25*)* and 50% *(*ρ=0.5*)*. It is not easy to estimate the proportion of under-reporting, especially at this stage of the epidemic.

However, we have based our assumptions on certain reports on scientific work in this regard; refer Jayan (2020).

3.6. Plotting predicted prevalence

MCMC posterior realizations on the prevalence of infected and removed are obtained from the output of *tvt.eSIR( )* function of the eSIR package. Posterior mean of predicted prevalence of infecteds is plotted against time along with daily estimated prevalence of mild to moderate, severe and critical cases among the total infecteds. To predict the cases belonging to the categories mild to moderate, severe and critical, we have considered the respective proportions, 80.1%, 13.8%, and 6.1%, as reported by the World Health Organisation; refer WHO (2020). To predict the number of deaths, we have used current proportion of deaths among the total removed cases in India, which is around 10%. Prevalence of removed is plotted against time along with estimated number of cases for the events recovered and death. Plots for case 3 of step-function modifier and for exponential modifier at two different values of γ are presented in Graphs 1-5.

4. Results and Discussion

Estimated values of time-dependent transmission rate adjusted for modifier, $\hat{\beta}_t$ (for phase 1, $\hat{\beta}_t = \hat{\beta}$) , rate of removal, $\hat{\gamma}$, and reproduction number R_0, along with their 95% credible intervals based on posterior realizations are reported for all models and all cases discussed in the implementation section. Expected total incidence (as % of total population), and forecasted dates for two crucial turning points of the epidemic are also reported for each case. The first turning point signifies the time at which the rate of increase in the number of infecteds starts decreasing (deceleration). The second crucial turning point is the peak time of the infected curve beyond which the prevalence of infecteds starts decreasing. Table 3 and Table 4 present results for all four cases of step-function modifier based state-space SIR models, at the prior infectious period estimates of 24 days and 14 days respectively. Results from the exponential modifier function based state-space SIR models, for both observed and calibrated data, are presented in Table 5. Table 6 contains prediction results for calibrated data using case 3 of the step-function modifier.

At γ_0 = 0.042 (24-day infectious period), estimated values of the production rate R_0 (unadjusted) consistently stays around 7 in all cases, and at γ_0 = 0.0714 (14-day infectious period) its estimates cluster around 4 for all cases. Consistency of the estimate of R_0 (unadjusted) and β (base value unadjusted for modifiers) under different hypothesized situations suggests that our estimates of modifiers are able to explain the changes in the transmission rate in their respective phases. Estimates of R_0 are comparatively very small in the containment phases, with that for the complete lockdown being the minimum. For example, citing results of case 1 from Table 3, under the assumption of 24-day infectious period, R_0 is estimated to be 1.29 for the lockdown period, and around 2.33 for the quarantine/ partial lockdown phase, as opposed to 7.1 in the no-intervention phase. Similar results are obtained for the case 3, at both levels of γ_0. R_0 values estimated from exponential modifier function based approach are slightly on the lower side as compared to those obtained from the step-function approach. This is expected as the use of exponential modifier function results in continuous decline in the transmission rates through time.

Estimate of (mean) total incidence is very sensitive to the choice of infectious period. Even under the assumption of complete lockdown till the end of epidemic, the estimated total incidence jumps from 0.35% to 7.7% of the population as we increase the infectious period from 14 days to 24 days. Similar jumps are seen in all cases. Unfortunately, as discussed in section 3.3, the existing reports on COVID-19 at this early stage of the epidemic are not conclusive about the duration of infectious period. In addition, current recovery/ death trends of different countries indicate that recovery rates and death rates can vary significantly between different regions.

Although we obtained most optimistic results under the assumption of complete lockdown like situation throughout the course of the pandemic, it is not practical to believe that our economy can sustain such a drastic measure for a long duration. Among the situations assumed for prediction using step-function modifiers, future assumptions for case 3 are practically most achievable. Also, daily predictions of number of infecteds for the month of May based on case 3 of step-function modifiers, with 24-day infectious period, are closest to the actual reported data as compared to those of any other scenario considered in this study; refer Graph 5. To restrict the COVID-19 spread within the limits predicted by case 3 results, we have to ensure that the post complete lockdown period should not let R_0 to go beyond 1.32 (14-day infectious period) or beyond 2.14 (24-day infectious period). In that situation, depending on the actual recovery time of COVID-19 patients in India we can expect around 9.1% to 31.8% of the total population to get infected with SARS-CoV-2 by the time the epidemic ends. Assuming 75% under-reporting of infected and recovered/ deceased cases, the range of expected total incidence becomes 30.1% - 67.2%, and for 50% under-reporting of cases, it is estimated as 16.4% - 49.7%. The rate of infection is expected to start decreasing around the end of August or start of September 2020, and the total number of active cases is expected to start declining towards the end of October 2020. If there is under-reporting of cases, these dates of turning points are expected to shift earlier by around a week.

It is also worth mentioning that if the lockdown measures had not been implemented and only quarantine and partial lockdown were continued (case 2), we would be expecting around 22% to 53% of the population to be infected by the end of the epidemic. And if there was no containment measure in place since start (case 4), 88% to 95% of the population would have contracted the infection till the time epidemic lasted.

Since the exponential modifier function assumes a continuous decline in the effective transmission rate, it may overlook some important real life factors while predicting the course of the epidemic, and hence it may result in underestimation of the overall impact of the epidemic. As expected, the results obtained using this approach is closer to those of case1 where complete lockdown is assumed beyond 17 May 2020. Use of exponential modifier function may not be the best way to describe the effects of sudden drastic measures like complete lockdown, travel ban *etc.*

Even in the best case scenario as depicted by the results obtained using the exponential modifier function, the total incidence is predicted to be up to 1.2% (without data calibration) and up to 2% (with data calibration). That is, around 16 million to 27 million people are expected to end up getting infected with SARS-CoV-2 by the time the epidemic ends. According to the

predictions from the exponential modifier approach, when the prevalence of infected reaches its peak around mid to end of July 2020, there will be around 60,000 (14-day recovery period) to 500,000 (24-day recovery period) severe cases who will need hospitalization, at once (Graph 3 and Graph 4). The picture becomes even more unsettling when we study the graphs of the case 3 predictions (Graph 1 and Graph 2). These figures range between 1 million to 8 million for the two cases, and the peak time is expected to be around the end of October 2020.

5. Conclusion

Substantial reduction in the reproduction rate R_0 during the partial lockdown and complete lockdown phases corroborates the effectiveness of these interventions in containing the spread of SARS-CoV-2 infection. Assuming an average recovery (or infectious) period of 14 days, R_0 is estimated to have reached below 1 in the complete lockdown phase. However, assuming a 24-day recovery period, the estimate of R_0 remained above 1 even during the complete lockdown. Under case 3, where we have considered existing situation of containment measures till 17 May 2020, the daily predictions for May 2020 are much closer to the actual reported values (followed till 27 May 2020) at $\hat{\gamma} = 0.042$ (24-day infectious period) as compared to those at $\hat{\gamma} = 0.0714$ (14-day infectious period). However, more clinical reports based on wider patient level data are imperative towards finding reliable estimates of recovery time for COVID-19 patients.

The fact, that instead of using pre-specified modifier for each phase we have estimated phase-specific modifiers from the observed data, improves our chances of obtaining more reliable estimates of transmission rate and R_0 as compared to other recent studies on India, like Ray *et al.* (2020). Use of lower than true values of modifiers may lead to over-estimation of transmission rate and vice-versa. Also, our procedure of defining cut-off dates for different phases of containment measures assimilates the effects of incubation period and initial lapses in the implementation of the lockdown.

Even under the most optimistic scenario, the time for flattening of the COVID-19 progression curve is still quite far. Number of infected cases is expected to increase at even a higher rate at the moment and by the time the peak is expected, we will need an extensive amount of medical and infrastructural preparedness. Quoting the results based on the assumptions of case 3, which we have repeatedly deemed as the most realistic case, we need to be prepared with enough health-care infrastructure to be able to handle between 1 million to 8 million severe cases around the end of October 2020.

Limitations

We have considered same estimates of phase-specific modifier for the entire country assuming that the lockdown and other containment protocols have been homogeneously implemented across India. However, because of significant differences in various socio-economic, demographic, cultural, and administrative level factors, actual transmission rates are bound to differ from region to region. Hence, the estimated parameters in our study are only valid for overall predictions of cases in India, on an average, and may fail to trace the dynamics of the epidemic in sub-regions, say districts or states.

Acknowledgements

We are extremely grateful to the reviewers and the editors for their invaluable comments and suggestions, which have helped us to improve the paper.

References

Biswas, S. www.bbc.com. (April 28, 2020). *India coronovirus: The 'mystery' of low covid-19 death rates*. Retrieved from https://www.bbc.com/news/world-asia-india-52435463.

Bjørnstad, O., Finkenstadt, B. and Grenfell, B. (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, **72(2)**, 169-184.

Bjørnstad, O. N. (2018). *Epidemics Models and Data Using R*. Springer, Switzerland (ISBN 978-3-319-97486-6).

Chatterjee, K., Chatterjee, K., Kumar, A. and Shankar, S. (2020a). Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. *Medical Journal Armed Forces India*, https://doi.org/10.1016/j.mjafi.2020.03.022.

Chatterjee, P., Nagi, N., Agarwal, A., Das, B., Banerjee, S., *et al.* (2020b). The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidence. *Indian Journal of Medical Research,* **151**, 147-159, https://doi.org/10.4103/ijmr.IJMR_519_20.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., *et al.* (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, https://doi.org/10.1126/science.aba9757.

Finkenstadt, B. F., Bjørnstad, O. N., and Grenfell, B. T. (2002). A stochastic model for extinction and recurrence of epidemics: Estimation and inference for measles outbreaks. *Biostatistics,* **3(4)**, 493–510.

Grenfell, B. T., Bjørnstad, O. N., and Finkenstadt, B. F. (2002). Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs, 72***(2),** 185–202.

Gupta, R., Pandey, G., Chaudhary, P. and Pal, S. (2020). SEIR and regression model based COVID-19 outbreak predictions in India. *medRxiv*, *preprint*, https://doi.org/10.1101/2020.04.01.20049825.

Jayan, T. V. www.thehindubusinessline.com. (April 17, 2020). *India may be detecting 1 in 4 covid-19 cases: Mathematical expert*. Retrieved from https://www.thehindubusinessline.com/news/science/india-may-be-detecting-1-in-4-covid-19-cases-mathematical-expert/article31366694.ece.

Jha, S. (April 1, 2020). *Nearly 600,000 workers migrated on foot during lockdown, govt tells SC*. Retrieved from https://mybs.in/2YLmm6L.

Kundapur, R., Rashmi, A., Sachin, M., Falia, K., Remiza, R.A. and Bharadwaj, S. (2020). COVID 19 – Observations and speculations – A trend analysis. *Indian Journal of Community Health*, **32(2**-Special Issue**)**, 300-305.

Lauer, S., Grantz, K., Bi, Q., Jones, F., Zheng, Q., *et al.* (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, https://doi.org/10.7326/M20-0504.

Mandal, S., Bhatnagar T., Arinaminpathy N., Agarwal, A., Chowdhury, A., *et al.* (2020). Prudent public health intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model-based approach. *Indian Journal of Medical Research*, **151(2),** 190-199, https**:**//doi.norg/10.4103/ijmr.IJMR_504_20.

Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D. and Valle, S. Y. D. (2017). Forecasting seasonal influenza with a state-space SIR model. *The Annals of Applied Statistics*, **11(1)**, 202-224.

Rao, A. S. R. S, Krantz S. G., Kurien T., Bhat, R. and Sudhakar, K. (2020). Model-based retrospective estimates for covid-19 or coronavirus in India: Continued efforts required to contain the virus Spread. *Current Science,* **118 (7)**, 1023-1025.

Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Mohammed, S., *et al.* (2020). Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: Data science call to arms. *medRxiv, preprint,* https://doi.org/10.1101/2020.04.15.20067256.

Shaikh, Z. www.indianexpress.com. (May 04, 2020). *Malegaon mystery: Covid count low but surge in overall deaths.* Retrieved from https://indianexpress.com/article/india/malegaon-death-coronavirus-count-covid-19-test-6392528/.

Singh, R. and Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. *arXiv, preprint*, arXiv:2003.12055v1 [q-bio.PE].

Tiwari, A. (2020). Modelling and analysis of COVID-19 epidemic in India. *medRxiv*, *preprint*, https://doi.org/10.1101/2020.04.12.20062794.

Verity, R., Okell, L., Dorigatti, I., Winskill, P., Whittaker, C., *et al.* (2020). Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases*, https://doi.org/10.1016/S1473-3099(20)30243-7.

Wang, L., Zhou, Y., He, J., Zhu, B., Wang, F., *et al.* (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. *medRxiv, preprint*, https://doi.org/10.1101/2020.02.29.20029421.

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

Wilder-Smith, A. and Freedman, D. O. (2020). Isolation, quarantine, social distancing and community containment: pivotal role for old style public health measures in the novel coronavirus (2019-nCoV) outbreak. *Journal of Travel Medicine,* **27(2),** https://doi.org/10.1093/jtm/taaa020.

www.indiatoday.in. (April 20, 2020). *Coronavirus: 80% cases asymptomatic, but no need to revise testing criteria, says ICMR.* Retrieved from https://www.indiatoday.in/india/story/80-of-coronavirus-cases-in-india-are-asymptomatic-icmr-1669073-2020-04-20.

Yuan, J., Li, M., Lv, G. and Lu, K. (2020). Monitoring transmissibility and mortality of COVID-19 in Europe. *International Journal of Infectious Diseases, In Press Journal Pre-Proof*., https://doi.org/10.1016/j.ijid.2020.03.050.

Table 1: Cut-off dates of major preventive interventions implemented in India

Preventive Measures	Actual Dates of Implementation	Cut-off Dates for Defining Phases in Our Study
No strict screening, quarantine measures, or surveillance.	2 March–12 March	2 March–17 March
Various social distancing measures, restrictions on public gathering, shutting down of academic institutions, international travel ban, and restriction on public transport.	13 March–24 March	18 March–3 April
Complete lockdown 1	25 March–14 April	4 April–30 April
Complete lockdown 2- lockdown continues along with certain exemption for selected activities.	20 April–3 May	4 May– 17 May

Table 2: Cases assumed for forecasting

Case	Phases	Modifiers	Assumption
Case1	P_1 : 02 March-17 March P_2: 18 March-03 April P_3: 04 April onwards	1 0.33 0.182	After 30 April 2020, the effect of containment measures will remain more or less the same with the modifier value equal to that of the third phase, P3.
Case2	P_1: 02 March-17 March P_2: 18 March onwards	1 0.33	Assuming that instead of complete lockdown only quarantine and partial lockdown measures were extended throughout after 12 March 2020.
Case3	P_1: 02 March-17 March P_2: 18 March-03 April P_3: 04 April-03 May P_4: 04 May-17 May P_5: 18 May onwards	1 0.33 0.182 0.2* 0.3*	Because of slight relaxations in green and orange zones, modifier value is assumed to increase a bit till 17 May. After that it is assumed that more economic and industrial work will start, bringing the modifier value closer to that of second phase, but will remain less than that as red zones will be strictly contained.
Case4	Basic Case- No phase modifiers assumed	---	If no containment measures were taken through the entire epidemic period.

* Value assumed according to the assumptions of the case, based on the estimates of prior phases

Table 3: Predictions using step-function modifiers (24-day infectious period)

Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	Posterior Mean Estimates (with 95% Credible Intervals)				Expected dates of important turning points	
Assumed Scenario	$\widehat{\beta}_t$	$\widehat{\gamma}$	R_0	**Expected Total Incidence (% of popn.)**	**Rate of infections start decreasing (deceleration)**	**Total number of infected start decreasing**
Case 1: Intervention effects similar to that of lockdown protocols to continue after 30 April 2020.	0.374 (0.205-0.601)	0.0528 (0.0302-0.0815)	7.1 (5.41-9.16)	7.7%	12 November 2020	20 March 2021
Adjusted for Quarantine & partial lockdown	0.123 (0.068-0.198)		2.33			
Adjusted for Lockdown	0.068 (0.037-0.109)		1.29			
Case 2: If only partial lockdown protocols were extended (no complete lockdown)	0.369 (0.204-0.592)	0.0525 (0.0308-0.0803)	7.05 (5.31-9.01)	53%	08 August 2020	24 September 2020
Adjusted for Quarantine & partial lockdown	0.122 (0.067-0.195)		2.32			
Case 3: If lockdown is slightly relaxed (only for green and orange zones) after 17 May 2020.	0.36 (0.191-0.594)	0.0505 (0.0293-0.0802)	7.14 (5.36-9.34)	31.8%	24 August 2020	30 October 2020
Adjusted for Quarantine & partial lockdown adjusted	0.119 (0.063-0.196)		2.36			
Adjusted for Lockdown	0.066 (0.035-0.108)		1.31			
Adjusted for Zone-wise protocols (Green/Orange/Red) (assumed)	0.072 (0.038-0.119)		1.43			
Adjusted for some more degree of relaxation in lockdown protocols for Green and Orange zones post 17 May 2020 (assumed)	0.108 (0.057-0.178)		2.14			
Case 4: If no quarantine/ lockdown measures were implemented from the beginning	0.252 (0.158-0.371)	0.0388 (0.0244-0.0578)	6.57 (5.03-8.46)	95%	10 June 2020	03 July 2020

Table 4: Predictions using step-function modifiers (14-day infectious period)

Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	Posterior Mean Estimates (with 95% Credible Intervals)				Expected dates of important turning points	
Assumed Scenario	$\widehat{\beta}_t$	$\widehat{\gamma}$	R_0	**Expected Total Incidence (% of popn.)**	**Rate of infections start decreasing (deceleration)**	**Total number of infected start decreasing**
Case 1: Intervention effects similar to that of lockdown protocols to continue after 30 April 2020.	0.349 (0.196-0.565)	0.0783 (0.056-0.1054)	4.46 (2.79-6.7)	0.35%	18 December 2020	15 February 2021
Adjusted for Quarantine & partial lockdown	0.115 (0.065-0.186)		1.47			
Adjusted for Lockdown	0.064 (0.036-0.103)		0.817			
Case 2: If only partial lockdown protocols were extended (no complete lockdown)	0.339 (0.193-0.532)	0.0776 (0.0567-0.104)	4.38 (2.78-6.41)	21.8%	15 August 2020	11 October 2020
Adjusted for Quarantine & partial lockdown	0.112 (0.064-0.176)		1.44			
Case 3: If lockdown is slightly relaxed (only for green and orange zones) after 17 May 2020.	0.343 (0.194-0.54)	0.078 (0.0568-0.1042)	4.42 (2.79-6.57)	9.1%	06 September 2020	30 October 2020
Adjusted for Quarantine & partial lockdown adjusted	0.113 (0.064-0.178)		1.45			
Adjusted for Lockdown	0.062 (0.035-0.098)		0.79			
Adjusted for Zone-wise protocols (Green/Orange/Red) (assumed)	0.069 (0.039-0.108)		0.88			
Adjusted for some more degree of relaxation in lockdown protocols for Green and Orange zones post 17 May 2020 (assumed)	0.103 (0.058-0.162)		1.32			
Case 4: If no quarantine/ lockdown measures were implemented from the beginning	0.254 (0.162-0.367)	0.0741 (0.0542-0.0995)	3.46 (2.35-4.91)	88.2%	13 June 2020	03 July 2020

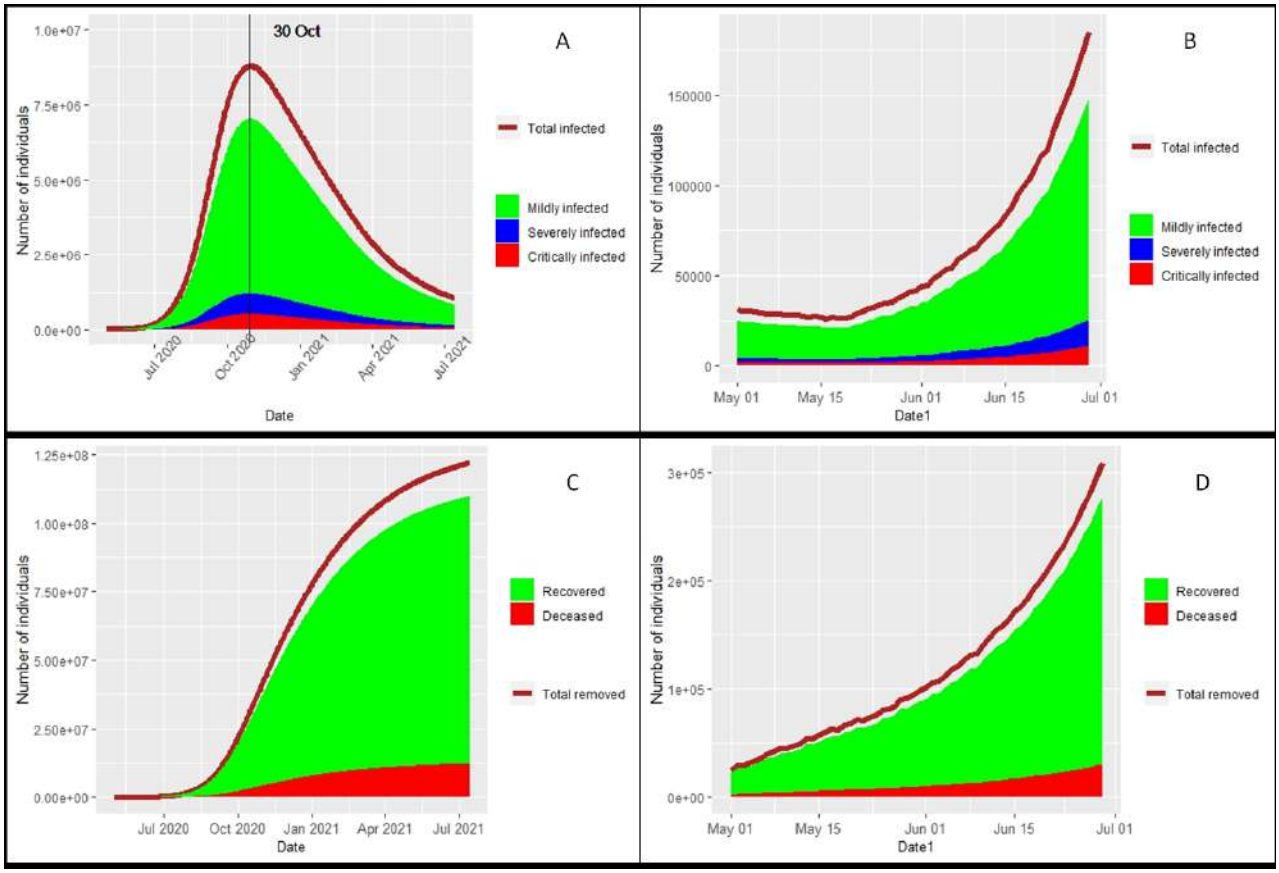
Table 5: Predictons using exponential modifier function

$\lambda_0 = 0.0131$	Posterior Mean Estimates (with 95% Credible Intervals)				Expected dates of important turning points	
Case	$\widehat{\beta}$	$\widehat{\gamma}$	R_0	**Expected Total Incidence (% of popn.)**	**Rate of infections start decreasing (deceleration)**	**Total number of infected start decreasing**
Without data calibration (assuming scale of under-reporting is not significant)						
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	0.329 (0.19-0.509)	0.0473 (0.0287-0.0767)	7.01 (5.24-8.94)	1.2%	26 June 2020	31 July 2020
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	0.326 (0.19-0.504)	0.0788 (0.0566-0.1046)	4.16 (2.65-6.07)	0.22%	08 June 2020	12 July 2020
With data calibration for under-reporting (assuming 75% under-reporting)						
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	0.312 (0.196-0.458)	0.0466 (0.0293-0.0694)	6.76 (5.2-8.71)	2%	22 June 2020	29 July 2020
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	0.307 (0.187-0.443)	0.0751 (0.0564-0.0981)	4.11 (2.78-5.73)	0.38%	30 May 2020	02 July 2020
With data calibration for under-reporting (assuming 50% under-reporting)						
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	0.326 (0.192-0.491)	0.0481 (0.0292-0.0727)	6.83 (5.22-8.86)	1.65%	24 June 2020	30 July 2020
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	0.322 (0.192-0.475)	0.0774 (0.0566-0.1050)	4.18 (2.67-5.9)	0.28%	04 June 2020	08 July 2020

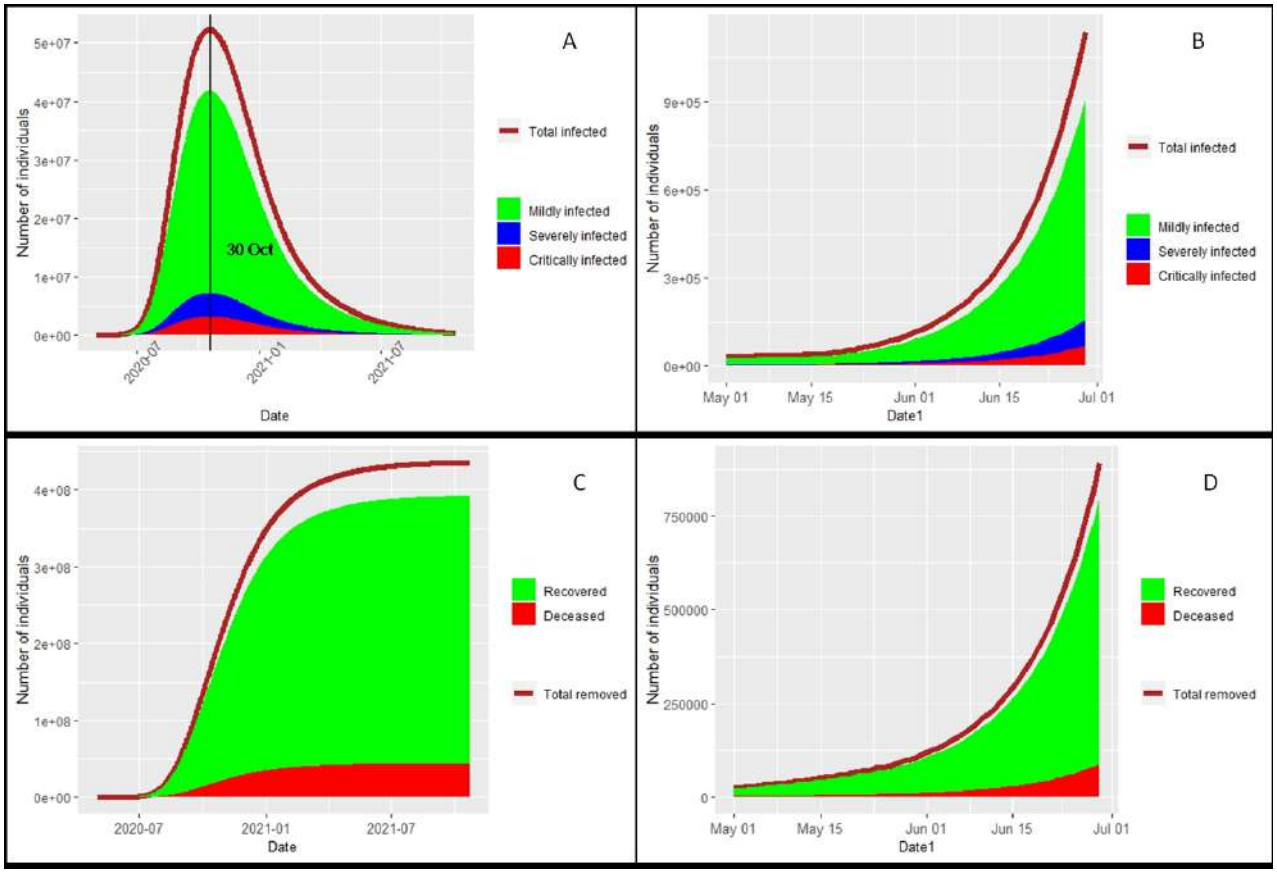
Table 6: Predictions using step-function modifiers and calibrated data

Case 3: If lockdown is slightly relaxed (only for green and orange zones) after 17 May 2020	Posterior Mean Estimates (with 95% Credible Intervals)				Expected dates of important turning points	
	$\widehat{\beta}$	$\widehat{\gamma}$	R_0	**Expected Total Incidence (% of popn.)**	**Rate of infections start decreasing (deceleration)**	**Total number of infected start decreasing**
With data calibration for under-reporting (assuming 75% under-reporting)						
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	0.39 (0.216-0.622)	0.054 (0.0319-0.0818)	7.24 (5.41-9.31)	67.2%	15 August 2020	04 October 2020
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	0.366 (0.206-0.59)	0.0759 (0.0559-0.0994)	4.82 (2.99-7.35)	30.1%	13 August 2020	14 October 2020
With data calibration for under-reporting (assuming 50% under-reporting)						
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.042$, $R_0 = 7.05$	0.389 (0.212-0.622)	0.0545 (0.0313-0.086)	7.17 (5.43-9.29)	49.7%	14 August 2020	12 October 2020
Hyper-parameters: $\beta_0 = 0.296$, $\gamma_0 = 0.0714$, $R_0 = 4.15$	0.357 (0.198-0.568)	0.0782 (0.0575-0.1045)	4.57 (2.86-6.87)	16.4%	23 August 2020	22 October 2020

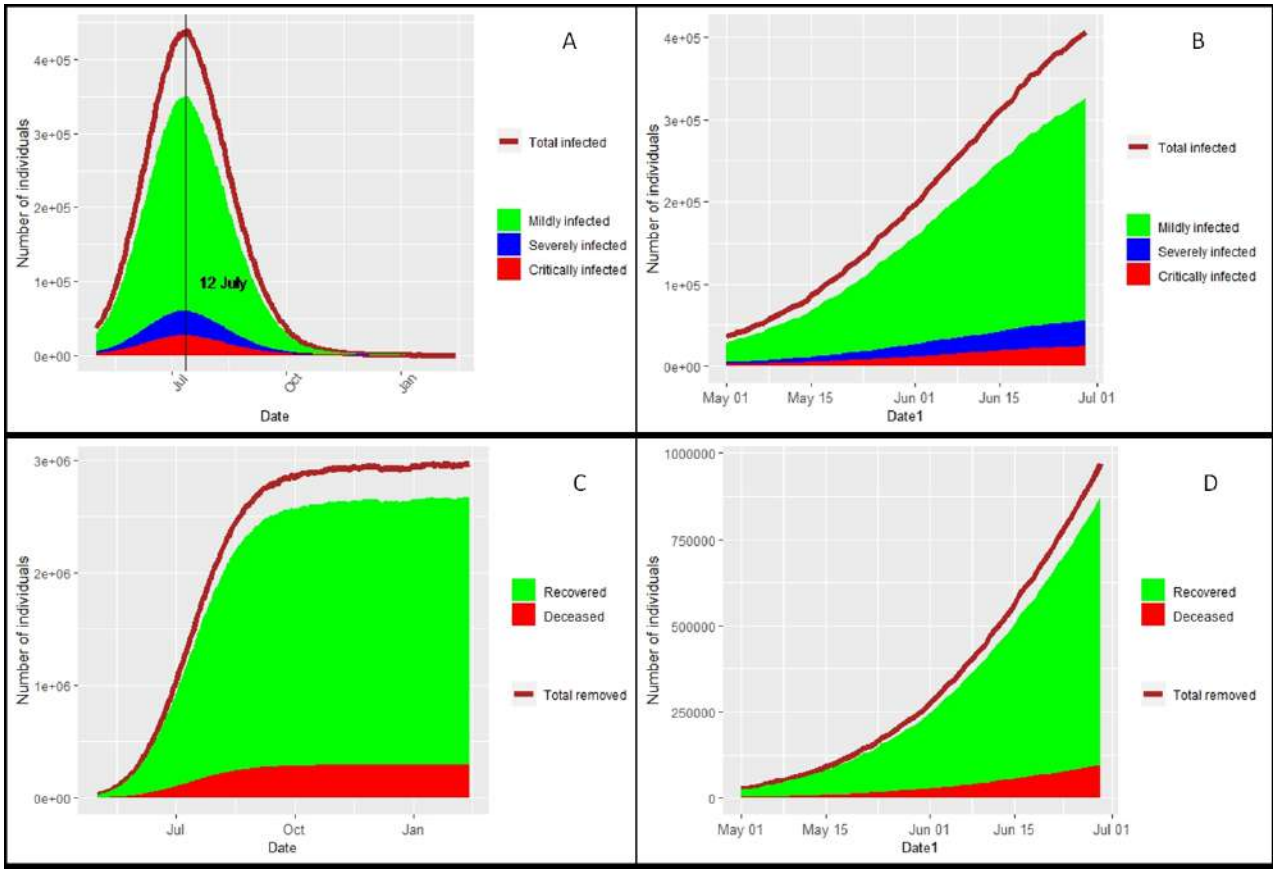
Graph 1: Predictions from case 3 of step-function modifier- (at $\hat{\gamma}$ = 0.0714)- Panel A- Number of infecteds predicted till the last day of the epidemic; vertical black line is the expected date for second turning point. Panel B- Number of infecteds shown till the end of June 2020. Panel C- Number of removed cases predicted till the last day of the epidemic. Panel D- Number of removed cases shown till the end of June 2020.



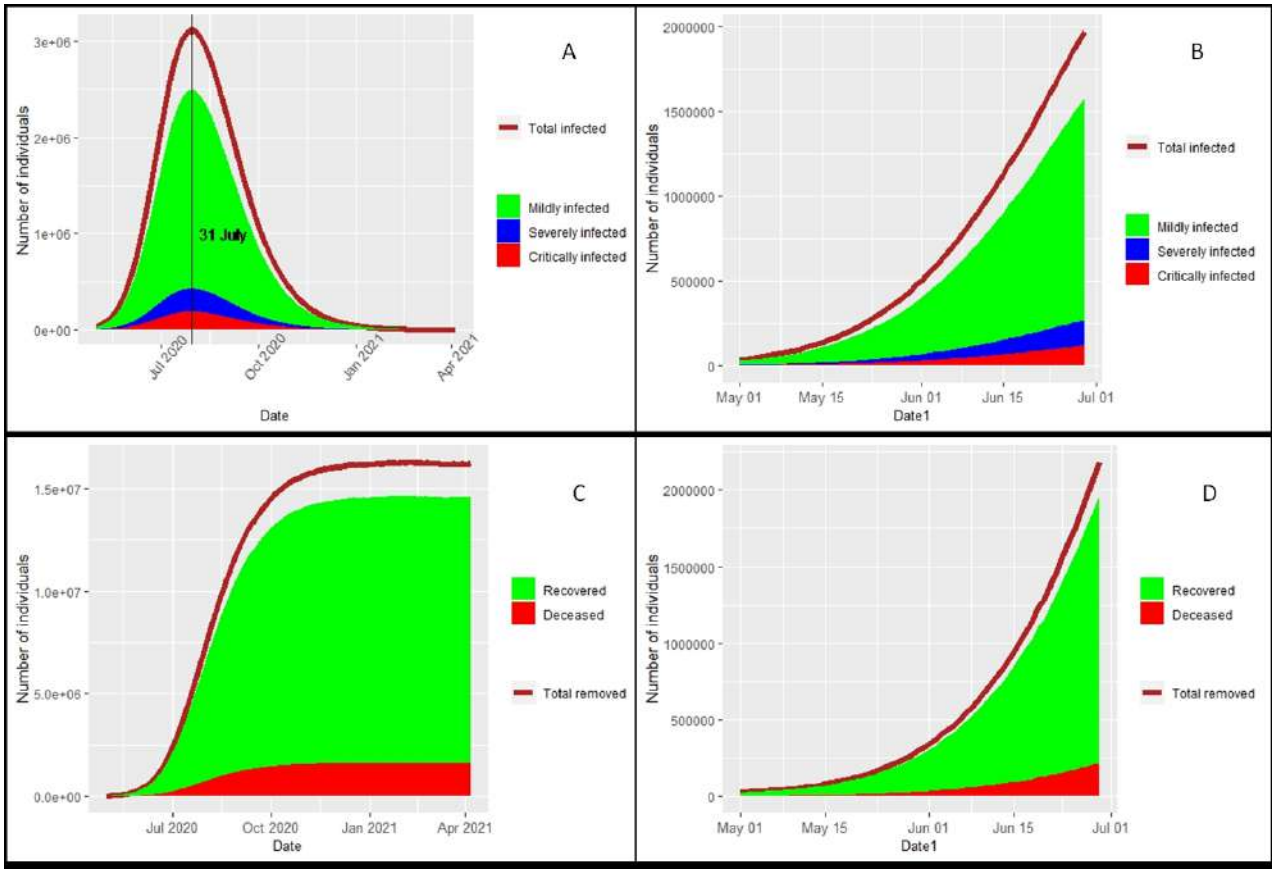
Graph 2: Predictions from case 3 of step-function modifier- (at $\hat{\gamma} = 0.042$)- Panel A- Number of infecteds predicted till the last day of the epidemic; vertical black line is the expected date for second turning point. Panel B- Number of infecteds shown till the end of June 2020. Panel C- Number of removed cases predicted till the last day of the epidemic. Panel D- Number of removed cases shown till the end of June 2020.



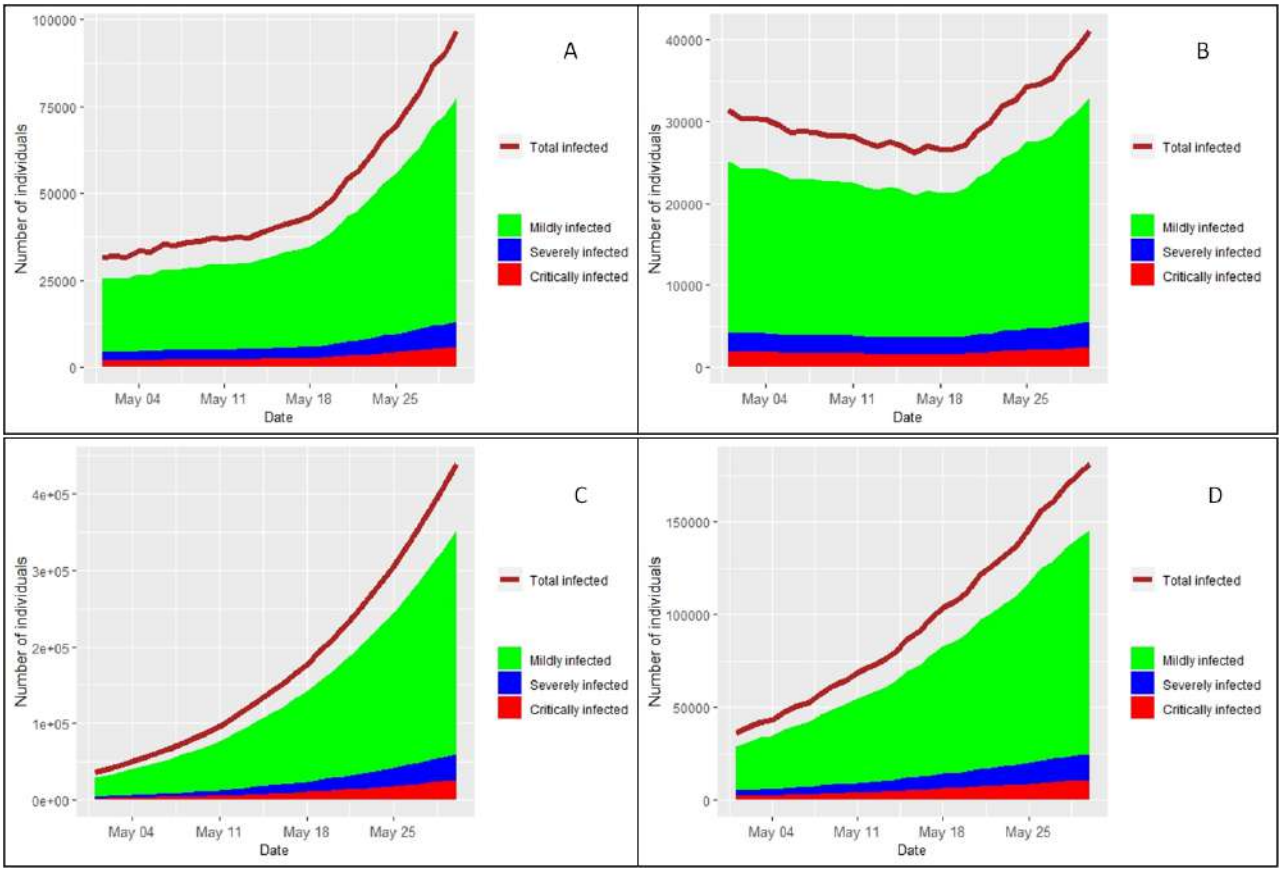
Graph 3: Predictions from exponential modifier function- (at $\hat{\gamma} = 0.0714$)- Panel A- Number of infecteds predicted till the last day of the epidemic; vertical black line is the expected date for second turning point. Panel B- Number of infecteds shown till the end of June 2020. Panel C- Number of removed cases predicted till the last day of the epidemic. Panel D- Number of removed cases shown till the end of June 2020.



Graph 4: Predictions from exponential modifier function- (at $\hat{\gamma} = 0.042$)- Panel A- Number of infecteds predicted till the last day of the epidemic; vertical black line is the expected date for second turning point. Panel B- Number of infecteds shown till the end of June 2020. Panel C- Number of removed cases predicted till the last day of the epidemic. Panel D- Number of removed cases shown till the end of June 2020.



Graph 5: Predictions of number of infecteds for May 2020- Panel A- Predictions using step-function modifiers (case 3) with 24-day infectious period; Panel B- Predictions using step-function modifiers (case 3) with 14-day infectious period; Panel C- Predictions using exponential modifier function with 24-day infectious period; Panel D- Predictions using exponential modifier function with 14-day infectious period.



Appendix-A

Detailed time-line of containment measures implemented in India

18 Jan 2020: Thermal screening of all passengers coming from China and Hong Kong started at three international airports.

30 Jan 2020: First Covid-19 case reported in India with travel history from Wuhan, China

04 Mar 2020: By this date thermal screening was initiated in a progressive manner for all international passengers (depending on the status of COVID-19 in the countries of origin) at all ports of entry (land, sea and air-ports) through various travel advisories.

13 Mar 2020 – 22 Mar 2020: During this period various state governments brought out notices for restricting social contacts - ban on public gatherings of any kind, shutting down of academic institutes, restrictions on public transportation, screening of interstate passengers at airports and complete lockdown of some states from 23 Mar 2020 till 31 Mar 2020.

15 Mar 2020 – 21 Mar 2020: Minimum 14-day quarantine made mandatory for all incoming travelers from Covid-19 infected countries (in progressive manner). Also, all Visa suspended till 15 Apr 2020

22 Mar 2020: Ban on all incoming international flights, except those already on transit. Suspension of all public transportation services, like metro, rail, domestic air, till 31 Mar 2020, except those that had started their journey before 22 Mar 2020.

25 Mar 2020: Complete lockdown till 14 Apr 2020. However, large-scale movements of migrant workers across various states started from 26 Mar 2020.

29 Mar 2020: Order issued to all state governments on this date to stop the migrants' movements and setting up of relief camps for those already in transit.

30 Mar 2020: Effective date of starting of measures to stop migrant movements across various states during complete lockdown till 14 Apr 2020.

15 Apr 2020: Complete lockdown extended till 03 May 2020 with new containment measures for hotspot areas.

20 Apr 2020: Apart from complete lockdown till 03 May 2020, new containment measures for hotspot areas, certain exemption for selected activities and limited movements of migrant workers with states/ UT .

Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 1, 2020 (New Series), pp 181-196

Critical Community Size for COVID-19: A Model Based Approach for Strategic Lockdown Policy

Sarmistha Das[1], Pramit Ghosh[2], Bandana Sen[3], Saumyadipta Pyne[4,5] and Indranil Mukhopadhyay[1]

[1]*Human Genetics Unit, Indian Statistical Institute, Kolkata, West Bengal, India*
[2]*Purulia Medical College, Purulia, West Bengal, India*
[3]*All India Institute of Hygiene & Public Health, Kolkata, West Bengal, India*
[4]*Public Health Dynamics Lab, and Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA*
[5]*Health Analytics Network, Pennsylvania, USA*

Received: 24 May 2020; Revised: 06 June 2020; Accepted: 08 June 2020

Abstract

Among the U.S. cities hit by the 1918 Spanish flu, social distancing played a pivotal role in flattening the pandemic curve. Similarly, to fight against COVID-19, restrictive mass quarantine or lockdown has been implemented as the most important controlling measure. India has already enforced a lockdown of 10 weeks and is extending the period depending on the current disease scenario. However, the idea that, if the susceptible population drops below certain threshold, the infection would naturally die out in small communities after a fixed time (following the outbreak), unless the disease is reintroduced from outside, was proposed by M. S. Bartlett in 1957. This threshold was termed as Critical Community Size (CCS).

We propose an Susceptible-Exposed-Infected-Recovered (SEIR) model that explains COVID-19 disease dynamics. Using our model, we have calculated state-specific Temporary Eradication of Spread Time (TEST) and CCS that would essentially determine the ideal number of lockdown days required and the size of quarantined population. With the given state-wise rates of death, recovery and other parameters, we have identified that, if at a place the total number of susceptible population drops below CCS, infection will cease to exist after a period of expected time to extinction (TTE), unless it is re-introduced from outside. The expected TTE suggests that the disease might take a long time to fade away from the human population in absence of pharmaceutical interventions. But we find that the disease might subside substantially after TEST. This would imply lockdown phases as much as TEST could be sufficient to contain COVID-19.

Key words: Critical community size; COVID-19; lockdown; quarantine; SEIR model; pandemic curve flattening.

AMS Subject Classifications: 00A05

Corresponding Author: Indranil Mukhopadhyay
Email: indranil@isical.ac.in

1. Introduction

In the face of COVID-19 pandemic, many countries have implemented restrictive mass quarantines or lockdown as the primary controlling measure to confine the number of secondary transmissions of the disease within countries. In absence of any specific medical treatment to treat the disease, patients are generally given only supportive care. Given the rapid Phase 3 transmission of the disease, health care systems of even developed countries are starting to face challenges within a week or two. Therefore, to prevent stage 4 transmission of the disease, along with many other countries India, which is densely populated, has resorted to complete lockdown already for more than 10 weeks and it is still counting. Available data confirms that the pandemic has already affected more than five million people in around 215 countries till date and already claimed more than 0.3 million lives across the world within approximately three months. After World Health Organisation (WHO) declared the outbreak as a pandemic, many countries initiated partial to complete lockdown as was done in some provinces of China after the outbreak started. By the end of March, one-third of the global population was under some form of lockdown.

Many countries implemented variable number of lockdown days, but none has come up with any magic figure for the ideal period of lockdown. No clear-cut guideline or rationale behind the number of lockdown days has been announced by any country or WHO till date to the best our knowledge. The initial phase of lockdown of 2-4 weeks was determined mostly on trial and error basis. The prediction on the number of trial lockdown days was possibly and partially based on the fact that an affected individual could be contagious in the first 14 days of contracting the disease and also on the information of the number of known positive cases at the time of taking decision.

The idea of quarantining a small group of people after an epidemic outbreak to arrest the disease dates back to 1950s when English statistician M.S. Bartlett introduced the term 'critical community size'. Probably the idea of such mathematical development was driven by the lessons of social distancing taught by the 1918 flu pandemic or Spanish flu. The cities with strong social distancing measures, successfully delayed its peak in deaths and maintained lower death rate (Markel et al., 2007). The flattening of 1918 flu pandemic curve that took approximately 24 weeks, was disrupted and the cities witnessed sharp increase in deaths when restrictions were temporarily relaxed after 8-10 weeks.

Bartlett (1957, 1960) proposed the idea that if the susceptible population is below some threshold, the infection is as likely as not to die out after a period of time (after the epidemic outbreak) in small communities, unless the disease is reintroduced from outside. Bartlett termed this threshold as Critical Community Size (CCS). Otherwise speaking, in absence of pharmaceutical interventions if the susceptible population that is quarantined together falls below CCS, the infection would die out from the population after a period of time unless the disease is re-introduced from outside. In the present context, CCS could guide government/health policy makers with an objective strategy of lockdown period as opposed to subjective trial and error phases of lockdown.

After an epidemic outbreak in a community, the infection persists long enough to engulf the entire susceptible population. Local extinction of the disease could be possible if the

susceptible population gets depleted. In large communities, the tendency of eventual damp down of the recurrent epidemics is balanced by random variability. But in small communities the infection would die out when the number of susceptible falls below a certain threshold, which is the CCS. Only a limited number of works (Nåsell, 2005; Anderson and Britton, 2000) including our work (under review) are available on CCS, may be because it involves complicated calculations even for simplest mathematical model viz. SI (S: Susceptible, I: Infected) model. However, since the actual extent of an epidemic can be assessed only retrospectively, it is essential to calculate the CCS for COVID-19 based on a realistic model that depends on the parameters which could be determined for a specific locality.

We propose an SEIR (S: Susceptible, E: Exposed, I: Infected, R: Recovered) model to explain the disease dynamics of COVID-19. We have derived with evidence the rationale behind the importance and extent of the lockdown period and also the number of people who could safely stay together in this lockdown phase. In absence of much prior knowledge on the disease, we have to rely on the mathematical predictions to combat the virus. In this article, we provide a cautionary note from the mathematical deductions, that this pandemic might take a very long time to fade away, in absence of any pharmaceutical interventions. Our work resonates the latest updates from WHO executive director stating "this virus may never go away". WHO also mentions that it may remain in the community as another endemic virus like human immunodeficiency virus (HIV). To have less disease transmission, WHO also stresses on enforcing withdrawal of the lockdown only when the day-to-day number of COVID-19 cases reaches the lowest possible level; otherwise, the transmission may accelerate (https://www.aninews.in/news/world/europe/who-executive-director-says-coronavirus-may-never-go-away20200514012424/).

But there is always a ray of hope. Apart from the fact that, we may learn much from Spanish flu, SARS, and MERS outbreaks, our deduction suggests that there should be no reason to panic as the lockdown, if properly followed, could contain the disease. Although we have to bear the burden of slow economic recovery or even a recession, the COVID-19 epidemic could be controlled and hopefully it would not cause a more severe public health emergency in the near future.

2. Methods

We propose an SEIR model to explain the dynamics of COVID-19 infection. The entire population is divided into four compartments. These compartments are mutually exclusive in the sense that no person can belong to more than one compartment at any time point. The four compartments are: susceptible individuals (S), individuals with and without symptoms of the disease but not yet tested positive for COVID-19 (E), infected individuals who are clinically tested positive (I), and individuals who are known to have recovered from the disease (R). Note that an individual belonging to class E may transmit the disease during

the incubation period. Under this situation, we consider the model as:

$$\frac{dS}{dt} = \Lambda - \beta\left(I(t) + \phi E(t)\right)\frac{S(t)}{N} - \mu S(t) \tag{1}$$

$$\frac{dE}{dt} = \beta\left(I(t) + \phi E(t)\right)\frac{S(t)}{N} - (\gamma + \mu)E(t) \tag{2}$$

$$\frac{dI}{dt} = \gamma E(t) - (\delta + \mu + d)I(t) \tag{3}$$

$$\frac{dR}{dt} = \delta I(t) - \mu R(t) \tag{4}$$

Here β (or $\beta\phi$) represents the contact rate for COVID-19 transmission from infected (or exposed) to susceptible individuals, an individual in E moves to I at the rate γ, δ is the recovery rate, d is death rate due to the disease and μ is the natural death rate in the population. Moreover, $\Lambda = \mu N(t)$ where $N(t)$ is the population size at time t.

Next we calculate the basic reproduction number (R_0) defined as the expected number of secondary cases produced by a single infection in a completely susceptible population. We calculate R_0 for the above model using next generation matrix $G = FV^{-1}$, where, $F = \left[\frac{\partial F_i(x_0)}{\partial x_j}\right]$ and $V^{-1} = \left[\frac{\partial V_i(x_0)}{\partial x_j}\right]$. Here, F_is are the new infections in the system, while V_i denotes the transfer of infections from one compartment to another and x_0 is the disease-free equilibrium state (section 2.2). In our model, $F = \begin{bmatrix} \beta\phi\frac{S}{N} & \beta\frac{S}{N} \\ 0 & 0 \end{bmatrix}$ and $V^{-1} = \begin{bmatrix} \gamma+\mu & 0 \\ -\gamma & \delta+\mu+d \end{bmatrix}$. R_0 is defined as the maximum eigen value of the matrix G. Based on the above model R_0 will be:

$$R_0 = \frac{\beta(\phi(\delta + \mu + d) + \gamma)}{(\gamma + \mu)(\delta + \mu + d)} \tag{5}$$

2.1. Stochastic model and quasi-stationarity

First we note the nature of transition and the respective transition rates from one compartment to another (Table 1).

We construct the fully stochastic version of the model in (1)-(4) using the transition rates in Table 1. Denoting $s = S/N$, $e = E/N$, $i = I/N$, $r = R/N$, the Kolmogorov forward equations for this process can be written as follows:

$$\begin{aligned} p'_{s,e,i,r}(t) =& \lambda_1 p_{s-1,e,i,r} + \lambda_2 p_{s+1,e,i,r} + \lambda_3 p_{s+1,e-1,i,r} + \lambda_4 p_{s+1,e,i-1,r} \\ &+ \lambda_5 p_{s,e+1,i,r} + \lambda_6 p_{s,e+1,i-1,r} + \lambda_7 p_{s,e,i+1,r-1} + \lambda_8 p_{s,e,i+1,r} \\ &+ \lambda_9 p_{s,e,i,r+1} - \kappa(s,e,i,r) p_{s,e,i,r} \end{aligned} \tag{6}$$

where $\kappa(s,e,i,r) = \sum_{j=1}^{9} \lambda_j(s,e,i,r)$.

We use Kolmogorov forward equations in order to find the expected time to extinction (TTE) and evaluate CCS based on our model. Now, conditioning on non-extinction, we

Table 1: Chart of transition rates

Event	**Transition**	**Transition rate**
Immigration of Susceptibles	$(s,e,i,r) \to (s+1,e,i,r)$	$\lambda_1 = \lambda_1(s,e,i,r) = \mu N$
Death of Susceptibles	$(s,e,i,r) \to (s-1,e,i,r)$	$\lambda_2 = \lambda_2(s,e,i,r) = \mu s$
Susceptible (S) to Exposed (E)	$(s,e,i,r) \to (s-1,e+1,i,r)$	$\lambda_3 = \lambda_3(s,e,i,r) = \beta\phi se/N$
Susceptible (S) to Infected (I)	$(s,e,i,r) \to (s-1,e,i+1,r)$	$\lambda_4 = \lambda_4(s,e,i,r) = \beta si/N$
Death of Exposed	$(s,e,i,r) \to (s,e-1,i,r)$	$\lambda_5 = \lambda_5(s,e,i,r) = \mu e$
Exposed (E) to Infected (I)	$(s,e,i,r) \to (s,e-1,i+1,r)$	$\lambda_6 = \lambda_6(s,e,i,r) = \gamma e$
Infected (I) to Recovered (R)	$(s,e,i,r) \to (s,e,i-1,r+1)$	$\lambda_7 = \lambda_7(s,e,i,r) = \delta i$
Death of Infected	$(s,e,i,r) \to (s,e,i-1,r)$	$\lambda_8 = \lambda_8(s,e,i,r) = (\mu+d)i$
Death of Recovered	$(s,e,i,r) \to (s,e,i,r-1)$	$\lambda_9 = \lambda_9(s,e,i,r) = \mu r$

have,

$$q_{s,e,i,r}(t) = P\left[S(t)=s, E(t)=e, I(t)=i, R(t)=r | E(t)\neq 0, I(t)\neq 0\right] = \frac{p_{s,e,i,r}(t)}{1-p_{\bullet 00\bullet}(t)}$$

where $p_{\bullet 00\bullet}(t) = \sum_{s=0}^{\infty}\sum_{r=0}^{\infty} P\left[S(t)=s, E(t)=e, I(t)=i, R(t)=r\right] = \sum_{s=0}^{\infty}\sum_{r=0}^{\infty} p_{s,0,0,r}(t)$.
Now, differentiating $q_{s,i,c,a}(t)$ with respect to t, we have,

$$q'_{s,e,i,r}(t) = \frac{p'_{s,e,i,r}(t)}{1-p_{\bullet 00\bullet}(t)} + \frac{p_{s,e,i,r}(t)}{\left(1-p_{\bullet 00\bullet}(t)\right)^2}.p'_{\bullet 00\bullet}(t). \tag{7}$$

Now, from (6), we have, after simplification,

$$p_{\bullet 00\bullet}(t) = \mu p_{\bullet 1\,0\,\bullet}(t) + (\delta+\mu+d)p_{\bullet 0\,1\,\bullet}(t) = p_{\bullet}^{(d,\mu,\delta)}(t) \;\; (say). \tag{8}$$

From (7-8) we have,

$$q'_{s,e,i,r}(t) = \frac{p'_{s,e,i,r}(t)}{1-p_{\bullet 00\bullet}(t)} + \frac{p_{s,e,i,r}(t)}{\left(1-p_{\bullet 00\bullet}(t)\right)}.q_{\bullet}^{(d,\mu,\delta)}(t) \;\text{ where } q_{\bullet}^{(d,\mu)}(t) = \frac{p_{\bullet}^{(d,\mu,\delta)}(t)}{1-p_{\bullet 0\,0\,0}(t)} \tag{9}$$

Now, $q'_{s,e,i,r}(t) = 0$

$$\implies p'_{s,e,i,r}(t) = -\frac{p_{s,e,i,r}(t)}{\left(1-p_{\bullet 0\,0\,\bullet}(t)\right)}.q_{\bullet}^{(d,\mu,\delta)}(t)(1-p_{\bullet 0\,0\,\bullet}(t)) = -q_{\bullet}^{(d,\mu,\delta)}(t)p_{s,e,i,r}(t)$$

$$\implies p_{s,e,i,r}(t) = ce^{-q_{\bullet}^{(d,\mu,\delta)}(t).t} = q_{\bullet}^{(d,\mu,\delta)}(0)e^{-q_{\bullet}^{(d,\mu,\delta)}(t).t} \tag{10}$$

Let τ_Q be the TTE when the initial distribution equals the quasi-stationarity distribution [Nåsell, 2005]. Hence for stationary distribution,

$$E(\tau_Q) = \frac{1}{q_{\bullet}^{(d,\mu,\delta)}}. \tag{11}$$

2.2. Equilibrium points

The disease-free equilibrium is obtained as: $\Sigma_0 = (S^0, I^0, C^0, A^0) = \Big(\frac{\Lambda}{\mu}, 0, 0, 0\Big)$.

To find the other endemic equilibrium, if exists, we put $N = N(0)$, $x_1(t) = S(t)/N$, $x_2(t) = E(t)/N$, $x_3(t) = I(t)/N$, and $x_4(t) = R(t)/N$. Then equilibrium point is obtained by equating the first differentiation to zero, i.e.

$$x_1'(t) = \mu - \beta\Big(x_3(t) + \phi x_2(t)\Big)x_1(t) - \mu x_1(t) = 0 \tag{12}$$

$$x_2'(t) = \beta\Big(x_3(t) + \phi x_2(t)\Big)x_1(t) - (\gamma + \mu)x_2(t) = 0 \tag{13}$$

$$x_3'(t) = \gamma x_2(t) - (\delta + \mu + d)x_3(t) = 0 \tag{14}$$

$$x_4'(t) = \delta x_3(t) - \mu x_4(t) = 0 \tag{15}$$

For simplicity we use the notations: $x_j(t) = x_j$ for $j = 1, \ldots, 4$.

Then solving (12) - (15), we have the endemic equilibrium as:

$$\hat{x}_1 = \frac{(\gamma + \mu)(\delta + \mu + d)}{\beta(\gamma + \phi(\delta + \mu + d))} = \frac{1}{R_0} \tag{16}$$

$$\hat{x}_2 = \frac{\mu(1 - \hat{x}_1)}{\gamma + \mu} = \frac{\mu}{\gamma + \mu}(1 - \frac{1}{R_0}) \tag{17}$$

$$\hat{x}_3 = \frac{\gamma\mu(1 - \hat{x}_1)}{(\gamma + \mu)(\delta + \mu + d)} = \frac{\gamma\mu}{(\gamma + \mu)(\delta + \mu + d)}(1 - \frac{1}{R_0}) \tag{18}$$

$$\hat{x}_4 = \frac{\gamma\delta(1 - \hat{x}_1)}{(\gamma + \mu)(\delta + \mu + d)} = \frac{\gamma\delta}{(\gamma + \mu)(\delta + \mu + d)}(1 - \frac{1}{R_0}) \tag{19}$$

2.3. Diffusion approximation

Stationary distribution of an epidemic process may be approximated with a specified multivariate normal distribution using Ornstein-Uhlenbeck process when the population size N is very large and $R_0 \geq 1$. This approximation is valid only in absence of any infection. We derive an approximate distribution of the quasi-stationarity by limiting Ornstein-Uhlenbeck process (Nåsell, 2005). We consider a diffusion approximation to the stochastic version of SEIR model.

Let the changes in the scaled state variables x_1, x_2, x_3, and x_4 during the time interval be denoted by δx_1, δx_2, δx_3, and δx_4 respectively, where $\delta x_i(t) = x_i(t + \delta t) - x_i(t)$, $i = 1, 2, 3, 4$.

Under the assumptions of the original process on sequence of transitions, we evaluate the mean vector and variance-covariance matrix for δx_i $(i = 1, 2, 3, 4)$ during the time interval $(t, t + \delta t)$ as follows.

First assume that we are in the state (S, E, I, R). Then the possible transitions from this state are:
(a) S increases by 1 at the rate μ

(b) S decreases by 1 at the rate μS
(c) S decreases by 1 and E increases by 1 at the rate $\beta\phi SE/N + \beta SI/N$
(d) E decreases by 1 and I increases by 1 at the rate γE
(e) E decreases by 1 at the rate μE
(f) I decreases by 1 at the rate $(\mu + d)I$
(g) I decreases by 1 and R increases by 1 at the rate δI
(h) R decrease by 1 at the rate μR.

The random variable δx_1 equals $\frac{1}{N}$ in case (1), $-\frac{1}{N}$ in cases (2), (3), and 0 in other cases. Similarly, δx_2 equals $\frac{1}{N}$ in case (3), $-\frac{1}{N}$ in cases (4), (5), and 0 in other cases. δx_3 equals $\frac{1}{N}$ in case (4), $-\frac{1}{N}$ in cases (6), (7), and 0 in other cases. δx_4 equals $\frac{1}{N}$ in case (7), $-\frac{1}{N}$ in case (8), and 0 in other cases.

$$\text{Then, } E(\delta \boldsymbol{x}) = b(\boldsymbol{x})\delta t + o(\delta t)$$

$$\text{where} \quad b(\boldsymbol{x}) = \begin{pmatrix} \mu - \beta(x_3 + \phi x_2)x_1 - \mu x_1 \\ \beta(x_3 + \phi x_2)x_1 - (\gamma + \mu)x_2 \\ \gamma x_2 - (\delta + \mu + d)x_3 \\ \delta x_3 - \mu x_4 \end{pmatrix} \tag{20}$$

Now to derive the variance-covariance matrix we find the Jacobian matrix of $b(\boldsymbol{x})$ at point $\boldsymbol{x}$,

$$B(\boldsymbol{x}) = \frac{\partial b(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} -\beta(x_3 + \phi x_2) - \mu & -\beta\phi x_1 & -\beta x_1 & 0 \\ \beta(x_3 + \phi x_2) & \beta\phi x_1 - (\gamma + \mu) & \beta x_1 & 0 \\ 0 & \gamma & -(\delta + \mu + d) & 0 \\ 0 & 0 & \delta & -\mu \end{pmatrix}$$

Approximating $B(\boldsymbol{x})$ at equilibrium point $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)$ by $B(\hat{\boldsymbol{x}})$, we get,

$$B(\hat{\boldsymbol{x}}) = \begin{pmatrix} -\mu R_0 & -\beta\phi x_1 & -\beta x_1 & 0 \\ \mu(R_0 - 1) & \beta\phi x_1 - (\gamma + \mu) & \beta x_1 & 0 \\ 0 & \gamma & -(\delta + \mu + d) & 0 \\ 0 & 0 & \delta & -\mu \end{pmatrix}$$

Therefore, variance-covariance matrix of $\delta \boldsymbol{x} = (\delta x_1, \delta x_2, \delta x_3, \delta x_4)'$ is,
$V(\delta \boldsymbol{x}) = \frac{1}{N} S(\boldsymbol{x})\delta t + o(\delta t)$ where,

$$S(\boldsymbol{x}) = \frac{1}{N} \begin{pmatrix} \beta(x_3 + \phi x_2)x_1 + \frac{\mu}{N} + \mu x_1 & -\beta(x_3 + \phi x_2)x_1 & 0 & 0 \\ -\beta(x_3 + \phi x_2)x_1 & (\gamma + \mu)x_2 + \beta(x_3 + \phi x_2)x_1 & -\gamma x_2 & 0 \\ 0 & -\gamma x_2 & (\delta + \mu + d)x_3 + \gamma x_2 & -\delta x_3 \\ 0 & 0 & -\delta x_3 & \delta x_3 + \mu x_4 \end{pmatrix}$$

Again approximating $S(\boldsymbol{x})$ by $S(\hat{\boldsymbol{x}})$, where $\hat{\boldsymbol{x}}$ is the equilibrium point, we obtain,

$$S(\hat{\boldsymbol{x}}) = \frac{1}{N} \begin{pmatrix} \frac{\mu}{N} + \mu & -\mu(1 - \frac{1}{R_0}) & 0 & 0 \\ -\mu(1 - \frac{1}{R_0}) & 2\mu(1 - \frac{1}{R_0}) & \frac{-\mu\gamma}{\gamma+\mu}(1 - \frac{1}{R_0}) & 0 \\ 0 & \frac{-\mu\gamma}{\gamma+\mu}(1 - \frac{1}{R_0}) & 2\frac{\mu\gamma}{\gamma+\mu}(1 - \frac{1}{R_0}) & \frac{-\delta\mu\gamma}{(\gamma+\mu)(\delta+\mu+d)}(1 - \frac{1}{R_0}) \\ 0 & 0 & \frac{-\delta\mu\gamma}{(\gamma+\mu)(\delta+\mu+d)}(1 - \frac{1}{R_0}) & 2\frac{-\delta\mu\gamma}{(\gamma+\mu)(\delta+\mu+d)}(1 - \frac{1}{R_0}) \end{pmatrix}$$

For large N, the process $\sqrt{N}(\boldsymbol{x}(t) - \hat{\boldsymbol{x}})$ is approximated by a multivariate Ornstein-Uhlenbeck (O-U) process with a local drift matrix $B(\hat{\boldsymbol{x}})$ and local variance-covariance matrix $S(\hat{\boldsymbol{x}})$.

The stationary distribution of this O-U process approximates the quasi stationary distribution. It is approximately normal with mean zero and variance-covariance matrix Σ, where Σ is obtained by solving

$$B(\hat{\boldsymbol{x}})\Sigma + \Sigma B'(\hat{\boldsymbol{x}}) = -S(\hat{\boldsymbol{x}}). \tag{21}$$

Exact analytical solution for Σ is not straightforward (Anderson and Britton, 2000). Since we are interested in calculating the CCS, we can easily solve the equation (21) numerically given the parameter values and the equilibrium point.

Let σ_{ij} be the solution for the (i,j)th element of Σ, where $i, j = 1, \ldots, 4$. Diffusion approximation guides us to consider the joint distribution of $x_1(t), x_2(t), x_3(t), x_4(t)$ as four-variate normal distribution with appropriate mean and variance-covariance matrix i.e.

$$\sqrt{N}(\boldsymbol{x}(t) - \hat{\boldsymbol{x}}) \sim N_4(\mathbf{0}, \Sigma), \text{ with } \boldsymbol{x}(t) = (x_1(t), x_2(t), x_3(t), x_4(t))', \hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4),$$

$$\text{and } \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} \tag{22}$$

An approximation for quasi-stationary distribution is obtained from truncated multivariate normal distribution. Thus in order to evaluate expected time to extinction and subsequently the CCS, we use results from conditional truncated multivariate normal distribution. Now define $\mu_2^* = \hat{x_2} + \frac{\sigma_{23}}{\sigma_{33}}(x_3 - \hat{x_3})$, $\sigma_{22}^* = \sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}^2}$.

To calculate $p_{\bullet 10 \bullet}$ (or $p_{\bullet 01 \bullet}$ or $p_{\bullet 00 \bullet}$) first note that for large N, $\sqrt{N}(\boldsymbol{x} - \hat{\boldsymbol{x}})$ approximately follows a four-variate multivariate normal distribution with mean zero and covariate matrix Σ, as obtained form equation (22). Now, we shall show that these terms contain product of $\frac{\phi(\nu)}{\Phi(\nu)}$ terms. Since N is unknown, we cannot evaluate its values exactly. Thus we use another approximation to $\frac{\phi(\nu)}{\Phi(\nu)}$ based on a logistic function only to make the calculation relatively simple. Using the idea that $|\sigma(\beta x) - \Phi(x)|$ is minimum when $\beta = \frac{16x}{15}$ (Birnbaum, 1963; Haley, 1952)) and putting $\sigma(z) = \frac{1}{1+e^{-z}}$ and $\beta = \frac{16}{15}\frac{\pi}{\sqrt{3}}$, we approximate $\frac{\phi(.)}{\Phi(.)}$ as,

$$\begin{aligned} \frac{\phi(\nu)}{\Phi(\nu)} &= \frac{\phi(\nu)}{\int_{-\infty}^{\nu} \phi(x)dx} = \frac{\beta\phi(\nu)}{\int_{-\infty}^{\beta\nu} \phi(\frac{y}{\beta})dy} \approx \frac{\beta\phi(\nu)}{\sigma(\beta\nu)} \text{ (Williams, 2005)} \\ &= \beta\phi(\nu)(1 + e^{-\beta\nu}) \approx \beta\Big[\frac{1 + cos(\nu)}{2\pi}\Big](1 + e^{-\beta\nu}) \text{ (Raab, 1961)} \\ &\approx \beta\frac{1 + cos(\nu)}{2\pi} \text{as, } \nu \to \infty \text{ as, } N \to \infty \end{aligned}$$

Therefore, for $y > 0$, $\Phi(y+h) - \Phi(y) \approx h.\phi(y)$, we obtain

$$
\begin{aligned}
p_{\bullet 10\bullet} &= \sum_{s=0}^{\infty}\sum_{r=0}^{\infty} P(S=s, E=1, I=0, R=r) = P(E=1, I=0) \\
&\approx P(0.5 < Nx_2(t) \le 1, 0 \le Nx_3(t) \le 0.5) \\
&\approx P(0.5 < Nx_2(t) \le 1 | 0 \le Nx_3(t) \le 0.5).P(0 \le Nx_3(t) \le 0.5) \\
&= \frac{\Phi(\frac{\sqrt{N}(\frac{1}{N}-\mu_2^*)}{\sqrt{\sigma_{22}^*}}) - \Phi(\frac{\sqrt{N}(\frac{1}{2N}-\mu_2^*)}{\sqrt{\sigma_{22}^*}})}{1-\Phi(\frac{\sqrt{N}(0-\mu_2^*)}{\sqrt{\sigma_{22}^*}})} . \frac{\Phi(\frac{\sqrt{N}(\frac{1}{2N}-\hat{x_3})}{\sqrt{\sigma_{33}}}) - \Phi(\frac{\sqrt{N}(0-\hat{x_3})}{\sqrt{\sigma_{33}}})}{1-\Phi(\frac{\sqrt{N}(0-\hat{x_3})}{\sqrt{\sigma_{33}}})} \\
&\approx \frac{\frac{0.5}{N}}{\frac{\sqrt{\sigma_{22}^*}}{\sqrt{N}}} \frac{\phi(\frac{\sqrt{N}(\frac{1}{2N}-\mu_2^*)}{\sqrt{\sigma_{22}^*}})}{\Phi(\frac{\sqrt{N}\mu_2^*}{\sqrt{\sigma_{22}^*}})} . \frac{\frac{0.5}{N}}{\frac{\sqrt{\sigma_{33}}}{\sqrt{N}}} \frac{\phi(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})}{\Phi(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})} \\
&\approx \frac{1}{2\sqrt{N}} \frac{1}{\sqrt{\sigma_{22}^*}} \beta \frac{1+cos(\frac{\sqrt{N}\mu_2^*}{\sqrt{\sigma_{22}^*}})}{2\pi} . \frac{1}{2\sqrt{N}} \frac{1}{\sqrt{\sigma_{33}}} \beta \frac{1+cos(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})}{2\pi}
\end{aligned}
$$

Thus we have,

$$
\begin{aligned}
p_{\bullet 10\bullet} &= \frac{1}{\sqrt{\sigma_{22}^*}} \frac{1}{2\sqrt{N}} \beta \frac{1+cos(\frac{\sqrt{N}\mu_2^*}{\sqrt{\sigma_{22}^*}})}{2\pi} \frac{1}{\sqrt{\sigma_{33}}} \frac{1}{2\sqrt{N}} \beta \frac{1+\cos(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})}{2\pi} \text{ when } x_3 = 0 \\
p_{\bullet 01\bullet} &= \frac{1}{\sqrt{\sigma_{22}^*}} \frac{1}{2\sqrt{N}} \beta \frac{1+\cos(\frac{\sqrt{N}\mu_2^*}{\sqrt{\sigma_{22}^*}})}{2\pi} \frac{1}{\sqrt{\sigma_{33}}} \frac{1}{2\sqrt{N}} \beta \frac{1+\cos(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})}{2\pi} \text{ when } x_3 = 1 \\
p_{\bullet 00\bullet} &= \frac{1}{\sqrt{\sigma_{22}^*}} \frac{1}{2\sqrt{N}} \beta \frac{1+\cos(\frac{\sqrt{N}\mu_2^*}{\sqrt{\sigma_{22}^*}})}{2\pi} \frac{1}{\sqrt{\sigma_{33}}} \frac{1}{2\sqrt{N}} \beta \frac{1+\cos(\frac{\sqrt{N}\hat{x_3}}{\sqrt{\sigma_{33}}})}{2\pi} \text{ when } x_3 = 0
\end{aligned}
$$

Once we find $q_{\bullet}^{(d,\mu,\delta)}$, we have an expression for expected time to extinction $\hat{E}(\tau_Q)$ using (11). Clearly, $\hat{E}(\tau_Q)$ will be a function of N. However, N is unknown. To obtain this N which is nothing but CCS, we equate median time to extinction with the quasi-period ($\hat{T_0}$) (Nåsell, 2005). The quasi-period is obtained as $\hat{T_0} = \frac{2\pi}{\theta}$ where θ is the angular frequency. The angular frequency is determined by linearisation about the critical point that corresponds to the endemic infection level (Dietz, 1975). The value of N from $E(\tau_Q)log2 = \hat{T_0}$ will be the CCS value (Nåsell, 2005).

We find the quasi-period of the oscillation about the critical point using linearisation method (Dietz, 1975). Note that for our model, the linearised system about the equilibrium point $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)'$ can be written as:

$$
\frac{d\boldsymbol{x}^*}{dt} = \begin{pmatrix}
-\frac{\beta\mu(1-\frac{1}{R_0})}{\gamma+\mu}(\phi + \frac{\gamma}{\delta+\mu+d}) & -\frac{\beta\phi}{R_0} & -\frac{\beta}{R_0} & 0 \\
\frac{\beta\mu(1-\frac{1}{R_0})}{\gamma+\mu}(\phi + \frac{\gamma}{\delta+\mu+d}) & \frac{\beta\phi}{R_0} - (\gamma+\mu) & \frac{\beta}{R_0} & 0 \\
0 & \gamma & -(\delta+\mu+d) & 0 \\
0 & 0 & \delta & -\mu
\end{pmatrix} \boldsymbol{x}^* \qquad (23)
$$

where $\boldsymbol{x}^* = \boldsymbol{x} - \hat{\boldsymbol{x}}$.

Now we can find the eigen values of the matrix in (23) and find the angular frequency, provided there are imaginary roots. Putting the values of the parameters, we can find the angular frequency (θ). The quasi-period is obtained as $\hat{T}_0 = \frac{2\pi}{\theta}$, which is independent of N.

From the relation $\hat{E}(\tau_Q)\log 2 = \hat{T}_0$ (Nåsell, 2005), we can solve for N, which is the CCS. Since we are dealing with a system consisting of more than two equations, the calculations become very complicated. Hence we find an approximate value of CCS numerically.

In a nutshell, our method at first develops a fully stochastic model corresponding to the deterministic model (1)-(4); then assuming quasi-stationarity and non-extinction of infection, expected time to extinction TTE ($\equiv E(\tau_Q)$) of the disease is derived. $E(\tau_Q)$ involves some probability terms that we evaluate using diffusion approximation of the scaled state variables (S, E, I, R). τ_Q is a function of the CCS. We derive quasi-period $\hat{T}_0$ in terms of angular frequency that is obtained using linearised system at equilibrium points. Then using the relation $E(\tau_Q)\log 2 = \hat{T}_0$ (Nåsell, 2005), we could finally evaluate the CCS for the disease dynamics of COVID-19.

3. Results

For COVID-19 transmission, we have calculated the CCS and TTE of the disease based on our proposed SEIR model. We note that the value of the CCS is approximate as we have applied some mathematical approximation while applying diffusion approximation to find the quasi-stationary distribution. The value of CCS for a community or a country, depends on its parameters which we deduce from the available information on COVID-19 till date.

We apply our method to different states in India. However, this is a general method and can be applied to any country or locality provided the values of the parameters are available. Actual fatality rate due to any epidemic could only be calculated after the epidemic gets over. But in the middle of the pandemic, it is difficult to assess. So we determine the state-specific death rate (d) at four time points at an interval of seven days based on the number of deaths in the duration of May $15-21$, May $8-14$, May $1-7$, and April $24-30$ and the total number of newly infected individuals during 7 days prior to these dates respectively. From hereon, we denote the four time points as T1, T2, T3, and T4 respectively.

Different countries have implemented varying criteria of discharging COVID-19 patients from hospitals making the actual recovery rate very difficult to calculate amid the pandemic. It is yet unknown whether all the discharged patients have fully recovered from the disease or some of them would get sick again, shortly afterwards. So we have assumed the recovery rate (δ) again at four time points at an interval of seven days based on the number of recovered patients during May $15-21$, May $8-14$, May $1-7$, and April $24-30$ and the total number of newly infected individuals during dates April 24-May 7, April $17-30$, April $10-23$, and April $3-16$ respectively. These Indian state-specific numbers for newly infected cases, death, and recovery at four time points are obtained from https://api.covid19india.org/.

Another very tricky and state dependent parameter is the rate of detection of positive cases from among the exposed pool of people, *i.e.* percentage of exposed people that are

actually tested to be COVID-19 positive. In absence of enough manpower and testing kits in this dire situation, we will not be able to know the actual proportion (γ) of the exposed who could later on become a COVID-19 patient. For calculating γ, we obtained daily state-wise test positivity rate (TPR) (https://api.covid19india.org/) that is the rate at which the exposed individuals are tested and reported to be infected daily. We calculate 7-day average TPR (using geometric mean) for each state at three time points mentioned above. Next we calculated at each time point the geometric mean of 7-day average TPRs from all states and took the maximum value as γ. We obtain, $\gamma \approx 0.04$. The rationale behind taking maximum value is due to the fact that, in India the number of tests done per million is 1823 (as on May 21, 2020), which is much less than many other countries (https://www.worldometers.info/coronavirus/). It is possible that if we had enough tests, the actual TPR could be different, rather higher. The per day rate of natural death that stabilises the population under normal scenario is $\mu = \frac{1}{70}\frac{1}{365} = 0.0000391$ (assuming average longevity of an Indian is 70 years). In absence of actual contact rate (β), we have assumed $\beta = 1.1$ (Senapati et al., 2020). Another difficult parameter to obtain is the contact rate for COVID-19 transmission from exposed to susceptible individuals ($\beta\phi$). We calculated a range of $\beta\phi$ values for all Indian states at all time points and assumed the most common rate. We obtained $\beta\phi \approx 0.0011$.

Table 2: CCS and TEST for Indian states at different time points

	T4			T3			T2			T1		
State	CCS	R_0	TEST	CCS	R_0	TEST	CCS	R_0	TEST	CCS	R_0	TEST
DL	230	5.081	19-23	350	2.186	16-19	380	1.836	15-19	720	1.563	18-21
GJ	170	2.389	13-16	350	2.121	16-19	280	1.782	14-17	70	2.575	9-13
JK	490	2.140	17-20	160	2.320	13-16	240	2.201	14-17	170	1.926	12-15
KA	40	2.416	6-10	460	1.908	17-20	380	2.808	17-20	140	2.518	12-16
MP	320	3.615	17-19	190	1.785	12-15	180	1.664	11-15	70	2.475	9-12
MH	190	2.992	14-17	170	3.317	14-17	60	2.733	9-12	260	2.166	14-18
RJ	590	2.401	19-22	670	1.616	18-21	260	2.088	14-18	310	1.913	15-18
TN	410	2.076	16-19	260	3.168	16-18	550	1.682	17-20	3380	1.037	23-26
TG	270	3.131	16-19	30	2.199	1-7	740	1.402	17-21	270	1.916	14-18
UP	570	2.124	18-21	470	1.715	16-20	450	1.855	17-20	950	1.483	19-23
WB	50	3.644	9-12	400	1.573	15-19	2510	1.179	22-26	70	2.667	10-13
PB	60	4.053	11-13	180	3.436	14-17	160	4.183	15-17			
HR	260	2.764	15-18	230	4.035	16-18				360	1.625	15-18
BR							160	2.072	12-15	470	2.632	18-21
CH				400	3.147	17-20	270	3.457	16-19			
OR				160	2.124	12-16						
UT				140	3.690	14-16						
AP	90	2.340	10-13	970	1.285	18-22						
KL	80	1.784	7-11									
AS										950	1.278	18-22
HP										2330	1.178	15-19

TEST (in weeks) gives a range of lockdown period across different Indian states at all time points; DL: Delhi, GJ: Gujarat, HR: Haryana, JK: Jammu and Kashmir, KA: Karnataka, MP: Madhya Pradesh, MH: Maharastra, PB: Punjab, RJ: Rajasthan, TN: Tamil Nadu, TG: Telangana, UP: Uttar Pradesh, WB: West Bengal, AP: Andhra Pradesh, BR: Bihar, CH: Chandigarh, OR: Odisha, UT: Uttarakhand, KL: Kerala

Table 3: TEST with minimum CCS across different time points for Indian states

States	min CCS	$T4_{lower}$	$T3_{lower}$	$T2_{lower}$	$T1_{lower}$	$T4_{upper}$	$T3_{upper}$	$T2_{upper}$	$T1_{upper}$
DL	230	16	14	13	12	19	17	16	15
GJ	170	13	13	12	9	16	16	15	13
JK	160	12	13	13	12	15	16	16	15
KA	40	6	3	7	6	10	8	10	10
MP	180	14	12	11	9	17	15	15	12
MH	60	9	10	9	7	12	13	12	11
RJ	260	15	13	14	14	18	17	18	17
TN	260	14	16	13	9	17	18	17	13
TG	30	6	1	1	1	9	7	3	6
UP	450	17	16	17	15	20	19	20	19
WB	50	9	2	1	8	12	8	5	11
PB	60	11	10	11	1	13	13	13	4
HR	230	15	16	6	12	18	18	10	16
BR	160		6	12	13		9	15	16
CH	270		16	16			19	19	
OR	160		12	6	2		16	10	6
UT	140		14			16			
AP	90	10	5			13	9		
KL	160	7				11			

$T1_{lower}$ (or $T2_{lower}$ or $T3_{lower}$): minimum TEST (in weeks) required at time point T1 (or T2 or T3), $T1_{upper}$ (or $T2_{upper}$ or $T3_{upper}$): maximum TEST (in weeks) required at time point T1 (or T2 or T3); DL: Delhi, GJ: Gujarat, HR: Haryana, JK: Jammu and Kashmir, KA: Karnataka, MP: Madhya Pradesh, MH: Maharastra, PB: Punjab, RJ: Rajasthan, TN: Tamil Nadu, TG: Telangana, UP: Uttar Pradesh, WB: West Bengal, AP: Andhra Pradesh, BR: Bihar, CH: Chandigarh, OR: Odisha, UT: Uttarakhand, KL: Kerala

Table 4: CCS and TEST for India at different time points

	δ	d	CCS	R_0	TEST
T4	0.4799	0.0327	620	2.1793	23-27
T3	0.4502	0.0321	470	2.3142	21-25
T2	0.4717	0.0305	690	2.2243	23-28
T1	0.5203	0.0270	380	2.0439	19-24

We have computed state-wise CCS at time points T1, T2, T3, and T4 (Table 2). For a few states, data were missing at some or all time points and so, CCS could not be obtained. While calculating CCS we observed that the expected TTE in absence of specific treatment or vaccine, is very large. But the most interesting observation from our study is that complete lockdowns or restrictive quarantines for a definite period might eradicate the disease almost completely. We term this period as Temporary Eradication of Spread Time (TEST) for the disease, which is immensely less than the expected TTE for the disease. Although the

disease might continue to exist for a very long time on the planet in absence of pharmaceutical interventions, soothing part from our deduction is that the virus might not be able to create any havoc on its return. Our study thus provides a rationale behind the determination of the lockdown period in different Indian states going through the catastrophic effect of the pandemic. This work may aid public health workers to strategise lockdown policies.

For example, we find in Table 2, CCS of Delhi (DL) at time T2 is 380 and TEST is $15-19$. This would mean that based on the demographic figures corresponding to time T2, if the susceptible population (or community size of quarantined people) of DL is below 380, the infection will subside substantially after around $15-19$ weeks of mass quarantine/ restrictive lockdown, unless it is re-introduced from outside. TEST for DL across time points T1, T2, T3, and T4 suggests that R_0 is decreasing. But to understand whether DL or the other states are improving from the lockdown or not, we need to note Table 3. In Table 3, we find for DL if the susceptible population is below minimum CCS value among all time points, both lower and upper limit of TEST decreases over time. This suggests DL is improving in the sense that the number of lockdown period is decreasing over time. If we observe that TEST is increasing over time, it would suggest apart from the fact that lockdown should be increased in those states, the level of infection is increasing.

Using the demographic data for India (`https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`), we find the overall CCS and TEST for India at time points T1, T2, T3, and T4 (Table 4). To compare the infection status of India, we obtained TEST at all time points with minimum CCS. We observe that if the susceptible population is below minimum CCS value (which is 380) among T1, T2, T3, and T4, TEST is almost $19-24$ weeks at all time points. So, it suggests India might have to wait at least another 5-6 months for the pandemic curve to flatten while maintaining maximum possible social distancing norms and in some situations complete lockdown, in absence of specific treatment.

4. Discussion

As things stand at present, the number of COVID-19 cases from many developed countries have surpassed those of China, from where this infection had originated. In such a dire situation, it is very difficult to propose any quintessential lockdown period specific to any country or state. The whole world is struggling to obtain unbiased data to predict on the pandemic. At the same time many questions arise in our minds as "Will the implemented number of lockdown days eradicate the virus?" or "Will it come soon again after the lockdown is over?" or like "How long should the lockdown be continued for the pandemic curve to flatten?" In this scenario of utter dilemma, with the available world-wide data, we provide state-wise estimates of the ideal lockdown phases using our proposed mathematical model for the Indian states. To the best of our knowledge, any guideline for country-wise mathematical prediction of lockdown days is not available till date. So, as the famous British statistician George E. P. Box pointed out, "All models are wrong but some are useful", we only hope that our deductions will provide some helpful suggestions to the policy-makers and public health practitioners, while we are all affected in the pandemic to varying degrees.

Our work suggests that if people are quarantined in limited groups presented by state-

specific CCS, the disease might become contained after the corresponding expected TTE, unless the disease is re-introduced from outside. We observe that although TTE in this case is a very long time (in absence of pharmaceutical interventions), the infection would subside almost completely after TEST. WHO also predicts that COVID-19 virus might continue to stay among us like the HIV. This fact matches with our observation. The TEST that we observed for Indian states or India as a whole shows a trajectory similar to Spanish flu virus. Markel et al. (2007) observed that during the 1918 Spanish flu, the overall deaths in the US cities roughly occurs over a period of 24 weeks. COVID-19 appears to have roughly a similar timeline as that of the Spanish flu. We also observe that as the contact rate (β) of infected with the susceptible increases, R_0 increases steeply. This would mean that if the lockdown is withdrawn before the infection level becomes substantially low, a second wave of infection may hit the society. During the Spanish flu, the cities that had terminated lockdown before the infection was substantially contained, witnessed another abruptly increasing death peak after a short while. In this direction, our work suggests that a lockdown should be taken very seriously to fight against COVID-19 pandemic. This paper provides evidence of the fact that even after the lockdown phase, the disease may recur but it is not expected to create a comparable pandemic situation. Presently, where contact tracing of the infected individuals can lead to tracing down of the exposed individuals, we suggest their quarantining in different feasible groups of sizes not exceeding the state-specific CCS, so that after TEST (as specified for each state) the disease may subside substantially unless any infection is re-introduced from outside. We understand that there could be some flexibility in the lockdown implementation strategies owing to the mathematical approximations made in our calculation of CCS and TEST of the disease. We note here that our model is robust to these approximations.

Another undeniable consequence of the current pandemic is the great negative impact on the economy. This is further magnified by the near paralysed state of transactions in many sectors due to lockdown. Moreover, there is a fear among the general population that if any infection recurs, it might lead to another round of spread of the disease. This fear, which is not unrealistic, may extend the lockdown further. However, after the scheduled lockdown period, if any individual gets infected and a few others get exposed to that person, we need to check whether the total number of such individuals is less than the CCS. If so, the group needs to be quarantined in smaller feasible groups to protect the rest of the population. A newly exposed group, if larger than the CCS, may be quarantined in separate, local subgroups of size no larger than CCS. Moreover, such localised lockdown or quarantine should help in preserving somewhat the daily flow of life and livelihood, and might thereby prevent, at least to some extent, the economy from being further weakened.

However, lockdown in its truest sense may not be feasible in a vast and diverse country like India. Therefore, a strategy of localised and limited lockdowns of objectively identified selected high risk population might be a cost-effective option compared to a generalised "blanket" lockdown. This would imply comprehensive screening for cases and thorough tracing of contacts. So, our take-home message during the still unfolding COVID-19 pandemic is that, till the end of TEST, we must be vigilant and careful. With any further onset of COVID-19 cases in the future, we should follow the quarantine guideline as objectively and

humanely as possible. Like any other epidemic, COVID-19 has the tendency to recur but it might not create any alarming pandemic in the future provided we keep a vigilant eye on our hygiene and have vaccinations and/or treatments. Surely, the realities on the ground − involving human life and death − are much more complex than any model can possibly ever capture. We humbly present the findings of our model as possible instruments of guidance in order to supplement relevant public policies based on ethics and ground realities.

Acknowledgments

SD and IM acknowledges the support by the NIH Fogarty International Center Award #R25 TW009717.

References

Bartlett, M.S. (1957). Measles periodicity and community size. *Journal of the Royal Statistical Society Series A (General).* **120(1)**, 48-70.

Bartlett, M.S. (1960). The critical community size for measles in the United States. *Journal of the Royal Statistical Society: Series A (General).* **123(1)**, 37-44.

Markel, H., Lipman, H.B., Navarro, J.A., Sloan, A., Michalsen, J.R., Stern, A.M., et al. (2007). Non-pharmaceutical interventions implemented by US cities during the 1918-1919 influenza pandemic. *JAMA.* **298(6)**, 644-654.

Nåsell, I. (2005). A new look at the critical community size for childhood infections. *Theoretical population biology.* **67(3)**, 203-216.

Andersson, H. and Britton, T. (2000). Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. *Journal of mathematical biology.* **41(6)**, 559-580.

Dietz, K. (1975). Transmission and control of arbovirus diseases. *Epidemiology.* 104-121.

Birnbaum, A. and Dudman, J., (1963). Logistic order statistics. *The Annals of Mathematical Statistics*, pp.658-663.

Haley, D.C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error *No. TR15, STANFORD UNIV CA APPLIED MATHEMATICS AND STATISTICS LABS.*

Williams, D., Liao, X., Xue, Y. and Carin, L. (2005, August). Incomplete-data classification using logistic regression. *InProceedings of the 22nd International Conference on Machine learning*, pp. 972-979.

Raab, D.H. and Green, E.H. (1961). A cosine approximation to the normal distribution. *Psychometrika* **26(4)**, pp.447-450.

Senapati, A., Rana, S., Das, T. and Chattopadhyay, J. (2020). Impact of intervention on the spread of COVID-19 in India: A model based study. *arxiv.org/abs/2004.04950.*

APPENDIX

Table 5 gives the values of δ and d for four time points considered in our calculation. 'NA' indicates that these parameters cannot be calculated based on the available data. As for example, if there is no death during the period under consideration, the value of d would be zero, as in case of Orissa. For such cases we did not provide CCS and TEST.

Table 5: δ and d values for different time pints

	T4		T3		T2		T1	
State	δ	d	δ	d	δ	d	δ	d
DL	0.212	0.005	0.505	0.004	0.594	0.014	0.700	0.016
GJ	0.422	0.043	0.464	0.061	0.59	0.037	0.395	0.036
JK	0.508	0.012	0.476	0.003	0.5	0.006	0.554	0.024
KA	0.44	0.02	0.552	0.032	0.376	0.019	0.427	0.014
MP	0.264	0.042	0.587	0.038	0.643	0.028	0.430	0.019
MH	0.336	0.035	0.302	0.032	0.378	0.028	0.488	0.025
RJ	0.443	0.02	0.664	0.028	0.515	0.018	0.569	0.013
TN	0.528	0.008	0.34	0.009	0.656	0.008	1.085	0.004
TG	0.348	0.006	0.503	0.003	0.766	0.033	0.553	0.029
UP	0.51	0.015	0.635	0.016	0.585	0.017	0.725	0.029
WB	0.253	0.051	0.487	0.224	0.896	0.059	0.389	0.027
PB	0.253	0.02	0.294	0.028	0.261	0.003	NA	NA
HR	0.392	0.01	0.25	0.024	NA	NA	0.682	0.006
BR	NA	NA	NA	NA	0.532	0.005	0.418	0.003
CH	NA	NA	0.333	0.019	0.302	0.019	NA	NA
OR	NA	NA	0.512	0.012	NA	NA	NA	NA
UT	NA	NA	0.25	0.05	NA	NA	NA	NA
AP	0.468	0.008	0.866	0.008	NA	NA	NA	NA
KL	0.615	0.011	NA	NA	NA	NA	NA	NA
AS	NA	NA	NA	NA	NA	NA	0.833	0.045
HP	NA	NA	NA	NA	NA	NA	3.333	0.029

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 197-208

Transition from Social Vulnerability to Resiliency vis-à-vis COVID-19

S. Pyne[1,2,3]*, S. Ray[4]*, R. Gurewitsch[1] and M. Aruru[3,5]
[1]*Public Health Dynamics Laboratory, Graduate School of Public Health, University of Pittsburgh, USA*
[2]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, USA.*
[3]*Health Analytics Network, PA, USA.*
[4]*Centrum Wiskunde & Informatica, 1098 XG Amsterdam, The Netherlands.*
[5]*Program Evaluation and Research Unit, University of Pittsburgh School of Pharmacy, USA.*

Received: 29 May 2020; Revised: 10 June 2020; Accepted: 11 June 2020

Abstract

The COVID-19 pandemic has revealed systemic deficiencies in preparing and planning for disasters, with profound health, economic, social, political, and humanitarian consequences. When preparing for pandemics, social vulnerability needs to be assessed using vulnerability indices to identify which populations are at greater risk. In this context, we examined the possible association of social vulnerabilities in U.S. cities with COVID-19 case fatality ratios. Post-pandemic return to normalcy is fraught with uncertainty over the ability of different communities to recover with varying degrees of resilience. Towards this, we recommend use of a community resiliency planning framework, along with modeling and evaluation of the required measures, which may be useful for the Indian scenario.

Key words: Social vulnerability; Case fatality ratio; Community resilience; COVID-19; Pandemic; Socio-economic covariates.

1. Introduction

The COVID-19 pandemic, which had its first reported case in Wuhan, China on 17 November 2019, has subsequently had profound global consequences on health, economic, social, political, and almost every major aspect of human life. It is unlike any other single phenomenon that has occurred in modern history since the end of World War II. The effects of the COVID-19 pandemic have spanned over a range that is so vast over space, and yet so condensed over time, that the dual blows of intensity and rapidity have exposed myriad systemic vulnerabilities in many societies around the world. Many countries with apparently robust systems have come under severe stress, and now expect to trudge a slow and painful path to recovery.

Evidently, such systemic deficiencies serve as a reminder of the complex interplay of anthropogenic factors that unfold daily in the form of a vast range of human activities that shape

* These authors contributed equally to this paper.
Corresponding Author: Meghana Aruru
Email: meghana.aruru@pitt.edu

the world around us. By 2050, the average urbanization rate is expected to reach 86% in developed countries, and 64% in developing countries (Liu *et al.*, 2019). At the 2002 biennial meeting of the International Society for Ecosystem Health at Washington DC, a report titled, "Unhealthy Landscapes: How Land Use Changes Affect Health" was published (Patz *et al*., 2004). We understand now how rapid and extensive land-use changes may activate cascades of risk factors involving deforestation, pollution, poverty, migration, and an alarming rise in new human-animal interfaces, which exacerbate the risk of emergence of novel communicable diseases, especially through zoonotic pathogens such as bat-borne coronaviruses (Pyne *et al.*, 2015).

A century after the 1918 "Spanish flu" pandemic, at a 2018 meeting in Geneva, the World Health Organization (WHO) warned us about the possibility of a zoonotic pandemic caused by a novel pathogen, which was enigmatically called, "Disease X" (WHO 2018 Annual review of diseases). In fact, Disease X was included in its "2018 list of diseases to be prioritized under the R&D Blueprint". Yet, as the COVID-19 crisis has clearly demonstrated, we have no choice but to identify, assess and address the systemic vulnerabilities not just at the level of select organizations, but indeed of entire societies.

2. Social Vulnerability: The Place to Begin

There is a lack of consensus on the definition of vulnerability in scientific literature. In the context of a pandemic and other disease outbreaks, we assume that vulnerability is a property of a system, which upon interaction with a given hazard produces an outcome, including a disaster. A stress to the system that has a high potential to harm people and places is termed as a hazard. A disaster refers to a singular large-scale event to which a local community finds it difficult to effectively adapt or cope with. Risk is defined as the likelihood that certain loss or damage could result from a disaster (National Research Council, 2006).

In recent years, the field of vulnerability assessment has shifted from qualitative conceptualization to precise quantitative measures of vulnerability (Cutter SL, *et al*., 2009). Index based measurement provides objectivity to analysis and allows assessment by integrating various indicators to represent different vulnerability scenarios. Known examples of vulnerability indices include the Environmental Sustainability Index (Esty *et al.*, 2005), and the Human Development Index (Burd-Sharps *et al*., 2008). Vulnerability assessments need not aim for quantification of any absolute level of potential damage but rather attempt to assess objectively which populations, and the corresponding systems, are more vulnerable to a particular hazard.

While different frameworks of vulnerability assessment appear in literature, here we extend an earlier classification (Karmakar *et al*., 2010) of human vulnerability with respect to pandemics to include the following types: (1) individual (age, education, nutrition, immune health, comorbidities, exposures, behavioral factors), (2) social (housing, household composition, minority status, community network), (3) economic (income and employment, health insurance, food security, ready government programs, monetary relief instruments designed for a lockdown), (4) infrastructure (regional level secure essential supply chains, energy and communication access, means of essential transport during a crisis, reserved medical stocks), (5) technological (platforms to monitor physical, cognitive and psychological well-being, dynamic information on available medical facilities, optimized diagnostic and vaccination strategies, protected healthcare personnel,

interactive apps and round-the-clock helplines, real-time data collection and visualization, digital connectivity), and (6) administrative (see below). The above list is, of course, not exhaustive.

In the context of a pandemic, useful administrative modes of action may consist of an assiduously data-driven apolitical style of leadership, dynamic and flexible decision-making, mandatory daily clear and accurate media updates on unfolding situations, active ongoing surveillance on the ground for both stationary and mobile populations, meticulous contact tracing with ethical protocols, recognizing sensitivity to key local needs, using empathy as a core criterion when dealing with minorities and vulnerable groups, taking swift steps to mitigate rumors and misinformation, meaningful engagement of communities to obtain regular feedback and respond accordingly and accountably, enact schemes of local and limited economic activity as equitably and cautiously as possible, and coordinate across a well-practiced disaster management plan.

To aid planning, comprehensive social vulnerability maps have been developed in many countries in North America and Europe, and also China. Recent studies have produced vulnerability indices for health risk (NITI Aayog, 2019) and hydro-climatic risk in India (Vittal H, *et al*, 2020). However, to the best of our knowledge, India lacks a comprehensive health risk atlas based on district-wise vulnerability indices that can lay out the key socioeconomic and environmental determinants of community-specific health. The recently published Vulnerability Atlas of India, Third Edition, 2019 includes hazard scenarios for natural disasters, should also be extended to address future epidemics or pandemics (Vulnerability Atlas of India, 2019). Such lacuna could undermine the capacity of an administration to confront a sudden pandemic situation as it might render any breakdown of its systemic responses unpredictable, and thus, result in confounding of priorities.

On quantitative assessment of vulnerability, we take the example of Social Vulnerability Index (SVI) developed by the Geospatial Research, Analysis, and Services Program (GRASP) within the United States Centers for Disease Control and Prevention (CDC) to help flag areas where residents will be in greatest need of support and recovery assistance in the case of a disaster or extreme weather event (CDC's Social Vulnerability Index). SVI provides four categories of vulnerability: socioeconomic status, household composition and disability, minority status and language, and housing and transportation based on data from the 2012-2016 American Community Survey. These four SVI indices along with an overall SVI score are available for different geographical units (e.g., all U.S. counties) at a national scale (CDC's Social Vulnerability Index).

3. Does Social Vulnerability Impact COVID-19 Fatality?

Since the first reported case of COVID-19 in the U.S. in Washington State on January 31, 2020, there have been in the U.S. over 1.7 million cases as of May 27, 2020, when the number of related deaths crossed the mark of 100,000 (Coronavirus in the U.S., New York Times, 2020). In this study, we take a look at the early stages of the pandemic, from 29 February to 15 March, in the COVID-19 affected U.S. cities for possible association between their socioeconomic vulnerabilities and their case fatality ratio (CFR), which is given by the number of deaths by the disease divided by the number of confirmed cases of the same.

It is generally agreed that CFR of COVID-19 has varied between 4.5% and 16% globally, with the U.S. experiencing an overall 6% CFR. However, some U.S. cities have experienced a disproportionate number of deaths compared to others. Since precise calculation of CFR can be made only after an outbreak is over, which is not yet the case, we computed a crude version of CFR as the ratio of the cumulative number of deaths to the cumulative number of cases at a given city on a given date. This dynamic CFR value over a time-period starting from 29 February up to 15 March 2020, for 110 U.S. cities that had at least 500 cases of COVID-19 by that end date, are shown as a heatmap in Figure 1 (Annexure).

To identify the different common temporal patterns of CFR in an unsupervised manner, we used agglomerative hierarchical clustering with linkage by Ward's distance. It revealed 4 clusters of cities having (1) early and prolonged, (2) intermediate, (3) mild, and (4) weak CFR profiles. The names of the cities in the clusters (1) through (4) are depicted in Figure 1 in brown, red, orange, and yellow, respectively. Given the generally weak profiles in cluster 4, we exclude it from further analysis.

CFR is probably better suited than either absolute mortality figures or the COVID-19-specific mortality rate to provide insights into systemic deficiencies that may affect a community's response to the health and other challenges presented by the outbreak it faces. Upon grouping the cities according to the clustered CFR profiles, we compared the social vulnerability indices of these groups. We used SVI for Socioeconomic Status (SES), SVI for households, and overall SVI as computed by the U.S. CDC. Further, we also included some basic social and economic indicators from the latest U.S. Census Bureau data such as the percentages of black population (considered a minority group) and poor population of a city, and its Gini index as a known measure of overall wealth inequality. Figure 2 (Annexure) shows the boxplots for each of these indicators for the cities belonging to clusters 1 (brown), 2 (red) and 3 (orange).

We note that the 3 clusters, as well as their respective medians, differ significantly for each of these variables as per Kruskal-Wallis 3-group test (p-value < 0.1). Notably, cluster 1 with its early and prolonged CFR profile has higher median social vulnerability values compared to the other two clusters, on each of the stated indicators. While we want to avoid making any ecological fallacy in drawing inferences about individual disease outcomes based on city level socioeconomic conditions, it is nonetheless difficult to ignore the common pattern – of higher median vulnerability in cluster 1 – across the various indicators shown in Figure 2.

It is possible that pre-existing or chronic socioeconomic vulnerabilities could directly or indirectly contribute to the increased health risk in many of these cities when faced with the additional burden of a sudden and severe pandemic. The underlying pathways starting from one's exposure to death are often diverse, e.g., many of the young black casualties had little choice but to go out to work on jobs that could not be done remotely from the safety of home. According to the U.S. Bureau of Labor Statistics 2017-2018 report on job flexibilities, while more than 60% of the top quarter of salaried employees could work from home, that figure is less than 10% of those in the bottom quarter (Economic News Release, 2019). Intense research to shed light on this complex topic will no doubt be conducted over the coming years.

4. Developing Community Resiliency to Pandemics

Post-pandemic return to normalcy is fraught with uncertainty over the ability of different communities to recover with varying degrees of resilience. Above, we discussed social vulnerability in the context of populations to determine who would be more impacted by a pandemic than others. Resiliency is a term related to vulnerability. While vulnerability focuses more on chronic stressors such as existing exposures and sensitivities, resiliency, by contrast, is a dynamic property of a population that involves transformative concepts such as learning, critical reflection, adaptation, and reorganization (Cimellaro *et al*., 2016). Rather than assessing the state of a system prior to a disaster, action-oriented questions such as how long it would take to respond, organize, incorporate the lessons learned, and resume normal activities, are asked to assess resiliency of a community.

To illustrate how to approach a community resilience planning process, we could take the example of the U.S. National Institute of Standards and Technology (NIST) Community Resilience Planning Guide (NIST Special Publication, 2016). It defines resilience as "the ability of a community to prepare for anticipated hazards, adapt to changing conditions, withstand, and recover rapidly from disruptions". Often, such ability relies on key components such as infrastructure, utilities, administration, and governance – each of which requires significant time and resources to re-build. Towards this, the NIST guide offers a template of community resilience measurement framework based on estimates of expected recovery times, especially for different communities and infrastructure sectors, which could now be adapted for pandemic resiliency. The expression of resiliency in terms of recovery of system functionality over time following disruption by a disaster event can be seen in the concept diagram [Figure 3 (Annexure)] adapted from NIST Community Resilience Planning Guide, 2016.

Community resiliency planning for a pandemic would require a population to adapt to the post-pandemic realities on the ground, allow backup measures and redundancies in the system, even at the cost of some efficiency, to halt cascades of avoidable losses and despair, restore supply chains for food and energy security, include built-in safety nets such as insurance plans, easy access to loans, medical reserves to limit avoidable losses of life and livelihood, activating new projects to generate economic vitality, resist various sources of rumors and misinformation, and support socializing activities as well as a variety of community-specific and locally relevant constructive measures.

While the technical experts and policy makers may want to develop such resiliency measures by proposing interventions, it is, however, challenging to conduct real-life testing and benchmarking of their impact, particularly among high-density urban populations. In this regard, agent-based modeling (ABM) offers a promising solution based on a computational simulation approach. (Willensky and Rand, 2015) ABM is modeled as a collection of autonomous, decision-making, and interacting entities called agents. An agent could represent an individual, an organization or, for that matter, any entity that can follow certain rules of behavior, and thus, interact with other agents and also the environment. As a result, we can observe macroscopic systemic behaviors – resulting from a large number of micro-level interactions among the agents – as bottom-up "emergent" properties. In an ABM, the stochastic behavior of each agent introduces a certain degree of randomness, which is compensated by conducting a large number of

simulations and aggregating the system responses at the end. By altering an input intervention for a fixed population that is subject to a fixed environment and disease conditions, and running over a given time-period, ABM can help in evaluating the impacts of different interventions.

Despite its ability to allow uncertainty in the model, the micro-level design of an ABM makes it difficult to include in model specification the large degree of detail required to accurately reproduce real-world phenomena. However, an investigation of underlying principles and basic mechanisms is still quite possible, and indeed, most valuable (Brudermann *et al*., 2016). For example, measles outbreaks were modeled with FRED, an ABM platform, under different rates of vaccine coverage vis-à-vis anti-vaccination stigma among selected communities (Sinclair *et al.*, 2019). An ABM was combined with the application of optimal control theories in order to assess resilience of complex systems during extreme events (Cimellaro *et al.*, 2016). A post-epidemic (Ebola) society was modeled with ABM to show how the original structure of the social network, severity of the disaster, and individual beliefs may affect the resilience of the community (Michel S, et al 2015). The emBRACE project used interdisciplinary, socially inclusive, and collaborative methods to develop an ABM based resiliency framework for Europe (Deeming *et al.*, 2019). Another ABM study observed that relationship among the individuals of a community is so vital that a community with less population and more empathy may be more resilient to a disaster than one with more population and less empathy (Valinejad *et al.*, 2020).

Land use and land cover change often exhibit specific community dynamics, which have been modeled by several ABMs. (Guzy *et al.*, 2008; Robinson, *et al.* 2007; Schwarz *et al.*, 2012). This includes modeling of a cooperative approach to mitigate severe risk trade-offs resulting from increase in forest land at the cost of agricultural land (Guzy *et al.*, 2008). The scenario closely resembles real-life in which trade-offs are negotiated between competing risks. Cooperation setting and establishing common grounds demonstrated better outcomes in the model. Such strategies need to be modeled to compare the faced risks and benefits during pandemics to determine policies that ultimately build resilience among the affected communities.

Finally, we arrive at the problem of how to calibrate a model with community-specific characteristics. This is important as conditions prior to a disaster determine the degree of damage and lost functionality, which, in turn, impact resiliency of a community to withstand and recover. Therefore, when assessing resiliency, an ABM should be calibrated with pre-disaster conditions with community-specific real or estimated data and vulnerability indicators. For instance, a model for earthquake evacuation of pedestrians was based on the behavioral rules of the agents derived from real earthquake evacuations (Bernardini *et al.*, 2014). Since local level estimates are not often available, small-area estimates may be used to quantify community-specific health outcomes (Das *et al.*, 2019; Kong *et al.*, 2020). For instance, such estimates for 500 U.S. cities were computed using 27 chronic health and behavioral risk factors (COVID-19 Pandemic Vulnerability Index, NIEHS of NIH). In India, data from national scale surveys such as the National Family Health Survey, Annual Health Survey, Comprehensive National Nutrition Survey, etc., (Dandona *et al.*, 2016) may be harnessed to compute suitable small area estimates for calibrating reliable models of community resiliency.

We believe that the full potential of ABMs for modeling resiliency to disasters is yet to be realized. ABMs could be used for modeling complex administrative cascades, including obstacles,

trade-offs, dogmas, etc. Human emotions such as stigma or empathy can provide us key insights in testing of resilience. Examples of different ABM models were demonstrated in India at 2016 and 2018 'Health Analytics and Disease Modeling' workshops conducted by Health Analytics Network, and the Public Health Dynamics Laboratory of University of Pittsburgh (International Symposium on Health Analytics & Disease Modeling, 2016, 2018; Raghav and Verma, 2018).

5. The Post-Pandemic Way Forward for India

Incidentally, when a super cyclone named Amphan hit parts of the eastern coast of India and Bangladesh on 20 May 2020, right in the midst of the pandemic, despite the significant damage to local infrastructure, a relatively small number of human lives were lost thanks to administrative preparedness and efficient action (Cyclone Amphan bears down on India and Bangladesh – New York Times, 2020) that had to balance the competing risks of mass evacuation against the ongoing lockdown. With climate change and various recurring and seasonal disasters, new multi-hazard indices may prove to be useful for assessing possible vulnerability to the emerging reality of multiple concurrent disasters of different types (Locusts, COVID-19, Flooding pose "Triple Threat" in Africa – New York Times, 2020).

As a model of a system that could be designed to perform like "well-oiled machinery" during a massive disaster, let us take the example of the 2004 tsunami in the Indian Ocean. WHO observed that despite the magnitude of the disaster that killed around 18,000 people in India, there was no significant disease outbreak. We can give credit for this to the state of Tamil Nadu (TN), which has, since 1922, legislated for an independent Directorate of Public Health with an administrative authority board and its own budget. Unlike other Indian states, TN keeps the delivery of public health and medical services distinct (Gupta *et al*., 2010). Importantly, it maintains a dedicated cadre of professionals who are trained in different public health activities, allowing TN to conduct annual "anticipatory planning" to prepare for recurring disasters such as floods, endemic diseases, and other public health emergencies (Krishnan and Patnaik, 2020). Thus, even if badly affected by the pandemic, TN is likely to rebound with its resilient system.

Indeed, the central importance of the human component in the design of any critical system, however technologically enhanced, cannot be over-emphasized, especially if such a system is expected to have its "ear on the ground". The Global Public Health Intelligence Network (GPHIN), developed by Health Canada in collaboration with WHO, is a secure Internet-based multilingual early-warning digital tool that continuously searches global media sources to identify information about disease outbreaks and other events of potential international public health concern. Interestingly, more than 60% of the initial outbreak reports in GPHIN come from unofficial informal sources, including non-electronic media, which are then verified by human experts (WHO Epidemic Intelligence).

We conclude with mentioning the "Sendai Framework for Disaster Risk Reduction 2015-2030" (Sendai Framework for Disaster Reduction, 2015), which was adopted at the Third UN World Conference in Sendai, Japan, on 18 March 2015, and is supported by the United Nations Office for Disaster Risk Reduction. This framework aims to reduce disaster risk and losses over the next 15 years based on its 4 priorities: (1) understanding disaster risk, (2) strengthening disaster risk governance to manage disaster risk, (3) investing in disaster risk reduction for resilience, and

(4) enhancing disaster preparedness for effective response and to "Build Back Better" in recovery, rehabilitation and reconstruction. However, in its current form, it does not explicitly address the disaster of a pandemic. In India, the National Health Mission publishes the Indian Public Health Standards that incorporate many disasters but also lack explicit planning for pandemics (Krishnan and Patnaik, 2018). Likewise, the National Disaster Management Authority formed in 2005 addresses most natural and human-made disasters except for pandemics (Krishnan and Patnaik, 2020). Clearly, this is a gap that India needs to fill in its national planning efforts in the wake of COVID-19 pandemic.

It is not commonly known that during the 1918 Spanish flu pandemic, more than half of all deaths worldwide took place in (then British) India – as many as 17 million deaths from the disease (Schoenbaum, 2001). The high risks of zoonotic and other emerging infectious disease outbreaks for this region is well understood (Jones *et al.*, 2008; Allen *et al.*, 2017). Yet, at the same time, India has the advantage of having many strong institutions including its civilian services, research labs, vibrant media, and well-knit communities. We believe that by adopting the formal structure and priorities (such as those of the Sendai framework) to fortify its systems, India can emerge as a global leader in setting response and recovery standards that are specific to pandemic disasters and cognizant of the strengths and vulnerabilities of its unique and diverse communities, and thus, become more resilient to the complex crises of the future.

References

Allen, T., Murray, K. A., Zambrana-Torrelio C., Morse S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J., and Daszal, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Naure Communications*, **8**(**1**), 1-10.

Bernardini, G., D'Orazio, M., Quagliarini, E. and Spalazzi, L. (2014). An agent-based model for earthquake pedestrians' evacuation simulation in Urban scenarios. *Transportation Research Procedia*, **2**, 255-263.

Brudermann, T., Hofer, C. and Yamagata, Y. (2016). Agent-based modelling - A tool for urban resilience research? In Y. Yamagata and H. Maruyama (Eds.), *Urban Resilience, Advanced Sciences and Technologies for Security Applications*, 135-151, Springer, Switzerland (ISBN: 978-3-319-39810-5).

Burd-Sharps, S., Lewis, K. and Martins, E. B. (2008). *The Measure of America : American Human Development Report*, *2008-2009.* Columbia University Press.

CDC's Social Vulnerability Index. *Centers for Disease Control and Prevention*. https://svi.cdc.gov/.

Cimellaro, G. P., Renschler, C., Reinhorn, A. M. and Arendt, L. (2016). *Urban Resilience for Emergency Response and Recovery. Fundamental Concepts and Applications*. Springer International Publishing, **41** (ISBN: 978-3-319-30656-8).

Coronavirus cases in the U.S.: Latest map and Case Count. (2020). *The New York Times*. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

COVID-19 Pandemic Vulnerability Index. *National Institute of Environmental Health Sciences.* National Institutes of Health. https://covid19pvi.niehs.nih.gov/

Cutter, S. L., Emrich, C. T., Webb, J. J. and Morath, D. (2009). Vulnerability to Climate Variability Hazards: A Review of the Literature. *Final Report to Oxfam America.*1-44.

Cyclone Amphan Bears Down on India and Bangladesh. (2020). *The New York Times*. https://www.nytimes.com/2020/05/19/world/asia/cyclone-amphan-bangladesh-india.html.

Dandona,. R, Pandey, A. and Dandona, L. (2016). A review of national health surveys in India. *Bulletin of the World Health Organization*, **94**, 286-296A.

Das, S., Chandra, H. and Saha, U. R. (2019). District level estimates and mapping of prevalence of diarrhoea among under-five children in Bangladesh by combining survey and census data. *PLoS One*, **14**(**2**), 1-19.

Deeming, H., Fordham, M., Kuhlicke, C., Pedoth, L. and Schneiderbauer, S. C. (2019). *Framing Community Disaster Resilience.* John Wiley & Sons (ISBN: 9781119165996).

Economics News Release.(2019). *Job Flexibilities and Work Schedules Summary*. U.S. Bureau of Labor Statistics. http://bls.gov/news.release/flex2.nr0.htm

Esty, D., Levy, M., Srebotnjak, T. and de Sherbinin A. (2005). Environmental sustainability index: benchmarking national environmental stewardship. *Yale Center for Environmental Law and Policy*. http://sedac.ciesin.columbia.edu/es/esi/ESI2005_Main_Report.pdf.

Gupta, M. D., Desikachari, B. R., Shukla, R., Somanathan, T. V., Padmanaban, P. and Datta, K. K. (2010). How might India's public health systems be strengthened? Lessons from Tamil Nadu. *Economic and Political Weekly*, **45**(**10**), 46-60.

Guzy, M. R., Smith, C. L., Bolte, J. P., Hulse, D. W. and Gregory, S. V. (2008). Policy research using agent-based modeling to assess future impacts of urban expansion into farmlands and forests. *Ecology and Society*, **13**(3), 37.

International Symposium on Health Analytics & Disease Modeling 2016. http://healthanalytics.net/HADM2016/.

International Symposium on Health Analytics & Disease Modeling 2018. http://healthanalytics.net/HADM2018/

Jones, K. E., Patel, N.G., Levy, M. A., et al. (2008). Global trends in emerging infectious diseases. *Nature*, **451**(**7181**), 990-993.

Karmakar, S., Simonovic, S. P., Peck, A. and Black, J. (2010). An Information System for Risk-Vulnerability Assessment to Flood. *Journal of Geographic Information System*. **02**(**03**),129-146.

Kong, A.Y. and Zhang, X. (2020). The Use of Small Area Estimates in Place-Based Health Research. *American Journal of Public Health*. **110**(**6**), 829-832.

Krishnan, S. and Patnaik, I. (2018). *Health and Disaster Risk Management in India.* NIPFP Working Paper Series. https://www.nipfp.org.in/publications/working-papers/1836/

Krishnan, S. and Patnaik, I. (2020). *Health and Disaster Risk Management in India.* In: Ardalan A, Ordun CY, Riley JM. *Public Health and Disasters*. Springer Singapore.

Liu, Y., Li, L., Chen, L., et al. (2019). Urban growth simulation in different scenarios using the SLEUTH model: A case study of Hefei, East China. *PLoS One*. **14**(**11**), e0224998.

Locusts, COVID-19, Flooding Pose "Triple Threat" in Africa. (2020). *The New York Times*. https://www.nytimes.com/aponline/2020/05/21/world/africa/ap-af-africa-locust-outbreak.html.

Michel, S. and Megerdoomian, K. (2015). Modeling Community Resilience for a Post-Epidemic Society Modeling Community Resilience for a Post-Epidemic Society. In: *Computational Social Science Society of the Americas*, Santa Fe, New Mexico. http://computationalsocialscience.org/wp-content/uploads/2015/10/CSSSA_2015_submission_37.pdf.

National Research Council. (2006). *Facing Hazards and Disasters: Understanding Human Dimensions.* The National Academies Press, Washington, DC.

NIST (2016). Special Publication 1190: Community Resilience Planning Guide for Buildings and Infrastructure Systems, Volume II. *National Institute of Standards and Technology*. **II**:126. https://www.nist.gov/topics/community-resilience/planning-guide.

NITI Aayog. (2020). *Healthy States, Progressive India: Report on the Ranks of States and Union Territories.*

Patz, J. A., Daszak, P., Tabor, G. M, et al. (2004). Unhealthy Landscapes : Policy recommendations on land use change and infectious disease emergence. *Environmental Health Perspectives,* **1092**(**10**),1092-1098.

Pyne, S., Lee, S. and McLachlan, G. (2015). Nature and Man: The Goal of Bio-security in the Course of Rapid and Inevitable Human Development. *Journal of Indian Society of Agricultural Statistics*, **69**(**2**), 117-125.

Raghav, P, and Verma, C. (2018). International Symposium in New Delhi Addresses Emerging Challenges in Health Analytics and Disease Modeling. *Science Reporter*, June, pp 49-50.

Robinson, D. T., Brown, D. G., Parker, D. C., et al. (2007). Comparison of empirical methods for building agent-based models in land use science. *Journal of Land Use Science*. **2**(1),31-55.

Schoenbaum, S. C. (2001). The impact of pandemic influenza, with special reference to 1918. *International Congress Series*. **1219** (**C**), 43-51.

Schwarz, N., Kahlenberg, D., Haase, D. and Seppelt, R. (2012). ABMland - A tool for agent-based model development on urban land use change. *Journal of Artifical Societies and Social Simulation*. **15**(**2**), 8.

Sendai framework for disaster risk reduction 2015-2030. (2015). *Australian Journal of Emergency Management*. **30**(**3**), 9-10.

Sinclair, D. R., Grefenstette, J. J., Krauland, M. G., et al. (2019). Forecasted zize of measles outbreaks associated with vaccination exemptions for school children. *JAMA Network Open*, **2**(**8**), e199768.

Valinejad, J., Mili, L., Triantis, K., von Spakovsky, M. and van der Wal C. N. (2020). *Stochastic Multi-Agent-Based Model to Measure Community Resilience-Part 2: Simulation Results.* 1-12. http://arxiv.org/abs/2004.05185.

Vittal, H., Karmakar, S., Ghosh, S. and Murtugudde, R. (2020). A comprehensive India-wide social vulnerability analysis: Highlighting its influence on hydro-climatic risk. *Environmental Research Letters*, **15**(**1**),14005.

Vulnerability Atlas of India. Third Edition, 2019. https://bmtpc.org/topics.aspx?mid=56&Mid1=180

WHO (2018). 2018 Annual review of diseases prioritized under the Research and Development Blueprint Informal consultation. *World Health Organization.* http://www.who.int/csr/research-and-development/documents/prioritizing_diseases_progress/en/.

WHO Epidemic intelligence – systematic event detection. *World Health Organization.* https://www.who.int/csr/alertresponse/epidemicintelligence/en/.

Willensky, U. and Rand, W. (2015). *An Introduction to Agent-Based Modeling. Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. (ISBN: 9780262731898).

ANNEXURE

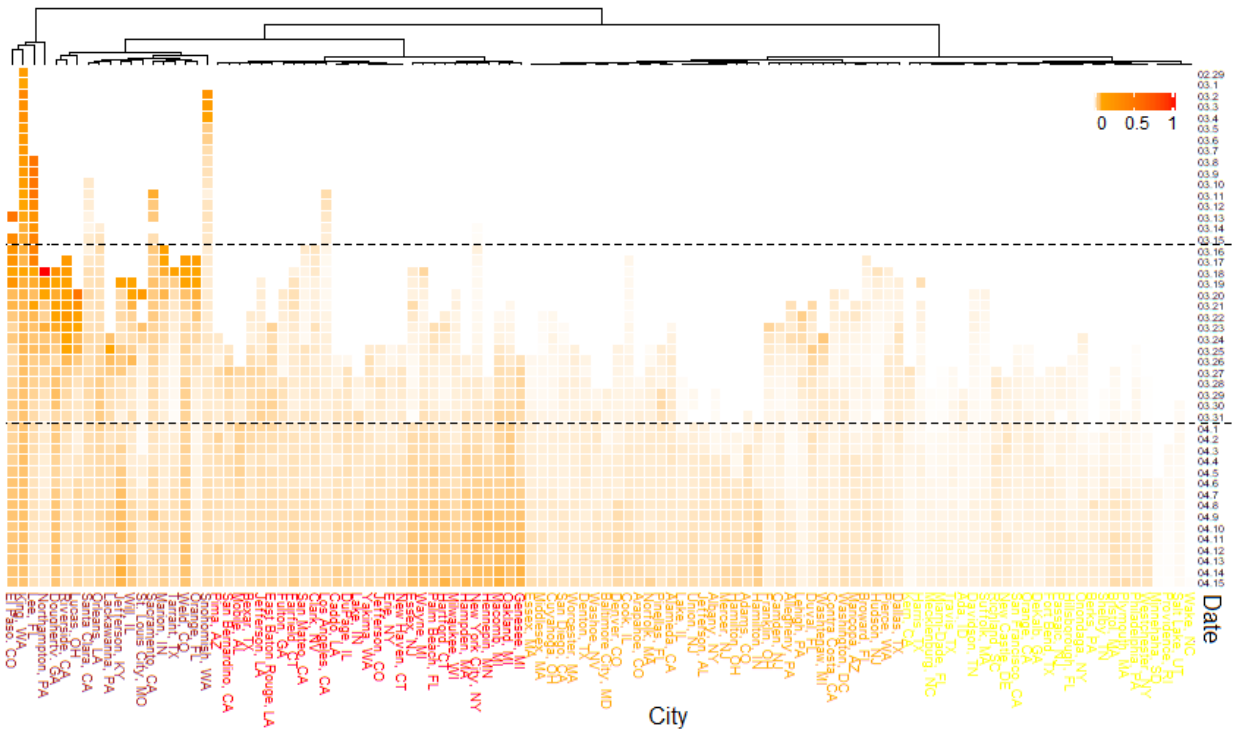


Figure 1: Clusters of COVID-19 affected U.S. cities. Unsupervised hierarchical clustering of dynamic CFR time series, shown in heatmap, of 110 U.S. cities (*x*-axis) revealed 4 clusters of cities, as named in 4 different colors. The dashed lines mark 15-day intervals over the time-period (*y*-axis) of February 29 to April 15, 2020. On top is a dendrogram showing the linkage among the clusters based on Ward's distance.

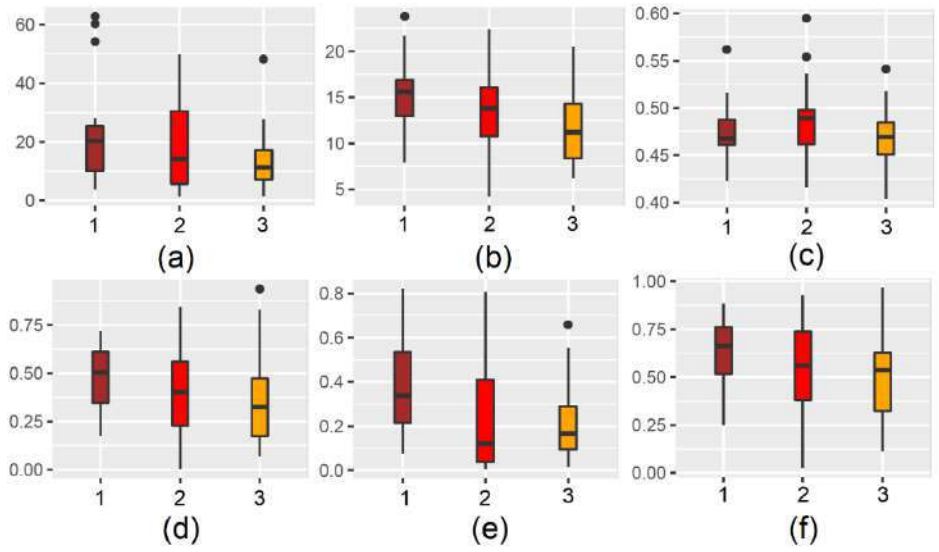


Figure 2: Comparison of socioeconomic covariates. For clusters 1, 2, and 3 of U.S. cities, the boxplots show (a) the percentage of black population, (b) the percentage of poor population, (c) Gini index, (d) SVI for SES, (e) SVI for households, and (f) overall SVI.

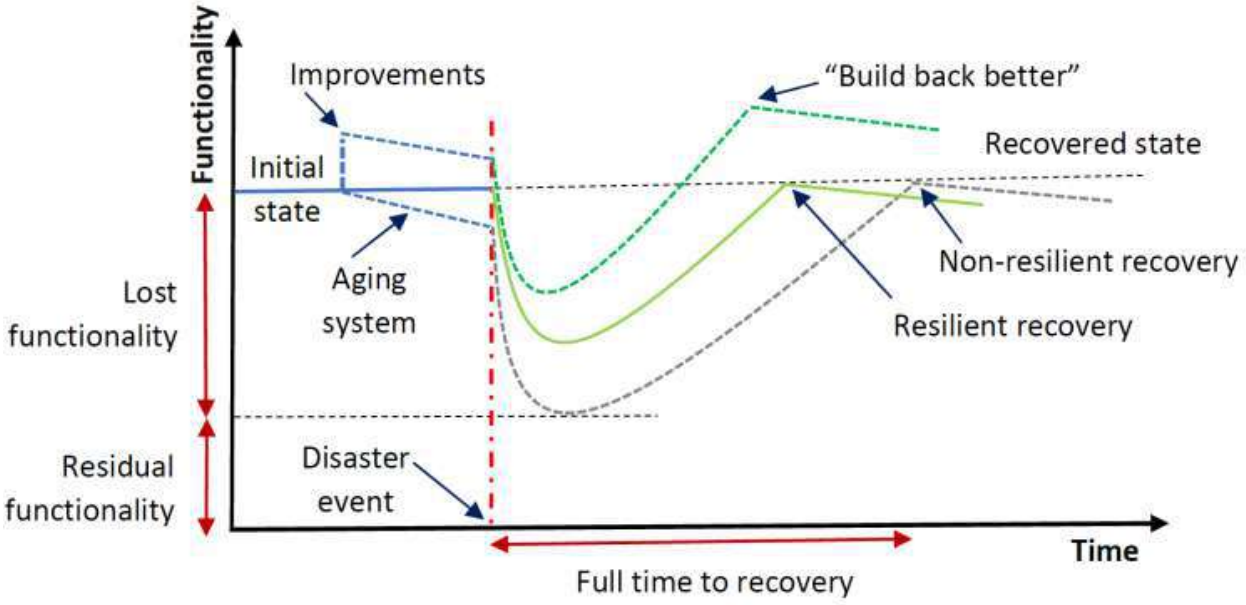


Figure 3: A Functional Concept of Resiliency. Resiliency can be expressed in terms of recovery of system functionality over time following disruption by a disaster event (adapted from NIST Community Resilience Planning Guide, 2016).

Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 1, 2020 (New Series), pp 209-221

What if Lockdown is Removed? District Level Predictions for Maharashtra and Gujarat

Srimanti Dutta[1], Kalyan Das[2], Kashinath Chatterjee[1] and Arindom Chakraborty[1]

[1]*Department of Statistics Visva-Bharati*
[2]*Department of Mathematics IIT Bombay, Powai*

Received: 29 May 2020; Revised: 11 June 2020; Accepted: 12 June 2020

Abstract

Throughout the world everyday, thousands of people are getting infected and hundreds are dying in each country due to an pandemic caused by the outbreak of COVID-19. Starting from Wuhan, this virus has almost travelled each and every country infecting millions of people. In the absence of any vaccine till date and no confirmation of herd immunity (D'Souza and Dowdy (2020)), if any, the world is depending on some non-pharmaceutical interventions (NPIs) to curb the spread of COVID-19. India is no exception. After three consecutive lockdown spells, India has entered into lockdown 4.0 from May 18, 2020, with some relaxations. Lockdown has a huge negative impact on the economy. This study aims to predict the future in few districts of two selected states (Maharashtra and Gujarat) of India if lockdown is removed or partially implemented. A statistical model based on renewal process has been used for prediction. A hierarchical Bayesian method has been used for this data. Predictions have been made till June 30, 2020, for each district of the two states (Maharashtra and Gujarat) under this study.

Key words: Average reproduction number; Renewal process; Hierarchical Bayes; Infections.

AMS Subject Classifications: 62M20, 62F15

1. Introduction

After ages, the world is experiencing something so lethal and frightening. Starting from a fish market in Wuhan, a novel coronavirus named SARS-CoV-2 or COVID-19 has travelled the globe within a short span of time. Over 45 lacs people are already infected and the death toll crossed 3 lacs throughout the world. In comparison with other deadly viruses like Ebola, SARS or MERS, this virus is less fatal but is more contagious in nature (Mahase (2020), Fox (2020)). A recent discovery of a particular mutation, found in India, claimed that this virus has become more contagious (See Korber *et al.* (2020)).

Till date, no vaccine is available for SARS-CoV-2. A lot of clinical trials are going on throughout the world to discover some "effective" vaccines. From previous experience, we have observed that it took years to find any "effective" vaccine. In this situation, the

Correponding Author: Arindom Chakraborty
Email: arindom.chakraborty@visva-bharati.ac.in

world is relying more on non-pharmaceutical interventions (NPIs) to curb the spread of this virus. Closing of educational institutions, ban on travel, maintaining physical distance were taken as NPIs by different countries at different point of time. However, these NPIs were not sufficient enough to tackle the spread. Knowing most of the adverse consequences, many countries were compelled to impose lockdown as the last resort to battle against COVID-19. Unlike most of the European countries, in India, the lockdown was not maintained properly in most of the parts. As a consequence, infection spread throughout the country.

It is always a challenge to epidemiologists to predict the progress of infection caused by an unknown virus. A lot of people are working in this direction individually as well as in groups. Commonly used SER, SEIR, adaptive SEIR are more common among epidemiologists which are basically based on differential equations. As an alternative, statisticians are using some probabilistic models to capture the uncertainty. It is imperative to use Bayesian inference in all models as we have little or no experience in this SARS-CoV-2. The data available from Wuhan gives us some idea of the spread. However, it is to be noted that, fatality or attack rate depend on several factors like genetic profile, particular mutation of the virus and on many confounding factors. It is well known that, even if in Bayesian inference, we borrow prior information from some other virus. As an example for SARS-CoV-2, priors are chosen on the basis of data on SARS infection which happened a few years ago. This sort of assumptions always put any prediction model under severe threat of reliability. In a recent paper published by Luo (2020) claimed that 97% of active cases would have been solved by May 27, 2020, which received strong criticism.

Specific to the Indian scenario, the first work that got some acclamation was done by a group of biostatisticians and epidemiologists at the University of Michigan (The Covid-19 India Group, 2020). They have used adaptive SEIR model and predictions have been made on the basis of this model. Gompertz model has been considered by Lee, Lei and Mullick (2020). A lot of work has been done by the researchers from Imperial College of London (ICL) (Seth *et al.* (2020), Walker *et al.* (2020)). The effects of non-pharmaceutical interventions have been studied for European countries. Another impressive model has been developed by the researchers belonging to IHME, University of Washington, Seattle (UW) (IHME, 2020). In a recent work done by Chatterjee (2020), it has been shown that the models developed by ICL and UW are so far giving good results in terms of prediction. In this work, we have implemented a model induced from ICL model. A question that people are interested in is that when this pandemic will come to an end. From previous studies, we have some idea about the disease progression of several viruses like SARS or MERS. Researchers are looking forward to the time point where asymptote will be achieved. It may be noted that the Government is trying to implement restrictions, if any, in micro levels to minimize economic loss. For the policymakers, it becomes necessary to have an idea in micro-level. The model that we are using in the present work is quite general and can be implemented at any region provided sufficient data is available.

The paper is structured as follows. In Section 2, we describe the methods briefly. Results of the districts are mentioned in Section 3. It is followed by a discussion section.

2. Methods

To predict the time-varying reproduction number (R_t) under various levels of lockdown, the reproduction number at time t is taken as a scale multiple of the baseline reproduction number. The multiplicity factor is a constant function of the lockdown levels. Note that we insert four dummy variables for the above four lockdown levels. The number of infected cases at any day can be predicted using the reproduction number and the weighted average of previous days' affected figures with the discretized serial interval distribution probability of secondary infection as weights (See Fraser (2007), Cori *et al.* (2013), Nouvellete (2018), Cauchemez *et al.* (2008)). The mortality at any day can similarly be predicted using the case fatality ratio (CFR) and the weighted average of previously affected figures with the chance of mortality as weights. Mathematically speaking, there are three models working together. These are models for infections, deaths and average reproduction rate. Let $I_{s,t}, D_{s,t}$ and $R_{s,t}$ be, at time t, number of new infections, number of new deaths and average reproduction number for the state s. Moreover, let $g(t)$ be the serial interval distribution. As explained above, we then express the infection on t th day as

$$I_{s,t} = R_{s,t} \sum_{\tau=0}^{t-1} I_{s,\tau} g_{t-\tau} \tag{1}$$

We discretize the serial interval as:

$$g_u = \int_{u-.5}^{u+.5} g(\tau) d\tau; u = 2, 3, 4, ... \tag{2}$$

with $g_1 = \int_0^{1.5} g(\tau) d\tau$ where from the past experience, $g(t)$ is assumed to be a gamma distribution with mean 6.5 (average time from onset in a primary infection to onset in a secondary infection) and a relatively small coefficient of variation 0.62 (Seth et al. (2020)).

For mortality, the observed number of deaths may be assumed to follow a Negative Binomial law where the expected deaths are assumed to be the weighted average of the daily infection, weights being a mixture of two gamma distributions that account for the incubation period and time between the onset of symptoms and death. In case of $R_{d,t}$ (the average reproduction number for a state d), we use levels of lockdown as covariates. Let $R_{d,0}$ be the baseline reproduction number for dth district. Then $R_{d,t}$ is modelled as

$$R_{d,t} = AF_d * R_{d,0} * (2 * \phi^{-1}(-\alpha * \delta)), \tag{3}$$

where AF_d is the adjustment factor considered for district d based on its population, δ indicates the level of lockdown: 0 (no lockdown), 1 (partial lockdown) or 2 (complete lockdown) and ϕ^{-1} is the inverse logit or Sigmoid function. It may be noted that partial lockdown is a policy decision which can be quantified. The Government or the local administration may decide the extent of lockdown which may qualify as "partial". We further assume

$$\alpha \sim N(0, 0.5)$$

To model the baseline reproduction number for a district d, we assume

$$R_{d,0} \sim N(3.28, \kappa)$$

where $\kappa \sim N^+(0, 0.5)$, N^+ denotes a half normal distribution with positive support.

To fit the model we use 6 sequential days of an equal number of infections: $I_{1,d} = .. = I_{6,d} \sim Exponential(1/\tau)$, where $\tau \sim Exponential(0.03)$.

MCMC samples are drawn from posterior distributions using Stan software and convergence criteria have been studied in details.

3. Results

Here we apply the proposed methodology for all the districts of Maharashtra and Gujarat. Daily infection and death data have been considered. In this study, for each district, data on 83 time points (from 2nd March, 2020 to 23rd May, 2020) are used. A few districts with no death or very few cases have been excluded from the study.

We have tried to predict the number of daily infections and daily deaths till 30th June, 2020 for each district under different levels of lockdown. Base reproduction rate and reproduction rate on 30th June, 2020 under different levels of lockdown, have also been reported.

In Figure 1 the values of reproduction rate R_t for different districts are given. Figure 1(a) depicts the baseline reproduction number R_0 values for each district belonging to Maharashtra. Aurangabad has the highest R_0 closely followed by Solapur. For Mumbai and Pune, R_0 values are over 3.3, which is also high.

As expected it is found that lockdown for a prolonged period has a positive effect in controlling R_t. After 31st May 2020 even if lockdown is completely removed, R_t values are less than 1.2 (Figure 1(b)) for all the districts of Maharashtra. Our purpose is to see which stage of lockdown pushes the value of R_t to smaller than 1. This may not be achieved, as in Figure 1(d), even if we impose complete lockdown for the entire month of June 2020. For districts like Pune and Mumbai, the situation remains almost identical even if lockdown is extended to its highest level from partial restrictions (Figure 1(c)). It may be noted that for a densely populated district like Mumbai, a reduction of a decimal place in R_t value may result in the reduction of a large number of infections.

In Figures 2(a)-2(l), predictions for four important districts (Mumbai, Nagpur, Pune and Thane) have been considered. Results for other districts are also available. Three different levels of lockdown have been considered for each of these four districts.

In the case of Mumbai, a surge may be observed if lockdown is withdrawn totally. Daily infection may cross 1000 marks within the first week of June and there will be exponential growth (Figure 2(a)). This may put the existing healthcare system into tremendous pressure. On the other hand, complete lockdown may reduce the rate of daily infection but may result in severe economic depression to this financial capital of India. In this tricky situation, the Government may opt for some intermediate solution where partial lockdown may be imposed. From Figure 2(c) it can be seen that the daily infection rate may increase slowly if partial lockdown is imposed. Similar features may be observed for the other three districts.

Pune may observe daily infected counts just over 100 if partial lockdown is maintained (Figure 2(h)). Removing lockdown may create a similar situation like Mumbai, where daily count may exceed 10,000 (Figure 2(g)). For Thane, it may take end of June to reach the daily count to be 10,000 if lockdown is removed completely (Figure 2(j)). On the contrary, complete lockdown may bring the rate of daily infection down to a great extent (Figure 2(l)). In the case of Nagpur, similar sort of inferences may be made (Figures 2(d)-2(f)).

In Figures 3(a)-3(l), we have plotted the predictions for a few districts (Aurangabad, Nashik, Akola and Palghar) which are less affected, compared to the districts we have already considered, till date. Here also it is seen that removal of complete lockdown may bring the less affected districts to highly affected districts. A partial lockdown may help the administration to maintain the situation which may be controlled with existing health care set up.

In terms of fatality, so far, COVID-19 is less fatal compared to MERS or SARS. A few probable confounding factors like the effect of malaria or BCG vaccine or any other factor or combination of factors may have contributed to this cause. For most of the districts, an increasing trend in the daily number of deaths may be seen. Unlike regions of Europe or states of the US, the death toll is relatively very low for the districts of India.

For Mumbai, the removal of lockdown may result in more than 100 daily deaths at the end of June (Figure 4(a)). A complete lockdown may bring the rate down significantly (Figure 4(c)), whereas a partial lockdown may result in less than 10 deaths per day.

Among all states in India, Gujarat has the highest baseline reproduction number for which the number of infected people will be higher. From Figure 5(a), it can be seen that for districts like Ahmedabad and Surat this value is above 3.5. Lockdown for more than two months has done something significant in reducing values to a great extent. Figure 5(c) shows that the expected reproduction number on 30th June will be less than 1 for most of the districts if partial lockdown is imposed. For the two effected districts Ahmedabad and Surat, the R_t may cross the threshold of 1. Further stringent intervention i.e. complete lockdown may further reduce R_t values.

Like Mumbai, Ahmedabad is also severely affected due to high population density. Removing lockdown completely would be very fatal and it is evident from Figure 6(a). A partial lockdown may help in curbing the spread (Figure 6(b)) whereas complete lockdown may contain the spread (Figure 6(c)). However, districts like Ahmedabad and Surat (Figures 6(d)-6(f)) complete lockdown may have a huge negative impact on economy. Some mitigation policies may be adopted. Vadodara, another important district shows a similar pattern (Figures 6(g)-6(i)). Daily death predictions for two districts, Ahmedabad and Surat are given in Figures 7(a)-7(f). If lockdown is removed completely, the daily death toll may cross 10 from 1st June 2020 and then an exponential growth may be anticipated. However, in a partial lockdown situation death rate will increase in a more constant manner. In the case of Surat, a similar trend is found.

It is to be mentioned here that in reality, the death figures may be much higher. We are assuming that the fatality rate is the same for all the age groups. From Wuhan data, it is found that the death rate for individuals above 60 is much higher compared to other age

groups. Moreover, patients with complex diseases like COPD, CVD, diabetes, hypertension have a higher probability of dying.

4. Discussion

In this study, we have considered different levels of lockdown. The proposed model is very general in nature and hence may incorporate several other covariates. For example, the effect of another important NPI *viz.* the role of intensive testing may also contribute to a reduction of R_t values which eventually reduce the number of infections. This study is under investigation and we plan to communicate it soon. While predicting, we have assumed that if there is any non-pharmaceutical intervention (NPI), it needs to be adhered to completely.

It may be noted that, after lockdown 4.0, complete lockdown may not be applicable any more. Relaxing lockdown norms will help the economy but at the same time will increase the chance of infection with increased mobility. We feel that this work may give an idea to the policymakers to chalk out future plan district-wise. As mentioned earlier, this can be done for subdivision, block or even a small region provided sufficient data is at our disposal.

The model we have used is based on a renewal process. A comparison may be of interest to see the change in inferences if models are changed. Another important issue in modelling the spread of any viral disease is the reproduction rate model. A comparison can also be made using some other types of modelling strategies like using Weibull model.

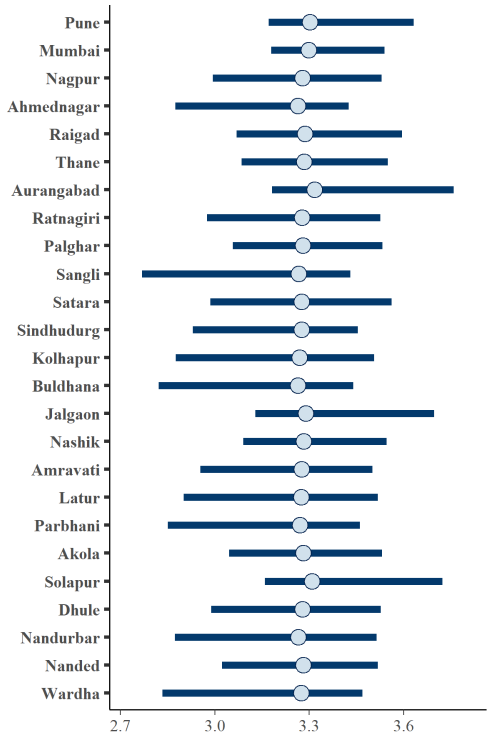
Another interesting and challenging job is to take care of asymptomatic cases which are very crucial in the spread of COVID-19. In recent times it has been observed that due to repeated mutations of this SARS-CoV-2 virus more infected people are becoming asymptomatic. This puts a lot of people at risk as the spreader does not know about his/her infection. More and more tests are desirable in this scenario. Pool testing may be a good option for the regions which are designated as containment zones. Pool testing will decrease the use of test kits and results may be available much sooner.

Getting good quality data is always a big challenge. Out of approximately 150000 individuals around 8% data do not contain district information. Same is true for data related deaths. This may result in biased estimates of the parameters involved and may hamper predictions.

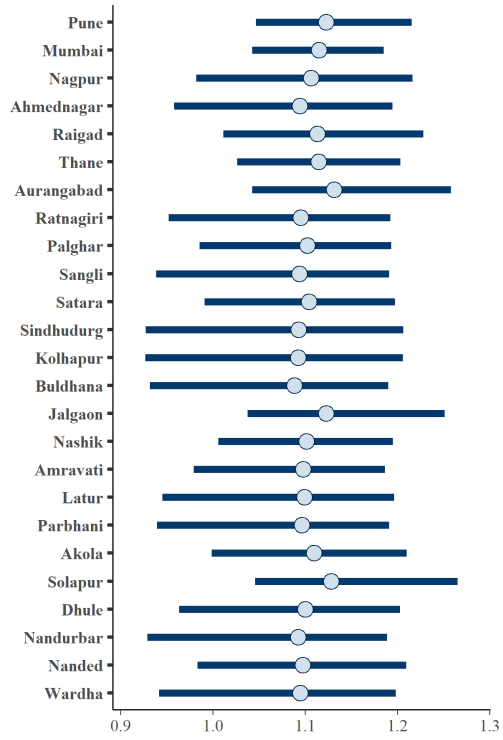
Whatsoever mitigation policies the Government may take, it is essentially the duty of each and every citizen to abide by all the interventions with utmost sincerity. Regulations given by Ministry of Health and Family Welfare or Ministry of Home Affairs or local administrator may be followed religiously. Trace, test and quarantine (TTQ) is the need of the hour.

5. Data and Software

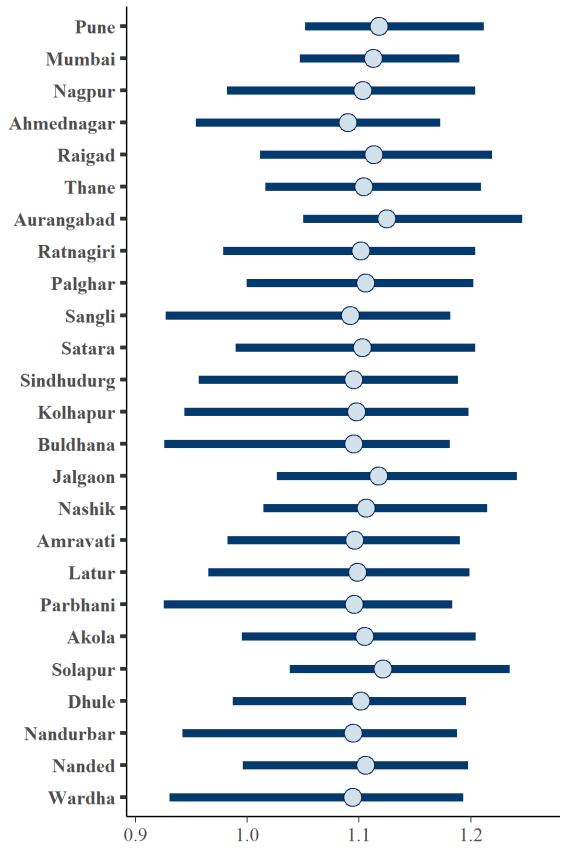
For this work we have used the data available in https://github.com/covid19india/api. All computations have been done using RStudio and Stan.



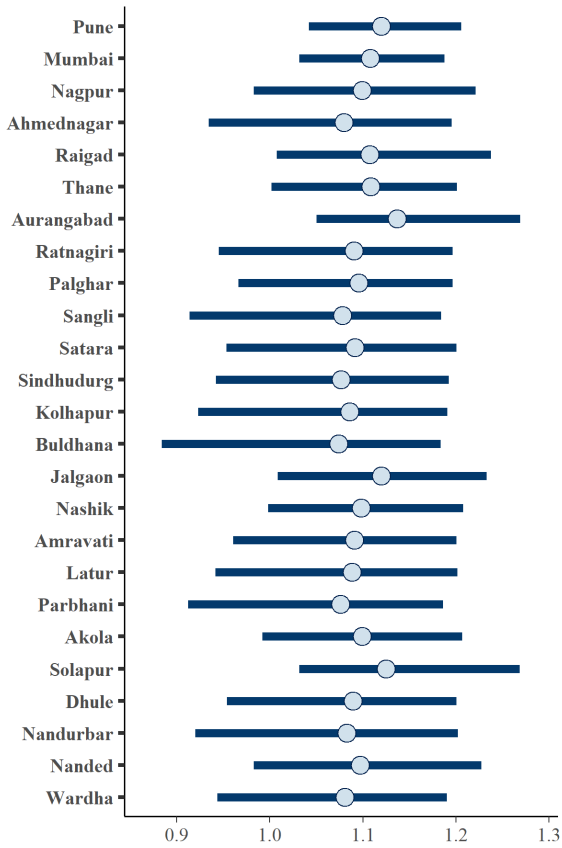
(a) Baseline reproduction number (R_0)



(b) Predicted R_t on 30th June, 2020 under no lockdown

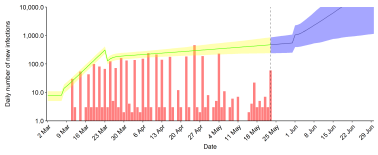


(c) Predicted R_t on 30th June, 2020 under partial lockdown

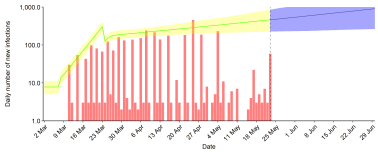


(d) Predicted R_t on 30th June, 2020 under complete lockdown

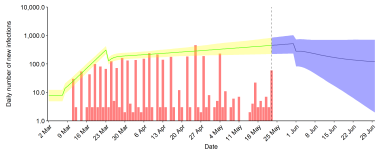
Figure 1: Average reproduction number for Maharashtra



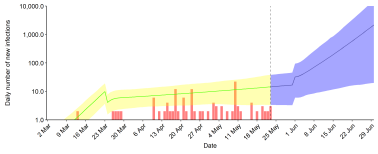
(a) Mumbai: under no further lockdown



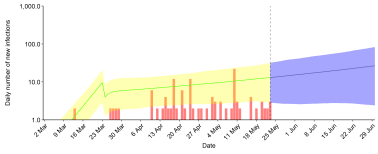
(b) Mumbai: under partial lockdown



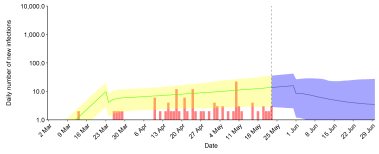
(c) Mumbai: under complete lockdown



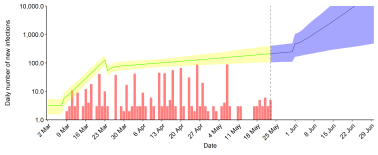
(d) Nagpur:Under no further lockdown



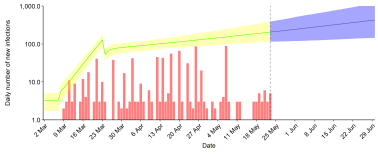
(e) Nagpur:Under partial lockdown



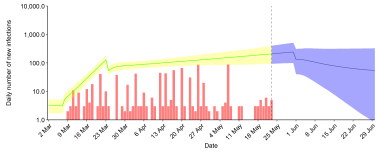
(f) Nagpur:Under complete lockdown



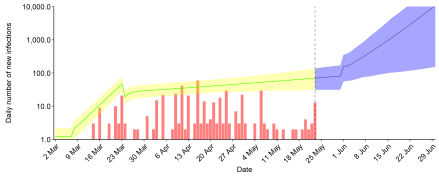
(g) Pune: under no further lockdown



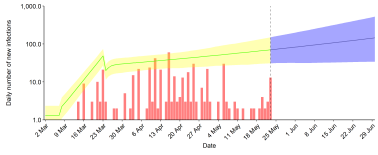
(h) Pune: under partial lockdown



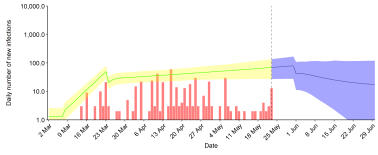
(i) Pune: under complete lockdown



(j) Thane: under no further lockdown

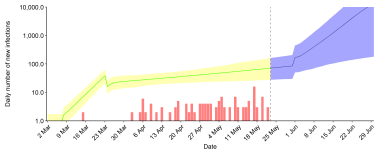


(k) Thane: under partial lockdown

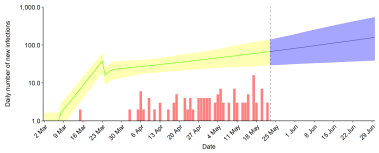


(l) Thane: under partial lockdown

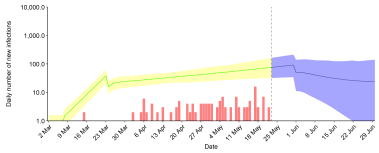
Figure 2: Prediction for different districts of Maharashtra



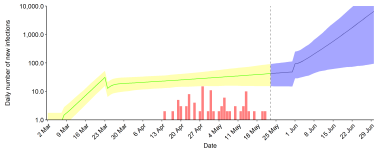
(a) Aurangabad: under no further lockdown



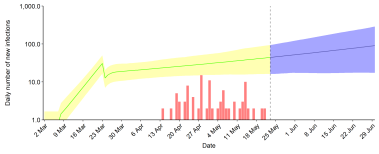
(b) Aurangabad: under partial lockdown



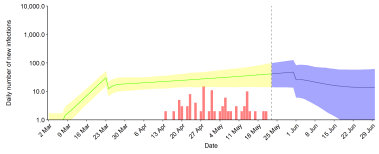
(c) Aurangabad: under complete lockdown



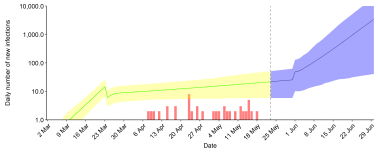
(d) Nashik:Under no further lockdown



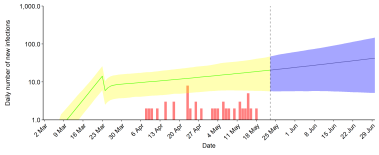
(e) Nashik:Under partial lockdown



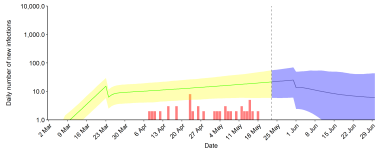
(f) Nashik:Under complete lockdown



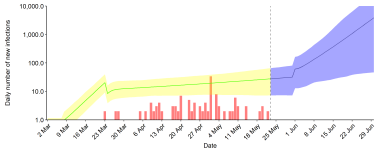
(g) Akola: under no further lockdown



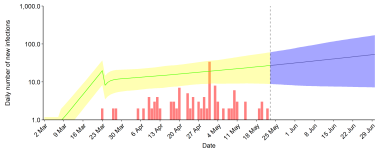
(h) Akola: under partial lockdown



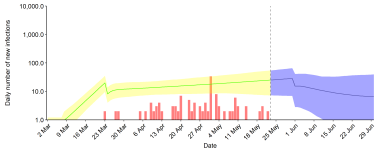
(i) Akola: under complete lockdown



(j) Palghar: under no further lockdown

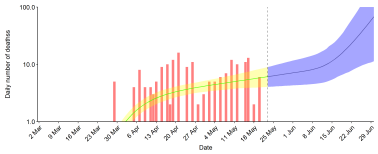


(k) Palghar: under partial lockdown

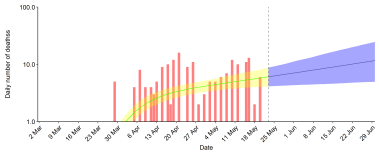


(l) Palghar: under partial lockdown

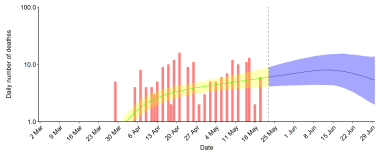
Figure 3: Prediction for different districts of Maharashtra



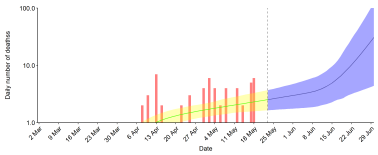
(a) Mumbai: under no further lockdown



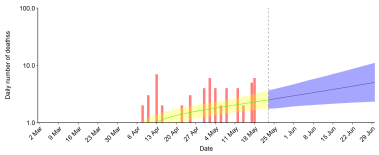
(b) Mumbai: under partial lockdown



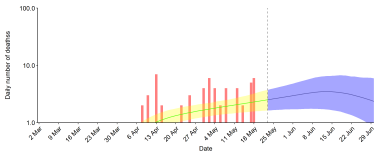
(c) Mumbai: under complete lockdown



(d) Pune: under no further lockdown

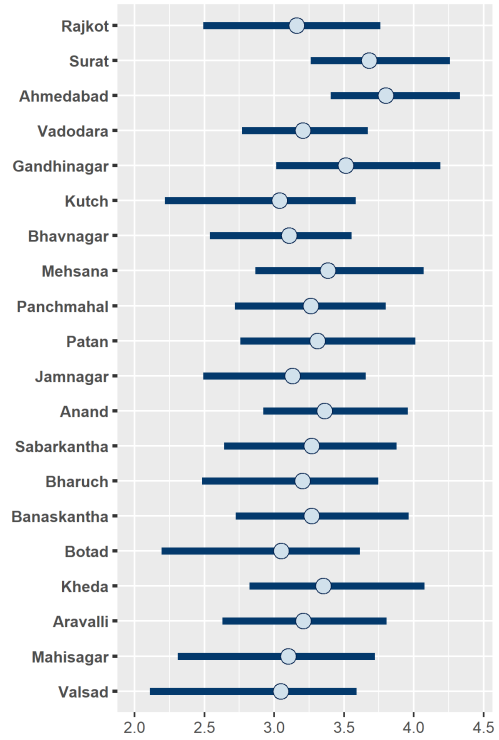


(e) Pune: under partial lockdown

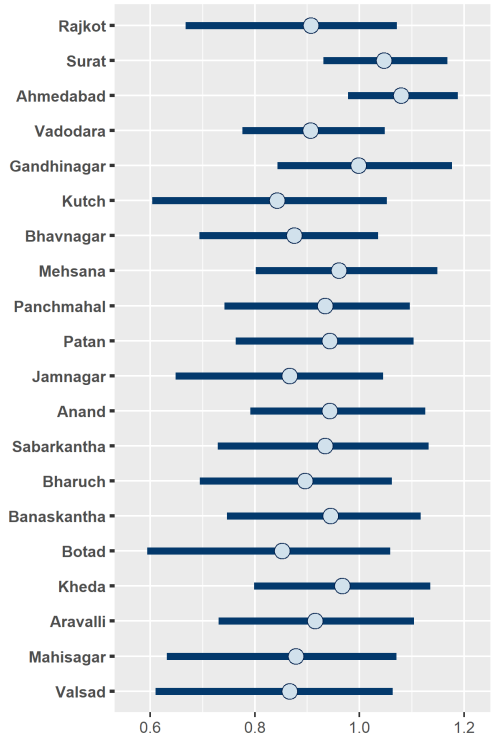


(f) Pune: under complete lockdown

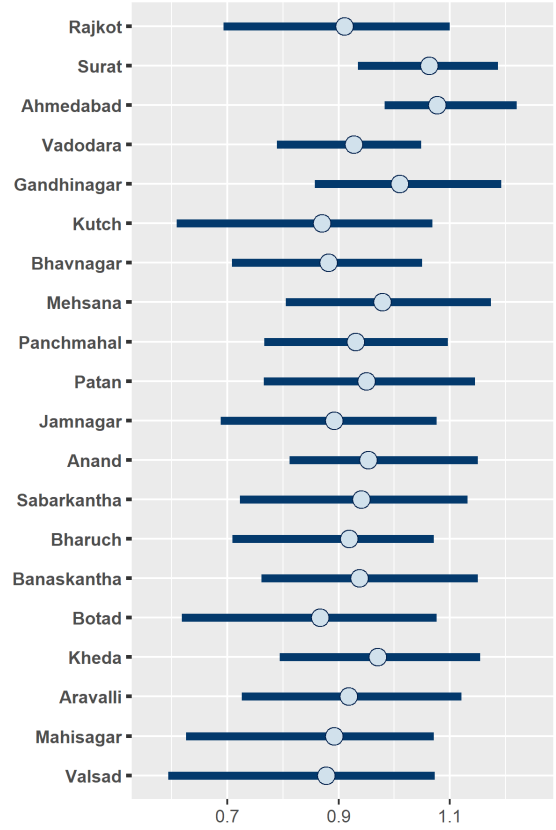
Figure 4: Daily death prediction for different districts of Maharashtra



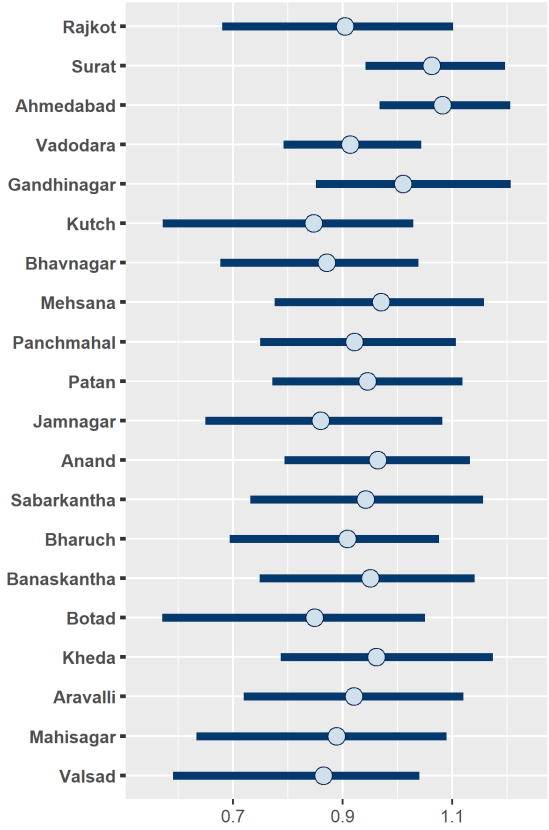
(a) Baseline reproduction number(R_0)



(b) Predicted R_t on 30th June, 2020 under no lockdown

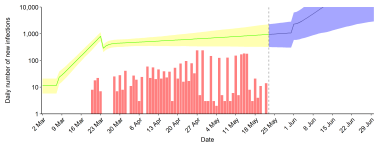


(c) Predicted R_t on 30th June, 2020 under partial lockdown

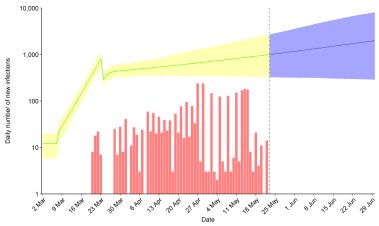


(d) Predicted R_t on 30th June, 2020 under complete lockdown

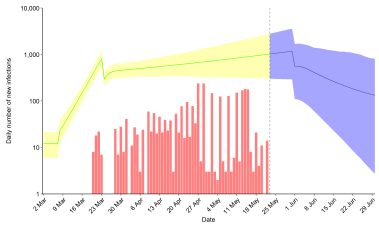
Figure 5: Average reproduction number for Maharashtra



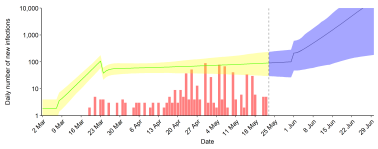
(a) Ahmedabad: under no further lockdown



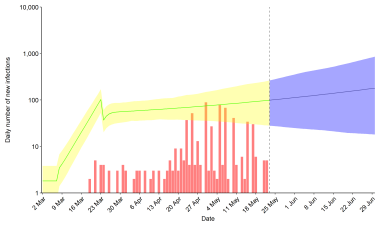
(b) Ahmedabad: under partial lockdown



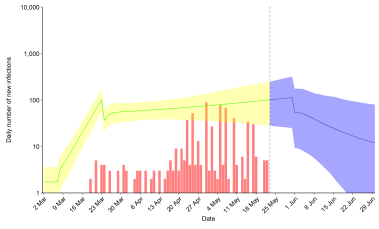
(c) Ahmedabad: under complete lockdown



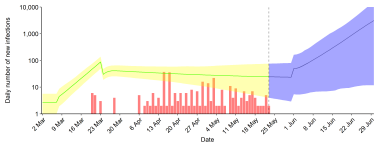
(d) Surat: under no further lockdown



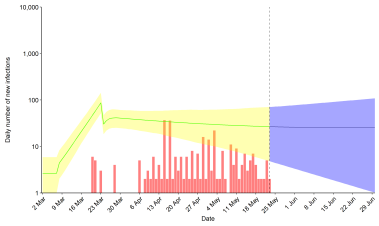
(e) Surat: under partial lockdown



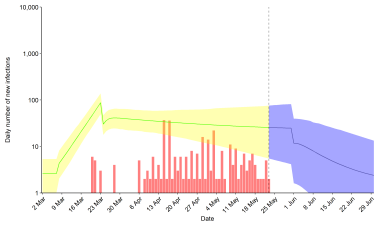
(f) Surat: under complete lockdown



(g) Vadodora: under no further lockdown

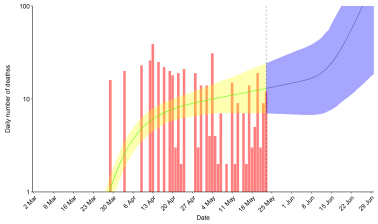


(h) Vadodora: under partial lockdown

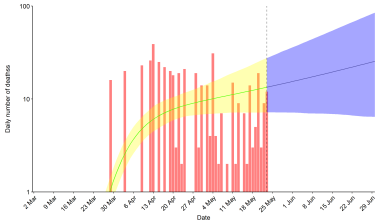


(i) Vadodora: under complete lockdown

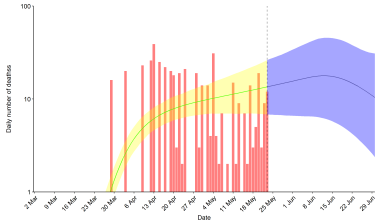
Figure 6: Daily infection prediction for different districts of Gujarat



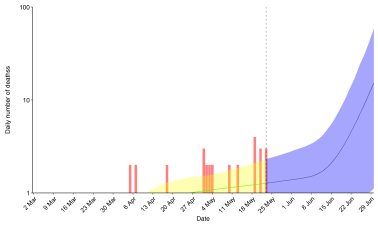
(a) Ahmedabad: under no lockdown



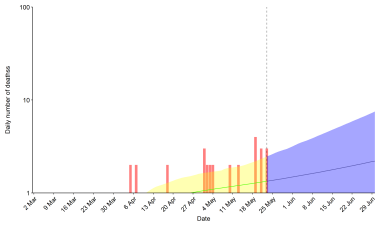
(b) Ahmedabad: under partial lockdown



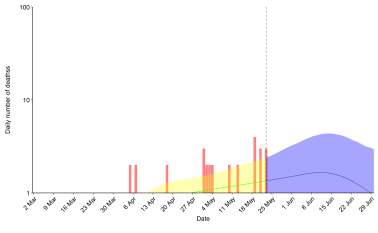
(c) Ahmedabad: under complete lockdown



(d) Surat: under no lockdown



(e) Surat: under partial lockdown



(f) Surat: under complete lockdown

Figure 7: Daily death prediction for different districts of Gujarat

Acknowledgments

The authors would like to thank the Editor and the referee for providing constructive suggestions leading to improvements in the presentation of this paper.

References

Cauchemez, S., Valleron, A. J., Boëlle, P. Y., Flahault, A. and Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature* **452**, 750–754.

Chatterjee, N. (2020). Transparency, Reproducibility, and Validity of COVID-19 Projection Models. *https://towardsdatascience.com/transparency-reproducibility-and-validity-of-covid-19-projection-models-78592e029f28*

Christopher, J. L. M. (IHME COVID-19 health service utilization forecasting team) (2020). Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. *https://doi.org/10.1101/2020.04.21.20074732*.

Cori, A., Ferguson, N. M., Fraser, C. and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, **178**, 1505–1512.

D'Souza, G. and Dowdy, D. (2020). What is Herd Immunity and How Can We Achieve It With COVID-19? *https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html* **April 10**.

Fox, Justin (2020). How Bad Is the Coronavirus? Let's Run the Numbers. *Bloomberg Opinions* https://www.bloomberg.com/opinion/articles/2020-03-05/how-bad-is-the-coronavirus-let-s-compare-with-sars-ebola-flu.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One*, **2(8)**, e758.

Korber, B., Fischer, W. F., Gnanakaran, S., Yoon, H. et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv 2020.04.29.069054. https://doi.org/10.1101/2020.04.29.069054*

Lee, Y. S., Lei, B. and Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *medRxiv 2020.04.23.20077065*, *https://doi.org/10.1101/2020.04.23.20077065*

Luo, J. (2020). When Will COVID-19 End? Data-Driven Prediction *https://www.in.gr/wp-content/uploads/2020/05/covid19predictionpaper.pdf* https://ddi.sutd.edu.sg/

Mahase, Elisabeth (2020). Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate*The British Medical Journal*, **368**. doi: https://doi.org/10.1136/bmj.m641

Nouvellet, P. et al. (2018). A simple approach to measure transmissibility and forecast incidence. *Epidemics,* **22**, 29–35, *doi:10.1016/j.epidem.2017.02.012*.

Seth, Flaxman, Mishra, S., Gandy, A. et al. (2020). Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries. *Imperial College London*, *https://doi.org/10.25561/77731*.

Walker, P. G. T., Whittaker, C. et al. (2020). The Global Impact of COVID-19 and Strategies for Mitigation and Suppression. *Imperial College London*, *https://doi.org/10.25561/* 77735.

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 223-238

Analysis and Forecasting of COVID-19 Cases Across Hotspot States of India

Khimya Tinani, K. Muralidharan, Akash Deshmukh, Bhagyashree Patil, Tanvi Salat and Rajeshwari Rajodia
Department of Statistics, Faculty of Science,
The Maharaja Sayajirao University of Baroda, Vadodara 390002, India.

Received: 18 May 2020; Revised: 15 June 2020; Accepted: 16 June 2020

Abstract

This paper attempts to develop a model to predict Novel Coronavirus affected cases in India. The virus is officially named as SARS-CoV-2 and was declared as a pandemic by WHO on 11th March 2020. This pandemic erupted in the Wuhan city of the People's Republic of China in December 2019. By now the whole world is in the grip of this virus. The first case of the COVID-19 in India was reported on 30th January 2020 in the state of Kerala. In India, the Ministry of Health and Family Welfare (MOHFW) keeps the track of COVID-19 cases daily. As of 14th June 2020, the total number of confirmed, recovered, and death cases in India are 332424, 169798 and 9520 respectively. The corresponding world statistics are 7900924, 3769712 and 433065 respectively. The disease is infectious and contagious and is affecting the health of people at large. The government and administration are trying hard to control the disease, and trying to find an effective treatment. This research aims to forecast the number of confirmed cases, recoveries and deaths of India and its six hotspot states (Maharashtra, Delhi, Tamil Nadu, Madhya Pradesh, Rajasthan, and Gujarat). To check the accuracy of the model, the first round of forecast is done from 15/4/2020 to 25/04/2020 based on the data available from 30th January 2020 to 14th April 2020. The second round of forecast is done from 16/05/2020 to 30/06/2020 based on the actual data from 30/01/2020 to 15/05/2020. Auto-Regressive Integrated Moving Average (ARIMA) model has been used to forecast the trend of COVID-19 cases in R programming.

Key words: COVID-19; Coronavirus; ARIMA; Forecast; Pandemic; Epidemic.

1. Introduction

Coronaviruses are commonly found in humans and animals. COVID-19 is an acronym that stands for the coronavirus disease of 2019. Common symptoms include fever, body ache, tiredness, and difficulty in breathing. Many affected people do not show any symptoms. The virus spreads within populations via respiratory droplets and close contact. Symptoms usually start 4 days after a person is infected with the virus. But in some people, it can take even longer for symptoms to appear or an infected person gets recovered without the appearance of any symptoms. The death rate of patients affected with COVID-19 is very less. The risk of becoming severely sick from COVID-19 increases with age. People who are critically ill are more prone to death if affected by COVID-19. The medicine for the treatment of COVID-19 is not found and the vaccine for COVID-19 is not available till 14th April 2020. However, the studies are being conducted by different countries. Since this is a statistical modeling-based

Corresponding Author: Khimya Tinani
E-mail: khimya27@yahoo.com

study, we deliberately avoid any detailed descriptions about the virus and its genesis. But to understand the inference part of this analysis, we need to supplement some information regarding its transmission and spread. The COVID-19 has four stages of transmission in line with other infectious diseases. In stage-1 the first appearance of the disease is through people with travel history, with everyone contained, their sources can be traced, and no local spread from those affected. The number of those infected would be quite low at this stage. Stage-2 is the local transmission when those who were infected and have a travel history spread the virus to close friends or family. At this stage, every person who comes in contact with the infected can be traced and isolated. Stage-3 is the community transmission when infections happen in public and a source for the virus cannot be traced. At this stage, large geographical lockdowns become important as random members of the community start developing the disease. Stage-4 is when the disease becomes an epidemic in a country, such as it was in China, with large numbers of infected people and the growing number of deaths with no end in sight. The World Health Organization declared it a pandemic. In the absence of a vaccine, social distancing has emerged as the most widely adopted strategy for mitigating and control of the virus. In India, the first novel coronavirus infection was reported on January 30 at Kerala. The cases increased to three by February 3. After this, no new cases were reported until March 1. On March 2, India reported two more positive cases, one each from Delhi and Hyderabad. By March 15, the total number of confirmed patients reached 107, most of which were linked to people with the travel history to affected countries and since then, the number of positive cases is continuously increasing. India observed a 14-hour voluntary public curfew on 22nd March 2020. This was followed by a nationwide lockdown for 21 days starting from 24 March 2020 and later extended to 3 May 2020, as the cases affected and deaths are increasing. The Indian Government feels that in the absence of lockdown this contagious disease may spread to a greater number of people and the number of hospitals may turn to be insufficient with limited equipment for the treatment of Covid-19 cases. However, understanding the seriousness of the issue, we feel that, constructing a good statistical model for inference and forecasting is the best we can contribute to this current subject. If the model fits well, then an estimate of the need for healthcare infrastructure, investment, and manpower can be anticipated.

In this paper, based on the data from January 30, 2020, till April 14, 2020, the first round of forecast was done day-wise for 11 days: 15/04/2020 till 25/04/2020 and the accuracy of the model was checked. The second round of forecast is done for 46 days: 16/05/2020 till 30/06/2020 based on actual data from January 30, 2020, till May 15, 2020. Since the forecasts for the number of days in the second round are more, we have presented only the weekly figures in the table. Auto-Regressive Integrated Moving Average (ARIMA) model has been used to predict the trend of COVID-19 cases using R programming.

2. Review of Literature

Petropoulos and Makridakis (March 2020) published the research article on forecasting the novel coronavirus COVID-19. Their paper describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides forecasts for the confirmed cases of COVID-19. Their study focuses on the cumulative daily figures aggregated globally of the three main variables like confirmed cases, deaths and recoveries. In their forecast, they predicted the cases for three variables in the period of 5 rounds. Kai Liu *et al*. (March 2020) studied that the mortality of elderly patients with COVID-19 is higher than that of young and middle-aged patients and elderly patients with COVID-19 are more likely to progress to severe disease. Khot and Nadkar (March 2020) published a valuable research paper on “The 2019 Novel Coronavirus Outbreak-A Global Threat’. They

had shown new insights into the pathophysiology, transmission dynamics, clinical features and management of this virus are developing. They said it is a highly transmissible infection but mortality is less compared to SARS and MERS. National and International health care agencies have shown appropriate co-ordination in the handling of this outbreak up till now and further international cooperation is the need of the hour. Lina *et al*. (March 2020) published a research paper on "A conceptual model for the coronavirus disease 2019 outbreak in Wuhan, China with individual reaction and governmental action". In this paper, their main purpose was to propose a conceptual model to address the individual reaction and governmental action, as well as the time-varying reporting rate. Schueller *et al*. (April 2020) had done research on COVID-19 in India on the potential impact of the 21-day Lockdown which was announced with effect from 25 March 2020 and other long-term policies. This lockdown is expected to avert a sudden and large increase in the number of infections in the short term. Additionally, interventions such as social distancing and isolation of infected individuals over several months could reduce peak infections and also interventions such as frequent hand washing, reduced mass gatherings, contact tracing, and quarantines could slow transmission and reduce overall infections. Read *et al*. (January 2020) studied and show the important information for the crisis management against the novel Coronavirus, early estimation of epidemiological parameters and epidemic predictions. Also, researchers proved that the SIR-family models at different complex levels can well capture the basic mechanism of the epidemic transmission. Liu *et al*. (February 2020) discussed on the reproductive number of COVID-19 is higher compared to SARS Coronavirus. They reviewed the basic reproduction number of the COVID-19 virus. Reproduction number is an indication of the transmissibility of a virus, representing the average number of new infections generated by an infectious person in a population. Khrapov and Loginova (2020) presented a research paper on mathematical modeling of coronavirus COVID-19, the authors used a modified system of differential equations constructed according to the SIR compartmental model. The optimal values of the model parameters, that describe the statistical data precisely, were found. Miller *et al*. (2020) published their study with an emphasis on the correlation between universal BCG vaccination policy and how it reduced morbidity and mortality of COVID-19 patients. They also found that countries without universal policies of BCG vaccination (Italy, Nederland, USA are some of them) have been more severely affected compared to countries with universal and long-standing BCG policies. BCG vaccination is a potential new tool in the fight against COVID-19. Probably a detailed statistical and mathematical treatment of modeling on this virus was done by Lin *et al*. (2019). For mathematical treatment, they used infectious disease prediction models based on differential equation prediction models and time series prediction models based on statistics and random processes. They also used the internet-based infectious disease prediction model and machine learning methods to substantiate the findings. Tania *et al*. (2020) published the research paper on "Forecasting of COVID-19 confirmed cases in different countries with ARIMA models". The aim of this study was first to find the best prediction models for daily confirmed cases in countries with a high number of confirmed cases in the world and second to predict confirmed cases with these models in order to have more readiness in healthcare systems. Ribeiro *et al*. (2020) developed efficient short-term forecasting models for forecasting the number of future cases. In their paper, they are using an autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR) and stacking-ensemble learning models for evaluating in the task of time series forecasting with one, three, and six-days ahead the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence. The models' effectiveness is evaluated based on the improvement index, mean absolute error, and symmetric mean absolute percentage error criteria. The ranking of models, from the best to the worst regarding

the accuracy, in all scenarios, is SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models.

3. Objectives

Forecast the number of COVID-19 confirmed cases for India as well as across the six hotspot states of India. Also, predict the number of deaths and recoveries amongst the number of cases of COVID-19 of India and across the hotspot states of India.

4. Data Source

This study has been conducted based on daily confirmed cases, deaths and recoveries of COVID-19 of India and only those states that are considered as hotspots of India. The data was collected from the official Indian website of COVID-19: https://www.mohfw.gov.in/ from 30 January 2020 to 15 May 2020.

5. Data Visualization

Table 1: Mortality rate and Recovery rate of six hotspots states of India for the period 30/01/2020 to 25/04/2020

Hotspots States of India	Confirmed cases	Death cases	Recovered cases	Mortality Rate per thousand	Recovery Rate per thousand
Maharashtra	7628	322	1076	42.2129	141.0593
Gujarat	3071	133	282	43.3083	91.82677
Delhi	2625	54	869	20.5714	331.0476
Rajasthan	2083	34	513	16.3226	246.2794
Madhya Pradesh	1945	100	281	51.4138	144.473
Tamil Nadu	1821	23	960	12.6304	527.1829

From the above table, it is observed that the mortality rate in Madhya Pradesh is highest when compared with other hotspots states of India. While the recovery rate in Tamil Nadu is highest and on other side mortality rate is minimal compared to other hotspot states of India.

6. Analysis and Forecasting

6.1. ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) is a stochastic approach of modeling which can be used for calculating the probability of a future value lying in a specified interval of limits. It consists of two models Autoregressive Process (AR) and Moving Average Process (MA) bind together by (I) the integration part. ARIMA models are generally used to analyze time series data for better understanding and forecasting. The ARIMA model is denoted as ARIMA (p, d, q), where the parameter p refers to the order of the AR process, q refers to the order of the MA process, and d refers to the order of differencing it takes to make the series stationary. In this study, the ARIMA model has been developed to forecast the confirmed cases, death cases and recovered cases of India cumulatively and its six hotspot states.

The ARIMA model for Confirmed cases is given as:

$$X_t = \Lambda + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots\ldots + \alpha_p X_{t-p} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \ldots\ldots + \Phi_q \varepsilon_{t-q} \quad (1)$$

where, X_t shows the forecasted values of confirmed cases, Λ is the intercept term, also estimated by the model, X_{t-i} is the lag variable at the time $t - i$ of the series, i=1, 2, …, p, α_i is the coefficient of AR process that the model estimates, ε_t is the error term and Φ_j is the coefficient of MA process where, j=1, 2, …, q.

The ARIMA model for Death cases is given as:

$$Y_t = \psi + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots\ldots + \beta_p Y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \ldots\ldots + \Theta_q \varepsilon_{t-q} \quad (2)$$

where, Y_t shows the forecasted values of death cases, ψ is the intercept term, also estimated by the model, Y_{t-i} is the lag variable at the time $t - i$ of the series, i=1, 2, …, p, β_i is the coefficient of AR process that the model estimates, ε_t is the error term and Θ_j is the coefficient of MA process where, j = 1, 2, …, q.

The ARIMA model for Recovered cases is given as:

$$Z_t = \zeta + \gamma_1 Z_{t-1} + \gamma_2 Z_{t-2} + \ldots\ldots + \gamma_p Z_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \ldots\ldots + \phi_q \varepsilon_{t-q} \quad (3)$$

where, Z_t shows the forecasted values of recovered cases, ζ is the intercept term, also estimated by the model, Z_{t-i} is the lag variable at the time $t - i$ of the series, i=1, 2, …, p, γ_i is the coefficient of AR process that the model estimates, ε_t is the error term and ϕ_j is the coefficient of MA process where, j =1, 2, …, q.

The first step to build an ARIMA model is to make the time series stationary. So, to make a series stationary, the most common approach is to difference it. Augmented Dickey Fuller test (ADF test) is a common statistical test used to test whether a given time series is stationary or not. The null hypothesis assumes that the series is non-stationary. ADF test is fundamentally a statistical significance test. That means, there is a hypothesis testing involved with a null and alternative hypothesis and as a result, a test statistic is computed and *p*-values get reported. It is from the test statistic and the *p*-value, we can make an inference as to whether a given series is stationary or not. For the identification of the model, the task is to find out the appropriate values of p and q with the help of autocorrelation function (ACF) and partial autocorrelation function (PACF) graph values. The initial number of the ARIMA model was guessed through the autocorrelation function (ACF) graph and partial autocorrelation (PACF) graph. ACF plot is merely a bar chart of the coefficients of correlation between a time series and lags of itself. The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself. According to these plots, the p and q parameters of ARIMA models were guessed. Then the guess models were compared according to Akaike Information Criterion (AIC) value, treating minimum as the best. The reason for choosing AIC is because of its wide acceptance as a statistical measure model selection. It is used to quantify the goodness of fit of the model. When comparing two or more models, the one with the lowest AIC is generally considered to be closer to real data. The appropriate ARIMA model then identified for the particular datasets and the parameters are estimated accordingly.

Having chosen the specific ARIMA model and its parameters estimated, the next step is to carry out a diagnostic check to see whether the model fits the data completely well. That is done by checking the residuals estimated from this model which are termed as white noise error or pure random error. This will decide if the chosen model fits the data well or not. For this, we use the Ljung-Box test introduced in (1978) which as a diagnostic tool to test the lack of fit of a time series model. The null hypothesis of the Ljung-Box test is given by H_0: The model does not show a lack of fit and the alternative hypothesis is H_1: the model does show a lack of fit. For a time series Y of length n, the Ljung-Box test statistic is defined as:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}_k}{n-k} \tag{4}$$

where $\hat{r}_k$ is the estimated autocorrelation of the series at lag k, and m is the number of lags being tested with a significant level α. We reject the null hypothesis and say that the model has significant lack of fit if $Q > {\chi^2}_{1-\alpha,h}$ where ${\chi^2}_{1-\alpha,h}$ is the chi-square distribution table value with h degrees of freedom and significant level α. Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that $h = m–p–q$, where p and q indicate the number of parameters from the ARIMA (p, d, q) model fit to the data. In Statistical package R, the Ljung-Box test can be run with the help of *Box.test* function.

After prediction, the accuracy is measured in percentage. We have used the Mean Absolute Error (MAE) method to compute the accuracy. Firstly, the predicted values and the actual values are stored in a single matrix with two columns, namely predicted value and actual value respectively. Then the error between the 2 columns is computed where, error =|actual value –predicted value|. The accuracy is calculated by,

$$Accuracy = 1 - \frac{error}{(actual\ value)} \tag{5}$$

which is generally reported in percentages.

6.2. First round of forecasts for the period: 15/04/2020 to 25/04/2020

Our focus is on the cumulative daily figures aggregated for India over the period from January 30, 2020 till April 14, 2020. While the data patterns show an exponential increase, the trend of confirmed cases comes to hold after it first entered India on 30th January 2020 until February 2020. From March beginning there was a sudden increase in cases, while deaths started to happen frequently only after 11th March onwards. The recovery of patients started to happen simultaneously from mid-February onwards. We have aimed our research to forecast the number of confirmed cases, recoveries and deaths of India and its six hotspot states (Maharashtra, Delhi, Tamil Nadu, Madhya Pradesh, Rajasthan and Gujarat). Based on the data from January 30, 2020, till April 14, 2020, the first round of forecast was done day-wise for the period of 11 days: 15/04/2020 to 25/04/2020 and the accuracy of the model was checked. The analysis is done in R programming and the necessary packages: library(hrbrthemes), library(dplyr), library(ggplot2), library(tseries), library(forecast) are loaded.

Now before we analyze the time series data for actual forecast, we use the Augmented Dickey Fuller test to check the stationary of the time series observations. The null hypothesis (H_0) for the test is that the data is not stationary whereas the alternative hypothesis (H_1) is that the data is stationary. The level of significance is taken to be 0.05. The output is obtained for confirmed cases using *adf.test* function in R programming. Here, the p-value turns out to be

0.99. We thus fail to reject our H_0 and conclude that the data is not stationary. We now have to work on the stationarity of the data. After differencing the time series for consecutively for two times, the *p*-value is obtained as 0.01, which is less than 0.05, and hence we reject the null hypothesis and conclude that the time series for confirmed cases is stationary. Since the order of differencing is 2, $d = 2$. Similarly, we have found that stationary time series for deaths and recoveries cases.

Figures 1 and 2 show the ACF and PACF plots for confirmed cases. These plots are used for choosing the model parameters for confirmed cases. Similarly, we have found model parameters for deaths and recoveries using ACF and PACF plots.

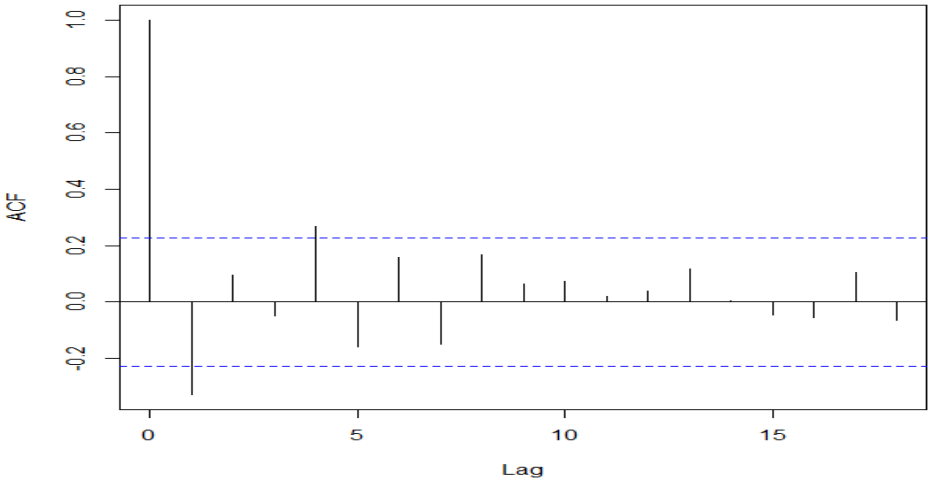


Figure 1: Plot of ACF for confirmed cases

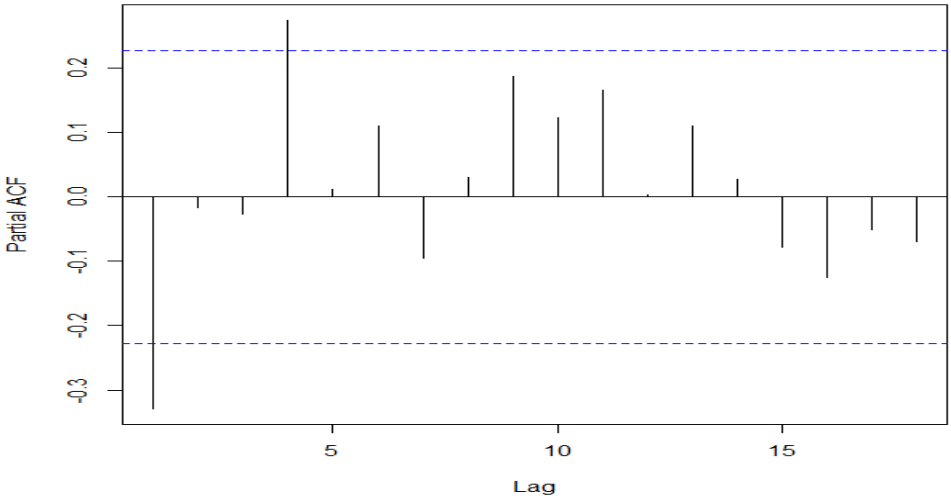


Figure 2: Plot of PACF for confirmed cases

According to ACF and PACF plots, the *p* and *q* parameters of ARIMA models are guessed. These guess models are compared according to AIC value. Table 2 presents all those ARIMA models with corresponding AIC values for all three types of cases.

Table 2: ARIMA models with all possible values of AIC for India

Confirmed	AIC	Deaths	AIC	Recovered	AIC
ARIMA(0,2,0)	853.724	ARIMA(0,2,0)	473.335	ARIMA(1,2,0)	627.8381
ARIMA(2,2,0)	848.897	ARIMA(1,2,2)	399.478	ARIMA(5,2,0)	599.7987
ARIMA(3,2,0)	850.859	ARIMA(1,2,1)	423.039	ARIMA(1,2,2)	617.8326
ARIMA(1,2,0)	849.184	ARIMA(1,2,0)	425.797	ARIMA(2,2,0)	606.2791
ARIMA(0,2,1)	850.707	ARIMA(0,2,1)	439.349	ARIMA(0,2,0)	627.6912
ARIMA(0,2,0)	853.724	ARIMA(0,2,0)	473.335	ARIMA(1,2,0)	627.8381

The model which has the least AIC is selected as the best model. Accordingly, the best ARIMA models for forecasting the number of daily confirmed, deaths and recovered cases for India are ARIMA(2,2,0), ARIMA(1,2,2), ARIMA(5,2,0) respectively for India. The first round of forecast is shown in figure 3. The same in actual numbers are presented in Table 3.

The equation corresponding to the best ARIMA(2,2,0) model for confirmed cases is given by

$$X_t = 15.3463 - 0.3524X_{t-1} + 50.1764X_{t-2} + \varepsilon_t \quad (6)$$

The equation corresponding to the best ARIMA(1,2,2) model for death cases is given by

$$Y_t = 0.4077 - 0.2613Y_{t-1} + \varepsilon_t - 0.7937\varepsilon_{t-1} + 0.7014\varepsilon_{t-2} \quad (7)$$

The equation corresponding to the best ARIMA(5,2,0) model for recovery cases is given by,

$$Z_t = 4.860 + 0.085Z_{t-1} + 0.261Z_{t-2} + 0.444Z_{t-3} + 0.632Z_{t-4} + 0.622Z_{t-5} + \varepsilon_t \quad (8)$$

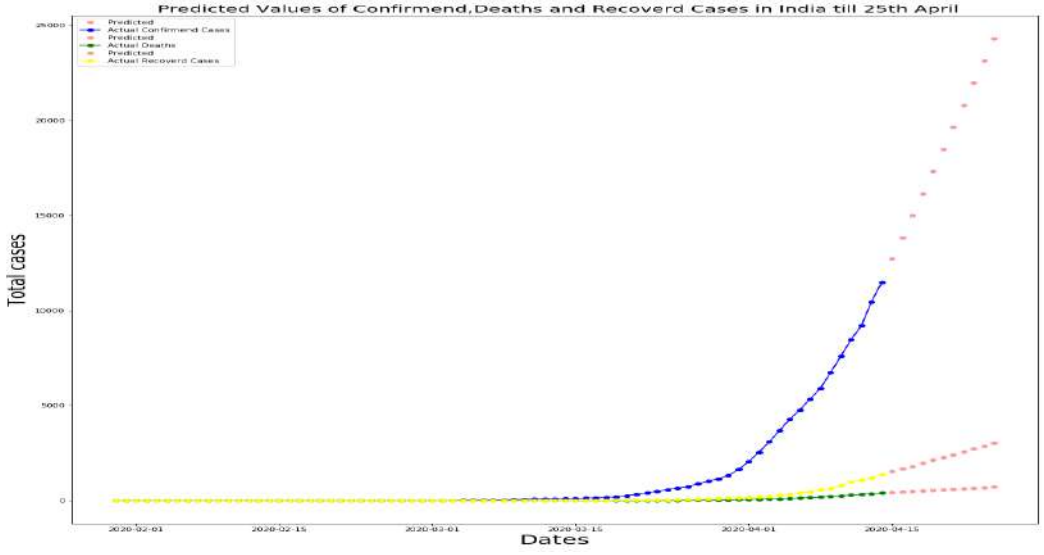


Figure 3: Plot of actual and forecasts of COVID-19 cases in India

The blue dots represent the actual confirmed cases, yellow dots represent recovered cases and green dots represent the actual deaths. The extended red dots represent forecasted COVID-19 cases.

Table 3: Actual and forecast values of COVID-19 with 95% CI for India

Date	Actual values			Forecast values		
	Confirmed	Death	Recovered	Confirmed	Death	Recovered
15-04-20	12370	422	1509	12707 (12564, 12850)	427 (420, 433)	1538 (1512, 1563)
16-04-20	13431	448	1767	13817 (13534, 14100)	456 (446, 465)	1664 (1613, 1715)
17-04-20	14353	486	2040	15012 (14525, 15498)	485 (469, 500)	1793 (1719, 1866)
18-04-20	15724	521	2466	16152 (15438, 16865)	514 (490, 537)	1965 (1861, 2069)
19-04-20	17304	559	2854	17330 (16352, 18308)	543 (510, 576)	2130 (1983, 2278)
20-04-20	18543	592	3273	18482 (17217, 19748)	572 (528, 615)	2261 (2065, 2458)
21-04-20	20080	645	3976	19653 (18072, 21234)	601 (546, 656)	2404 (2158, 2649)
22-04-20	21372	681	4370	20811 (18894, 22728)	630 (562, 698)	2575 (2273, 2877)
23-04-20	23039	721	5012	21978 (19702, 24253)	659 (578, 740)	2735 (2367, 3102)
24-04-20	24447	780	5496	23138 (20485, 25792)	688 (593, 783)	2871 (2434, 3308)
25-04-20	26282	824	5939	24303 (21253, 27354)	717 (607, 827)	3021 (2512, 3530)

From the above table, it is noted that the day-wise estimated figures for confirmed cases from 15 April 2020 to 25 April 2020 are nearly the same. However, the day-wise estimated recoveries are less than the actual values. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The output is obtained by using the *Box.test* function in R programming. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. Here, for confirmed cases *p*-value is 0.7315, for deaths *p*-value is 0.49863 and for recoveries, the *p*-value is 0.9585. As for all the cases, *p*-value is greater than 0.05, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit. The accuracy of prediction for India is computed by averaging the accuracies obtained by the algorithm of ARIMA modeling. As per this modeling, the accuracy for confirmed cases is 98%, for the deaths 97% and for the recoveries is 78%.

Now we will forecast the figures for the highly affected states in India assuring that the data is stationary and reliable to forecast. The final models that are reported in table 4 have the lowest AIC values for all hotspot states of India. The equations of best ARIMA model can be mentioned for all the hotspots states of India in the same way as we mentioned for India. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The outputs for all six hotspot states of India are given in table 4. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. The *p*-value for state Rajasthan is less than 0.05 for deaths and recoveries, hence we reject the null hypothesis and conclude that model does show lack of fit whereas *p*-value for Rajasthan is more than 0.05 for confirmed cases, hence we do not reject the null hypothesis and conclude that model does not show lack of fit for confirmed cases. For

the other hotspot states *p*-value is greater than 0.05 for confirmed, deaths and recovery cases, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Table 4: The best ARIMA models with least AIC for six hotspots states of India

Hotspot States	Cases	ARIMA Model	AIC	Ljung-Box *p*-value
Maharashtra	Confirmed	ARIMA(1,2,0)	695.1979	0.9091
	Deaths	ARIMA(2,2,2)	333.7695	0.9026
	Recovered	ARIMA(1,2,1)	519.3081	0.9072
Delhi	Confirmed	ARIMA(2,2,0)	741.8226	0.3344
	Deaths	ARIMA(3,2,0)	151.5864	0.7458
	Recovered	ARIMA(0,2,2)	233.4595	0.8812
Madhya Pradesh	Confirmed	ARIMA(2,2,0)	565.2183	0.1515
	Deaths	ARIMA(1,2,1)	210.2508	0.9744
	Recovered	ARIMA(0,2,1)	322.8164	0.1021
Tamil Nadu	Confirmed	ARIMA(2,2,2)	614.3582	0.1434
	Deaths	ARIMA(0,2,2)	57.20068	0.9963
	Recovered	ARIMA(0,2,5)	344.3379	0.8514
Gujarat	Confirmed	ARIMA(3,2,1)	557.0561	0.9976
	Deaths	ARIMA(3,2,0)	74.23162	0.9909
	Recovered	ARIMA(0,2,2)	297.4889	0.9622
Rajasthan	Confirmed	ARIMA(0,2,3)	551.4806	0.6519
	Deaths	ARIMA(0,2,1)	168.3248	0.0320
	Recovered	ARIMA(2,2,1)	496.3972	0.0283

Forecast values of ARIMA models with a confidence interval for six hotspot states of India are given in Table 5.

Table 5: Forecast values of COVID-19 cases with 95% CI for six hotspot states of India

Date	Cases	Maharashtra	Delhi	Madhya Pradesh	Tamil Nadu	Gujarat	Rajasthan
15-04-20	Confirmed	3028 (2977, 3079)	1801 (1732, 1869)	764 (744, 785)	1255 (1226, 1283)	735 (715, 754)	1093 (1074, 1112)
	Death	191 (187, 196)	32 (31, 34)	59 (57, 61)	12 (12, 13)	30 (29, 31)	11 (10, 13)
	Recovered	277 (262, 293)	32 (30, 35)	72 (68, 76)	85 (81, 90)	64 (61, 68)	163 (150, 176)
16-04-20	Confirmed	3375 (3275, 3475)	1998 (1906, 2090)	842 (816, 869)	1318 (1263,1373)	803 (758, 849)	1195 (1155, 1234)
	Death	206 (200, 212)	35 (33, 38)	63 (60, 66)	13 (12, 14)	32 (31, 33)	12 (10, 14)
	Recovered	299 (276, 322)	34 (31, 37)	81 (74, 88)	94 (87,101)	70 (66, 75)	172 (154, 189)
17-04-20	Confirmed	3723 (3561, 3884)	2140 (2003, 2276)	947 (911, 983)	1367 (1278, 1455)	868 (797, 939)	1291 (1221, 1361)
	Death	222 (211, 232)	39 (35, 43)	68 (64, 73)	14 (12, 16)	34 (33, 36)	13 (10, 16)
	Recovered	320 (288, 353)	35 (31, 39)	89 (79, 99)	102 (94, 111)	76 (70, 83)	190 (168, 212)

18-04-20	Confirmed	4070 (3839, 4301)	2366 (2178, 2555)	974 (921, 1028)	1404 (1272, 1537)	937 (843, 1032)	1387 (1285, 1489)
	Death	237 (221, 252)	42 (37, 47)	73 (67, 79)	15 (13, 17)	36 (34, 38)	14 (10, 17)
	Recovered	342 (299, 384)	37 (32, 42)	98 (85, 111)	116 (105, 126)	82 (74, 91)	206 (177, 236)
19-04-20	Confirmed	4417 (4109, 4726)	2530 (2292, 2767)	1068 (1003, 1133)	1454 (1271, 1637)	1014 (893, 1135)	1483 (1347, 1619)
	Death	252 (231, 272)	45 (39, 52)	77 (70, 85)	16 (13, 18)	39 (36, 41)	15 (11, 19)
	Recovered	363 (309, 417)	38 (32, 44)	106 (90, 123)	126 (112, 140)	88 (78, 99)	219 (183, 256)
20-04-20	Confirmed	4765 (4371, 5158)	2716 (2418, 3015)	1152 (1072, 1232)	1512 (1277, 1748)	1086 (935, 1238)	1579 (1407, 1752)
	Death	267 (240, 294)	48 (40, 56)	82 (72, 92)	16 (13, 20)	41 (38, 44)	16 (11, 21)
	Recovered	384 (318, 450)	39 (32, 47)	115 (95, 135)	137 (118, 155)	94 (82, 107)	235 (192, 279)
21-04-20	Confirmed	5112 (4628, 5596)	2912 (2551, 3272)	1189 (1090, 1288)	1562 (1271, 1852)	1157 (972, 1342)	1676 (1464, 1887)
	Death	282 (249, 315)	51 (42, 61)	87 (75, 98)	17 (13, 21)	43 (39, 47)	17 (11, 22)
	Recovered	405 (327, 483)	41 (32, 50)	123 (100, 147)	147 (125, 170)	100 (85, 115)	251 (199, 303)
22-04-20	Confirmed	5459 (4878, 6041)	3085 (2658, 3511)	1289 (1175, 1403)	1603 (1252, 1955)	1226 (1006, 1446)	1772 (1519, 2024)
	Death	297 (257, 337)	55 (43, 66)	91 (77, 105)	18 (14, 22)	45 (41, 50)	17 (11, 24)
	Recovered	426 (335, 518)	42 (32, 52)	132 (104, 160)	158 (130, 186)	106 (89, 124)	265 (205, 326)
23-04-20	Confirmed	5807 (5122, 6491)	3278 (2780, 3775)	1357 (1224, 1491)	1652 (1235, 2070)	1299 (1043, 1555)	1868 (1572, 2164)
	Death	312 (265, 360)	58 (44, 71)	96 (80, 112)	19 (14, 23)	46 (42, 53)	18 (11, 25)
	Recovered	447 (342, 553)	44 (32, 55)	140 (109, 172)	169 (136, 201)	112 (92, 132)	281 (211, 350)
24-04-20	Confirmed	6154 (5362, 6947)	3461 (2890, 4031)	1408 (1254, 1562)	1708 (1222, 2193)	1371 (1077, 1665)	1964 (1623, 2305)
	Death	328 (272, 383)	61 (45, 76)	101 (82, 119)	19 (14, 25)	50 (44, 56)	19 (12, 27)
	Recovered	469 (349, 589)	45 (32, 58)	149 (113, 185)	179 (141, 217)	118 (96, 141)	296 (217, 375)
25-04-20	Confirmed	6502 (5595,7408)	3644 (2996, 4292)	1506 (1334, 1678)	1757 (1201, 2313)	1442 (1108, 1777)	2060 (1672, 2449)
	Death	343 (279, 406)	64 (46, 82)	105 (84, 126)	20 (14, 26)	52 (45, 58)	20 (12, 28)
	Recovered	490 (355, 625)	47 (32, 61)	158 (117, 198)	190 (146, 233)	124 (99, 150)	311 (222, 400)

From the above table, it can be noted that the day-wise estimated figures for confirmed cases from 15 April 2020 to 25 April 2020 are nearly the same. However, the day-wise estimated recoveries and deaths are less than the actual values. The accuracy of prediction for six hotspots states of India is computed by averaging the accuracies obtained by the algorithm of ARIMA modeling. The result is given below in Table 6.

Table: 6 Model Accuracy for six hotspot states of India

Hotspot States	Confirmed	Deaths	Recovered
Maharashtra	92%	95%	71%
Delhi	68%	91%	66%
Madhya Pradesh	81%	80%	79%
Tamil Nadu	98%	85%	53%
Rajasthan	76%	26%	31%
Gujarat	82%	77%	78%

For all six hotspots states of India, the ARIMA model accuracy of confirmed cases forecasted is 83% on an average, which indicates that ARIMA gives good accuracy of prediction. On the other hand, model accuracy for death and recovery cases in six hotspots states of India is around 76% and 58% respectively. This seems we need a better model for forecasting death and recovery cases in hotspots states of India.

6.3. Second round of forecasts for the period: 16/05/2020 till 30/06/2020

While writing this paper, the number of cases in India is doubling up every day, and hence the prediction after 25th April may not match with our estimated values. The forecast of COVID-19 cases until 25thApril 2020 is nearly the same as per our actual cases. This needs further investigation. One of the reasons could be to revise the base data for prediction, a lot of the administrative level containment measures started in between. For instance, the complete lockdown for three weeks from March 24, 2020, onwards. Similar forecasting is done for India and the hotspots states of India based on actual data from January 30, 2020, till May 15, 2020, and forecast is done for the period of 46 days: 16/05/2020 till 30/06/2020. The model summary for India with the least AIC is presented in Table 7.

Table 7: ARIMA models with all possible values of AIC for India

Confirmed		Deaths		Recovered	
ARIMA model	AIC	ARIMA model	AIC	ARIMA model	AIC
ARIMA(1,2,0)	1421.854	ARIMA(0,2,2)	789.6459	ARIMA(1,2,0)	1282.373
ARIMA(0,2,1)	1426.392	ARIMA(0,2,0)	823.0364	ARIMA(0,2,0)	1313.382
ARIMA(2,2,0)	1423.777	ARIMA(1,2,0)	793.7753	ARIMA(0,2,1)	1293.886
ARIMA(0,2,0)	1441.791	ARIMA(1,2,1)	791.0409	ARIMA(2,2,0)	1283.265
ARIMA(1,2,1)	1423.807	ARIMA(1,2,3)	792.4254	ARIMA(1,2,1)	1284.015

The model which has the least AIC is selected as the best model. The best ARIMA models for forecasting the number of daily confirmed, deaths and recovered cases are ARIMA(1,2,0), ARIMA(0,2,2), ARIMA(1,2,0) respectively for India. Weekly forecasts of COVID-19 with confidence interval are presented in Table 8.

The equation corresponding to the best ARIMA(1,2,0) model for confirmed cases is given by

$$X_t = 19.9474 - 0.4333X_{t-1} + \varepsilon_t \quad (9)$$

The equation corresponding to the best ARIMA(0,2,2) model for death cases is given by

$$Y_t = 0.6201 + \varepsilon_t - 0.6618\varepsilon_{t-1} + 0.2145\varepsilon_{t-2} \quad (10)$$

The equation corresponding to the best ARIMA (1,2,0) model for recovery cases is given by

$$Z_t = 10.904 - 0.5950Z_{t-1} + \varepsilon_t \quad (11)$$

Table 8: Weekly forecast values of COVID-19 with 95% CI for India

Date	Confirmed	Deaths	Recovered
16-05-20	89742 (89335, 90149)	2861 (2841, 2880)	32098 (31888, 32307)
23-05-20	116778 (112499, 121056)	3602 (3432, 3772)	46119 (44118, 48120)
30-05-20	143821 (133520, 154122)	4343 (3939, 4748)	60084 (55293, 64875)
06-06-20	170864 (153005, 188723)	5085 (4387, 5783)	74051 (65762, 82340)
13-06-20	197908 (171220, 224595)	5826 (4785, 6868)	88018 (75644, 100391)
20-06-20	224951 (188327, 261575)	6568 (5140, 7996)	101984 (85016, 118953)
27-06-20	251994 (204438, 299550)	7309.9 (5457, 9162)	115951 (93927, 137975)

Forecasted confirmed COVID-19 cases would be 263584, deaths would be 7627 and recoveries would be 121937 on 30th June 2020. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: the model does show a lack of fit. Here, for confirmed cases *p*-value is 0.6307, for deaths *p*-value is 0.8192 and for recoveries, *p*-value is 0.1003. As for all the cases, *p*-value is greater than 0.05, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Now we will forecast the figures for the highly affected states in India assuring that the data is stationary and reliable to forecast. The final models that are reported in table 9 have the lowest AIC values for all hotspot states of India. To estimate model adequacy, the Ljung-Box test which is a diagnostic tool is used to test the lack of fit of a time series model. The output for all six hotspot states of India is given in Table 9. The null hypothesis, H_0: The model does not show a lack of fit. The alternative hypothesis, H_1: The model does show a lack of fit. The *p*-value for state Rajasthan is less than 0.05 for deaths and recoveries, hence we reject the null hypothesis and conclude that model does show lack of fit whereas *p*-value for Rajasthan is more than 0.05 for confirmed cases, hence we do not reject the null hypothesis and conclude that model does not show lack of fit for confirmed cases. For the other hotspot states *p*-value is greater than 0.05 for confirmed, deaths and recovery cases, hence we do not reject the null hypothesis and conclude that our model does not show a lack of fit.

Table 9: The best ARIMA models with least AIC for six hotspots states of India

Hotspot States	Cases	ARIMA Model	AIC	Ljung-Box p-value
Maharashtra	Confirmed	ARIMA(2,2,2)	1338.927	0.7898
	Deaths	ARIMA(2,2,3)	571.9736	0.9937
	Recovered	ARIMA(2,2,2)	1101.577	0.3162
Delhi	Confirmed	ARIMA(2,2,2)	1159.906	0.8447
	Deaths	ARIMA(0,2,2)	449.952	0.9321
	Recovered	ARIMA(1,2,2)	1166.805	0.9672
Madhya Pradesh	Confirmed	ARIMA(1,2,1)	1030.119	0.9634
	Deaths	ARIMA(0,2,1)	456.2507	0.3309
	Recovered	ARIMA(1,2,2)	936.19	0.2378
Tamil Nadu	Confirmed	ARIMA(3,2,0)	1126.77	0.1057
	Deaths	ARIMA(2,2,0)	234.3007	0.8082
	Recovered	ARIMA(2,2,2)	1049.927	0.9991
Gujarat	Confirmed	ARIMA(0,2,1)	1008.264	0.8712
	Deaths	ARIMA(2,2,1)	563.1694	0.9569
	Recovered	ARIMA(4,2,2)	1032.743	0.9571
Rajasthan	Confirmed	ARIMA(1,2,0)	944.7558	0.9514
	Deaths	ARIMA(1,2,2)	396.4093	0.0182
	Recovered	ARIMA(2,2,2)	980.4154	0.0397

Forecast values of ARIMA models with confidence interval for six hotspot states of India are given in table 10.

Table 10: Weekly forecast values of COVID-19 with 95% CI for hotspot states of India

Date	Cases	Maharashtra	Delhi	Madhya Pradesh	Tamil Nadu	Gujarat	Rajasthan
16.05.20	Confirmed	30501 (30233, 30770)	9258 (9144, 9372)	4760 (4697, 4823)	10605 (10507, 10704)	10270 (10213,10327)	4956 (4914,4998)
	Deaths	1116 (1109, 1123)	132 (128, 136)	245 (241, 249)	75 (74, 82)	628 (621, 634)	128 (125, 131)
	Recovered	7064 (6978, 7150)	3874 (3755, 3993)	2407 (2368, 2447)	2832 (2765, 2899)	4119 (4058, 4181)	2810 (2761, 2859)
23.05.20	Confirmed	41154 (39120,43188)	11838 (11152, 12523)	5897 (5401, 6393)	14184 (12888, 15480)	12639 (11977, 13300)	6430 (6013, 6848)
	Deaths	1451 (1380, 1521)	194 (162, 226)	280 (258, 303)	102 (90, 110)	772 (705, 839)	151 (134, 168)
	Recovered	10802 (10045, 11559)	6008 (5158, 6858)	3291 (3054, 3528)	4094 (3712, 4476)	5921 (5314, 6529)	3411 (3098, 3724)
30.05.20	Confirmed	51836 (4669, 56981)	14565 (12730, 16401)	7031 (6029, 8032)	17737 (14685, 20790)	15007 (13412, 16603)	7905 (6903, 8906)
	Deaths	1786 (1609, 1963)	256 (187, 325)	316 (270, 363)	129 (101, 137)	917 (757, 1078)	174 (141, 207)
	Recovered	14514 (12581,16448)	8167 (6333, 10001)	4177 (3554, 4800)	5466 (4463, 6470)	7504 (6200, 8808)	4007 (3277, 4737)
06.06.20	Confirmed	62516 (53417, 71615)	17283 (13927, 20639)	8164 (6572, 9756)	21288 (16043, 26532)	17376 (14609, 20143)	9379 (7644,11113)
	Deaths	2121 (1809, 2433)	319 (205, 433)	352 (276, 428)	157 (108,165)	1062 (784,1340)	198 (145, 250)
	Recovered	18221 (14794, 21648)	1032 (7287,13359)	5063 (3932, 6194)	6846 (5048, 8644)	9192 (7019, 11366)	4604 (3348, 5860)

13.06.20	Confirmed	73196 (59454, 86938)	20002 (14840, 25164)	9298 (7036, 11559)	24838 (17040, 32637)	19745 (15609, 23881)	10853 (8262, 13443)
	Deaths	2456 (1985, 2926)	381 (215, 548)	388 (279, 497)	184 (112, 192)	1207 (792, 1622)	221 (146, 296)
	Recovered	21926 (16745, 27108)	12480 (8056,16903)	5949 (4215, 7684)	8226 (5492, 10959)	10932 (7742, 14122)	5201 (3332, 7070)
20.06.20	Confirmed	2790 (2142, 3439)	22720 (15508, 29932)	10431 (7428, 13434)	28389 (17719, 39059)	22114 (16438, 27790)	12327 (8773, 15881)
	Deaths	2790 (2142, 3439)	444 (219, 668)	424 (278, 570)	212 (113, 219)	1353 (783, 1922)	244 (145, 344)
	Recovered	25631 (18469, 32793)	14636 (8661, 20611)	6835 (4416, 9254)	9605 (5812, 13398)	12625 (8299, 16951)	5798 (3239, 8357)
27.06.20	Confirmed	94557 (69798, 119316)	25438 (15959, 34917)	11564 (7754, 15375)	31939 (18113, 45766)	24482 (17111, 31854)	13801 (9188, 18414)
	Deaths	3125 (2280, 3971)	506 (217, 795)	459 (273, 646)	239 (111, 247)	1498 (759, 2237)	268 (141, 394)
	Recovered	29336 (19990, 38682)	16792 (9119, 24466)	7721 (4545, 10898)	10985 (6023,15947)	14312 (8742, 19882)	6395 (3078, 9712)

7. Discussion and Conclusions

In this paper, we have conducted a two-round study of COVID-19 cases in India and six hotspots states of India. Model accuracy is checked for the first round and then the predication is verified from 15 April 2020 to 25 April 2020. The first-round model is built on data of cumulative confirmed, recovery and death cases from 30 January 2020 to 14 April 2020. We have evaluated the accuracy of the ARIMA model in predicting cumulative confirmed, recovery and death cases. For all six hotspots states of India, the ARIMA model in predicting cumulative confirmed cases is 83% on average which indicates that ARIMA has given good accuracy of prediction. If we discuss country India, forecasted cumulative confirmed cases give 98% model accuracy using the ARIMA model. While model accuracy of cumulative recovery cases and death cases are 97% and 78% respectively. On the other hand, model accuracy for death and recovery cases in six hotspots states of India is 76% and 58% respectively. This seems we need a better model for forecasting death and recovery cases in hotspots states of India. Thus, through this model forecasted confirmed cases are more reliable than with death cases and recovery cases in six hotspots states of India. We hope that our forecasts will be a useful tool for governments and individuals towards making decisions and taking the appropriate actions to curb the spreading of the virus.

There are certain limitations in the numbers of COVID-19 cases forecasted. The forecast is based on past data and information, whereas the technology changes with time and medical science are in the process of doing inventions for the betterment of mankind. If new methods or medicines are invented for the treatment of COVID-19, the figures forecasted may vary. The numbers forecasted may also vary if the effective methods are not adopted or medicines or vaccines are not invented for the treatment of COVID-19 cases. Depending upon the resources, if a greater number of tests are conducted nationwide, the better management of the disease can be done and more spread of disease can be avoided. While considering figures forecasted, we should understand that we have not considered urban-rural variations, stratification of age, occupation, pre-existing co-morbidities, travel history which alters the outcomes. The testing rate is lower in India than in different countries, so our absolute numbers might below. If there is a substantial increase in tests, it may also affect the numbers forecasted. If healthcare facilities are increased, the forecasted figures may alter.

Acknowledgements

The authors express their gratefulness to the reviewer and the editor for their invaluable comments and suggestions, which have helped us to improve the paper substantially.

References

Kai, L., Ying, C., Ruzheng, L. and Kunyuan, H. (March, 2020). Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *Journal of Infection*, **6,** *e*14-*e*18.

Khot, W. Y. and Nadkar, M. Y. (March, 2020). The 2019 novel coronavirus outbreak – A global threat. *Journal of the Association of Physicians of India*, **68**.

Khrapov, P.V. and Loginova, A.A. (2020). Mathematical modelling of the dynamics of the Coronavirus COVID-19 epidemic development in China. *International Journal of Open Information Technologies*, **8(4),** 13-16.

Lin, J., Kewen, L., Jiang, Y., Xin, G. and Ting, Z. (March, 2020). Prediction and analysis of coronavirus disease 2019. *arXiv.org>q-bio>arXiv*, 2003.05447.

Lina, Q., Zhaob, S., Daozhou, G., Loue Y., Shu, Y., Musa, S., Wangb, M. H., Caig, Y., Wang, W., Yangh, L. and Hee, D. (March, 2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases*, **93**, 211-216.

Liu, Y., Gayle A. A., Wilder-Smith, A. and Rocklöv, J. (February, 2020). The reproductive number of COVID-19 is higher compared to SARS Coronavirus. *Journal of Travel Medicine*, **27,** 1-4.

Miller, A., Reandelar, M. J., Fasciglione, K., Roumenova, V., Li, Y. and Otazu, G. H. (March 2020). Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. *CC-BY-ND 4.0 International license.* https://doi.org/10.1101/2020.03.24.20042937.

Petropoulos, F. and Makridakis, S. (March, 2020). Forecasting on the novel coronavirus COVID-19. *PLOS ONE*, https://doi.org/10.1371/journal.pone.0231236.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A. and Jewell, C. P. (January, 2020). Early prediction of the 2019 novel coronavirus outbreak in the mainland China based on simple mathematical model. *IEEE*, **8,** 51761-51769.

Ribeiro, M.H.D.M., Da Silva, R.G., Mariani, V.C. and Coelho, L.S. (May, 2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons and Fractals,* **135,** 1-10.

Schueller, E., Klein, E., Tseng, G. K., Balasubramanian, R., Kapoor, G., Joshi, J., Sriram, A., Nandi, A. and Laxminarayan, R. (April, 2020). COVID-19 in India: Potential Impact of the Lockdown and Other Longer-Term Policies. *The Centre for Disease Dynamics, Economics and Policy.*

Tania D., Mardani-Fard, H.A. and Paria, D. (March 2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA. *MedRxiv,* preprint. doi: https://doi.org/10.1101/2020.03.13.20035345.

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 239-251

Impact of COVID-19 on Payment Transactions

Ashish Das[1], Suchismita Das[2], Aashima Jaiswal[2] and Tushar Sonthalia[2]
[1]Department of Mathematics, Indian Institute of Technology Bombay, Mumbai 400076, India
[2]Department of Data Science, S P Jain School of Global Management, Mumbai 400070, India

Received: 10 June 2020; Revised: 26 June 2020; Accepted: 27 June 2020

Abstract

Due to the extreme contagious nature of the COVID-19 virus, the Government of India has had to implement several restrictions to curb the outbreak. The brunt of the economic consequences of the restrictions has been faced majorly by the aviation, tourism and hospitality sector. Nevertheless, many other small/medium/large services are facing the economic consequences.

With the lockdown allowing mostly the essential sectors to continue to function, the spending of disposable income is expected to reduce drastically. The cash withdrawals at ATMs is a prominent measure of the general retail economic activity. We show its impact along with the impact on other digital payment modes such as Debit/Credit cards, UPI, IMPS, NEFT and RTGS. We present the extent to which COVID-19 and the subsequent nationwide lockdown and slow unlocking, during April-May-June 2020, has impacted the financial transactions in the country. The sheer slowdown of the economy gets depicted by the drastic reduction of retail payment activities in the country.

Based on transaction data during Q4 of FY20 and Q1 of FY21, we relate the economic impact of COVID induced lockdown and its subsequent relaxations. Although the economy is likely to suffer in FY21, we have begun to see some form of cautious and calibrated opening up of economic activity, as we see payment transactions picking up starting June 2020. The positive effect of COVID can be seen in form of increased levels of BHIM-UPI usage, as more and more people learn to use this mobile app-based, easy to use, digital mode of payment.

Key words: ATM cash withdrawal; Digital transactions; Polynomial trend.

1. Introduction

The first case of COVID-19 was reported in January end. While the number of cases remained low in February, the number of cases started to rise in March, which led the government to implement a nationwide lockdown in the country. As on date, the number of new cases is constantly increasing every day.

Corresponding Author: Ashish Das,
Email: ashish@math.iitb.ac.in

On 1st June, the nationwide lockdown (which started from 25th March) was extended till 30th June to combat COVID-19. This increases the total lockdown period to 98 days. We provide an update on the latest COVID-19 status for India as a whole.

Some of the key observations include:

• Confirmed cases: As of 25th June, India's COVID-19 confirmed cases stood at 4.9 lakh. Before the first phase of the lockdown was initiated from 25th March, India's total confirmed COVID-19 cases stood at 519. Currently, confirmed cases are doubling in about 18 days. The timeline for the doubling of confirmed cases has shown a steady increase.

• Active cases: Active cases currently stand at 1.9 lakh and the doubling of such cases followed trends similar to confirmed cases, but on a higher side.

• Recovery cases: As of 25th June, a total of 2.9 lakh COVID positives have recovered, resulting in a nationwide recovery rate of 58%.

• Deaths: COVID-19-related deaths stood at 9 before 25th March, which has increased to more than 15,300 as on 25th June. Deaths are now doubling in about 20 days *vs*. the previous average of 9-18 days in the second half April and month of May and June. The current mortality rate based on the number of closed cases (recovered + death) is 5.1%.

• The daily new confirmed cases of COVID-19 is still at its increasing trajectory with the highest, till date, being 18,000 plus.

• State-wise trend: Maharashtra remains the worst-affected state (confirmed cases have crossed the 147,000 mark), followed by Delhi, Tamil Nadu, Gujarat, Uttar Pradesh, Rajasthan, West Bengal, Madhya Pradesh, Haryana and Karnataka. They contribute to over 85% of the total COVID-19 confirmed cases. On a combined basis, these ten states contribute about 70% to India's total GDP.

With the nationwide lockdown now being extended to a total of 98 days and fiscal stimulus is proving to be inadequate, the economy is likely to suffer a deep recession in FY21. Although we have begun to see some form of cautious and calibrated opening up of economic activity in some regions/areas starting June 2020, return to normalcy or near normalcy depends on how quickly the COVID curve flattens out.

Over the past couple of years, the GDP growth rate has been falling, 2018 (6.8%), 2019 (4.2%) and was expected to be 4.6% percent in 2020. However, due to the imposed lockdown recent conservative estimates place the growth rate to be in negative territory (for the first time since 1979).

Due to the extremely contagious nature of the COVID-19 virus, the Government of India has had to implement several restrictions to curb the outbreak. The brunt of the economic consequences of the restrictions has been faced majorly by the aviation, tourism and hospitality sector. Nevertheless, many other small/medium/large services are facing the economic consequences that include real estate, constructions, textiles, passenger/commercial vehicles, poultry/meat, etc.

According to Willis Towers Watson India COVID-19 Readiness (Survey Results and Key Insights April 2020), the following gets highlighted:

• 57% of organisations have indicated that there will be a moderate to large negative impact on their business in the next 6 months

• 46% of organisations have indicated that there will be a moderate to large negative impact on their business in the next 12 months

• 19% of organisations have indicated that there will be a moderate to large negative impact on their business in the next 12-24 months

• 5% of organisations responded that there will be a positive business impact within the next 12 to 24 months

The economic activity of the country is gauged by the retail payments activity, be it the ATM usage for cash, or other digital means of payment. To see the impact of COVID-19 on the payment systems during the past three months, we use data provided by the Reserve Bank of India (RBI) and the National Payments Corporation of India (NPCI).

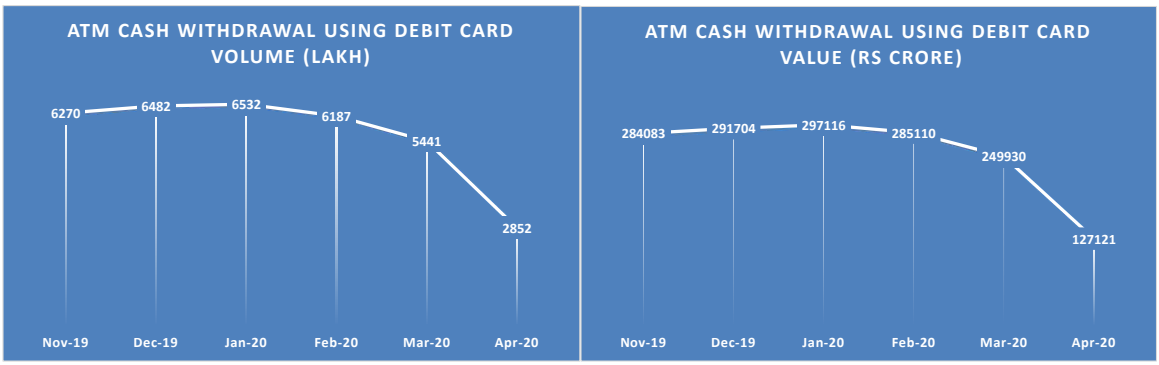
With the lockdown allowing mostly the essential sectoRs. to continue to function, the spending of disposable income is expected to reduce drastically. The cash withdrawals at ATMs is a prominent measure of the general retail economic activity. The inactivity in the economy and its extent in terms of requirement of cash is depicted in Section 2, using the ATM cash withdrawal data. In Section 3, we show the impact of lockdown on other digital payment modes such as Debit/Credit cards, UPI, IMPS, NEFT and RTGS. We present the extent to which COVID-19 and the subsequent nationwide lockdown has impacted the financial transactions in the country. The sheer slowdown of the economy gets depicted by the drastic reduction of retail payment activities in the country. Finally, in Section 4 we give some concluding remarks.

2. Impact of COVID-19 Related Lockdown on ATM Usage

We study the impact of lockdown on cash withdraws at ATMs. ATM transactions are an important indicator of the day-to-day economic activity. The ATM cash withdrawal transactions constitute the on-us and the off-us transactions. In case of on-us transactions, the ATM and the debit card that is used for cash withdrawal are of the same bank. While for the off-us transactions, the debit card of a bank is used in an ATM of a different bank. Such off-us transactions are routed through a switch, called the National Financial Switch (NFS), maintained by NPCI. We first present trends on cash withdrawals at ATM for the combined on-us and off-us transactions. This is followed by studying the trends for off-us transactions ATM only.

2.1. ATM usage for combined on-us and off-us transactions

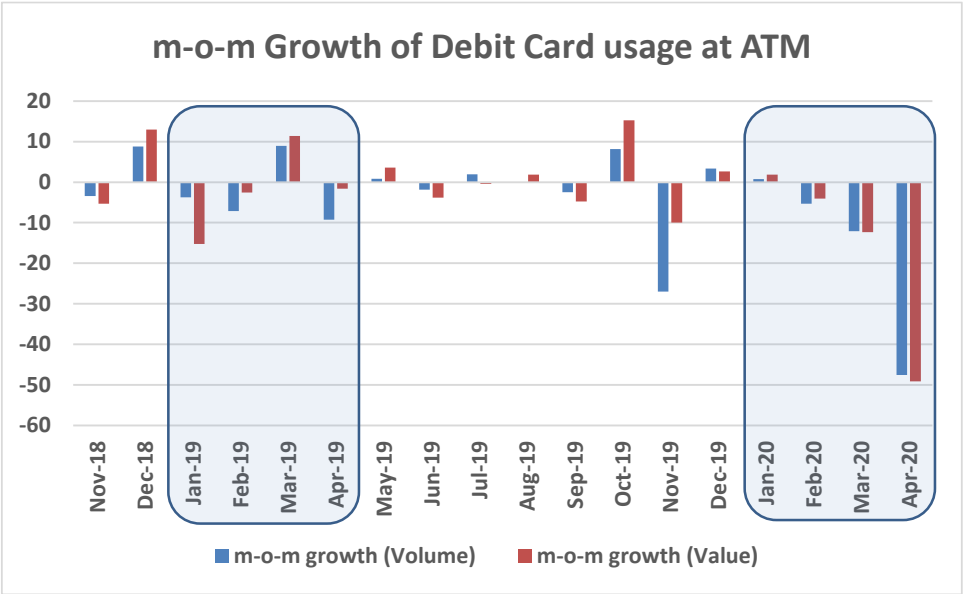
We primarily look at the volume and value data of ATM transactions for the period November 2019 through April 2020. Chart 1 shows that in April 2020, there had been Rs. 1.27 lakh crore of ATM cash withdrawals unlike monthly average of Rs. 2.77 lakh crore during Q4 of FY20. Thus, there had been less cash withdrawal to the tune of over Rs. 1.5 lakh crore in April 20, over the average monthly withdrawal in Q4 of FY20.



Source: RBI data

Chart 1: Trend of cash withdrawal at ATM using debit card

The impact of lockdown on ATM cash withdrawals are better judged based on growth curves. Charts 2 provides month-on-month (m-o-m) percentage growths for the period November 2018 through April 2020. The m-o-m growth of ATM transactions (in value terms) for March 2020 over March 2019 had been (-)24% while for April 2020 over April 2019 had been (-)48%.



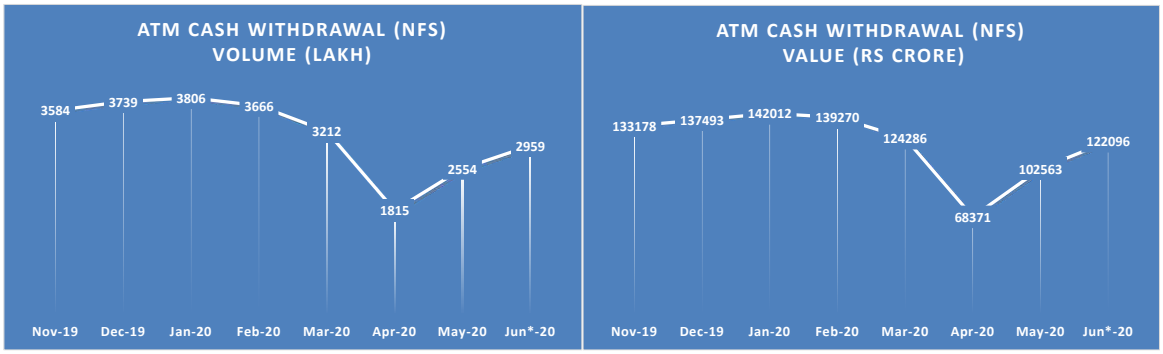
Source: RBI data and authors' computation

Chart 2: m-o-m growth of debit card usage at ATM

2.2. ATM usage for off-us transactions

Unlike RBI data for the combined on-us and off-us transactions, NPCI provides the monthly NFS data for off-us cash withdrawals at ATM. The latest data available is for the month of May 2020. Additionally, since June 2020, RBI has started disseminating daily data for cash

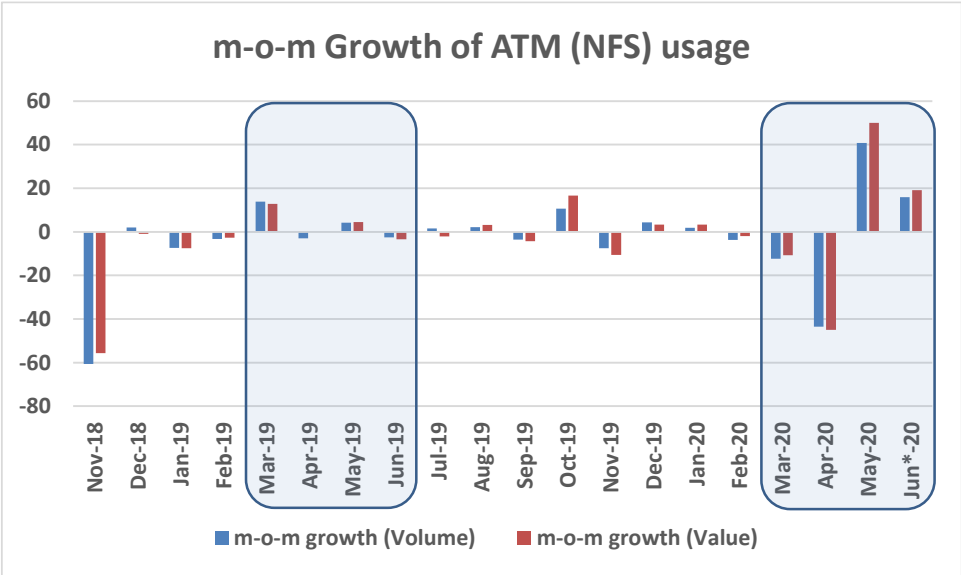
withdrawal using the NFS (through ATM). We use such daily data till June 23, 2020, to project the June figures. The off-us ATM cash withdrawals till June 23rd had been 2268.33 lakh (volume) and Rs. 93606.94 crore (volume). Therefore, based on proportions, for the 30 days of June 2020, an estimate of the cash withdrawal volume is 2268.33*30/23 = 2959 lakh and that the value is 93606.94*30/23 = Rs. 1.22 lakh crore.



Source: NPCI/RBI data and authors' computation

Chart 3: Trend of cash withdrawal at ATM for off-us transactions

Chart 3 shows that April 2020 showed a trough and thereafter things are slowly returning to normal with significant improvements in June 2020. During Q4 of FY20, the monthly average of off-us cash withdrawal at ATM had been Rs. 1.35 lakh crore, while April, May and June 2020 reflect off-us cash withdrawals of Rs. 0.68 lakh crore, Rs. 1.03 lakh crore and Rs. 1.22 lakh crore, respectively. This demonstrates the effect of some form of cautious and calibrated opening up of economic activity during the months of May and June.



Source: NPCI/RBI data and authors' computation

Chart 4: m-o-m growth of off-us ATM usage

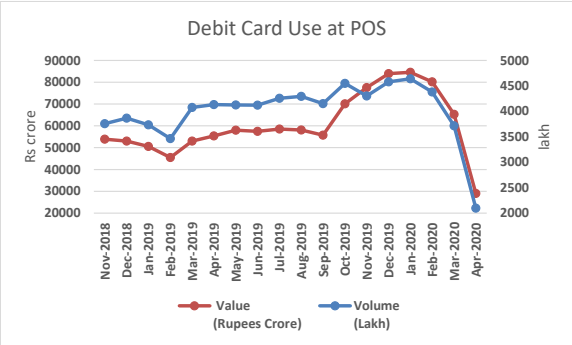
Charts 4 provides m-o-m growths for the period November 2018 through June 2020. The m-o-m growth of off-us ATM transactions (in value terms) for March 2020 over March 2019 had been (-)24%, for April 2020 over April 2019 had been (-)45%, for May 2020 over May 2019 had been 45%, while for June 2020 over June 2019 is estimated to be 22%.

3. Impact of COVID-19 Related Lockdown on Digital Payments

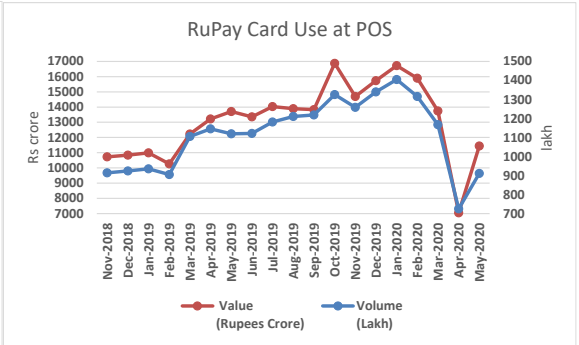
We study the impact of lockdown on some important digital payment modes such as Debit/Credit cards, UPI, IMPS, NEFT and RTGS.

3.1. Debit card usage at POS (includes e-Com)

Post demonetization, debit cards (primarily mastercard/VISA/RuPay debit cards) have seen a significant jump in usage at merchant POS, which includes e-Com mobile/computer-based online transactions. Primarily, e-Com constitutes e-commerce transactions and digital bill payments through ATMs, etc. for credit/debit cards, while for debit cards it additionally includes a card to card transfers. With lockdown in place, we see a significant drop in retail economic activity and this is seen clearly in Chart 5, where we present debit card and RuPay card transaction data for the period November 2018 through April 2010 (for RuPay debit card, we have the additional data for May 2020).



Source: RBI data



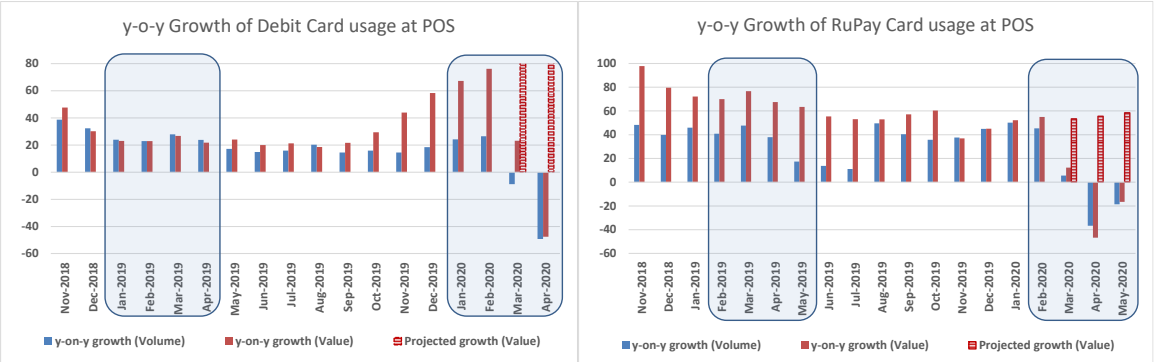
Source: NPCI data

Chart 5: Debit card and RuPay card transactions

The impact of lockdown on debit/RuPay card-based POS transactions are now judged based on growth patterns. The year-on-year (y-o-y) growth percentages for the period November 2018 through April/May 2020 are plotted in Chart 6. The impact of lockdown on POS transactions (in value terms) is now judged based on growth estimates in the absence of COVID-19 and the actual COVID-19 impacted figures. The estimates for March-April-May 2020 are based on a third-degree polynomial trend fitted from the growth (value) figures of November 2018 through February 2020. The period November 2018 through May 2020 are associated to

the variable x, taking values 1,2,…,16,17,18,19, respectively. Let y denote the y-o-y growth in transaction values at POS. The fitted curves for debit card and RuPay card are

$$\text{Debit Card:} \quad y = 47.938 - 6.934x + 0.251x^2 + 0.019x^3 \text{ with } R^2 = 0.939$$
$$\text{RuPay Card:} \quad y = 97.341 - 6.774x + 0.223x^2 + 0.001x^3 \text{ with } R^2 = 0.832$$



Source: RBI data and authors' computation *Source: NPCI data and authors' computation*

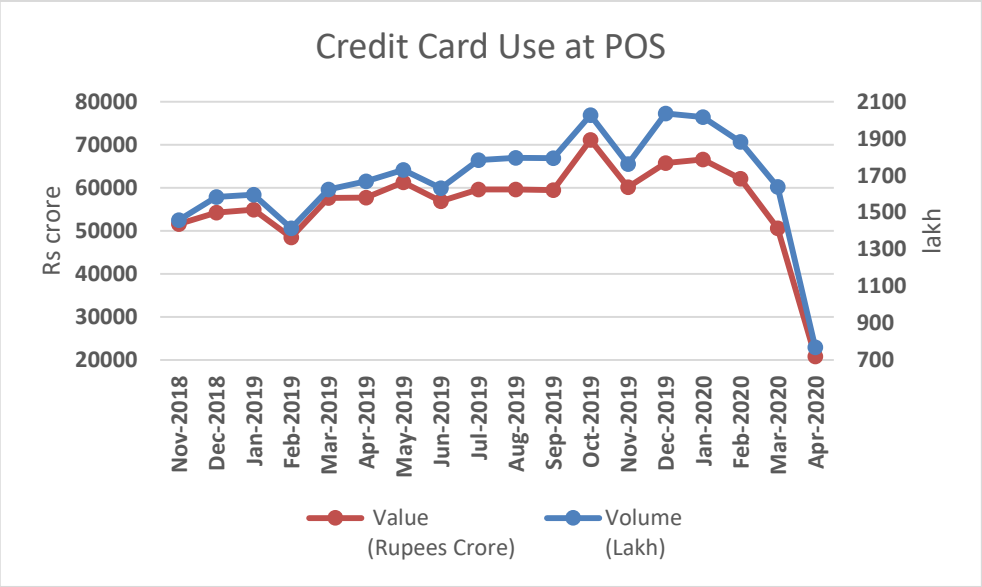
Chart 6: y-o-y growth patterns for debit/RuPay card usage at POS

The y-o-y growth of debit card transactions (in value terms) for March 2020 over its trend estimate had been (–)73%, while for April 2020, it had been (–)163%. For RuPay card the same is (–)41% and (–)102% for March 2020 and April 2020, respectively. Finally, the y-o-y growth of RuPay card transactions (in value terms) for May 2020 over its trend estimate had been (–)75%.

3.2. Credit card usage at POS (includes e-Com)

Prior to COVID outbreak, credit card (primarily mastercard/VISA credit cards) usage at POS (includes e-Com) have been increasing consistently. Lockdown induced a significant drop in retail economic activity and this is clearly reflected in Chart 7, which provides credit card transactions for the period November 2018 through April 2010.

As in case of debit cards, the impact of lockdown on credit card based POS transactions are now judged based on growth patterns. The y-o-y growth percentages for the period November 2018 through April 2020 are plotted in Chart 8, and the impact of lockdown on POS transactions is judged based on growth estimates in the absence of COVID-19 and the actual COVID-19 impacted figures. As earlier, the estimates for March-April 2020 are based on a third-degree polynomial trend fitted from the growth (value) figures of November 2018 through February 2020.

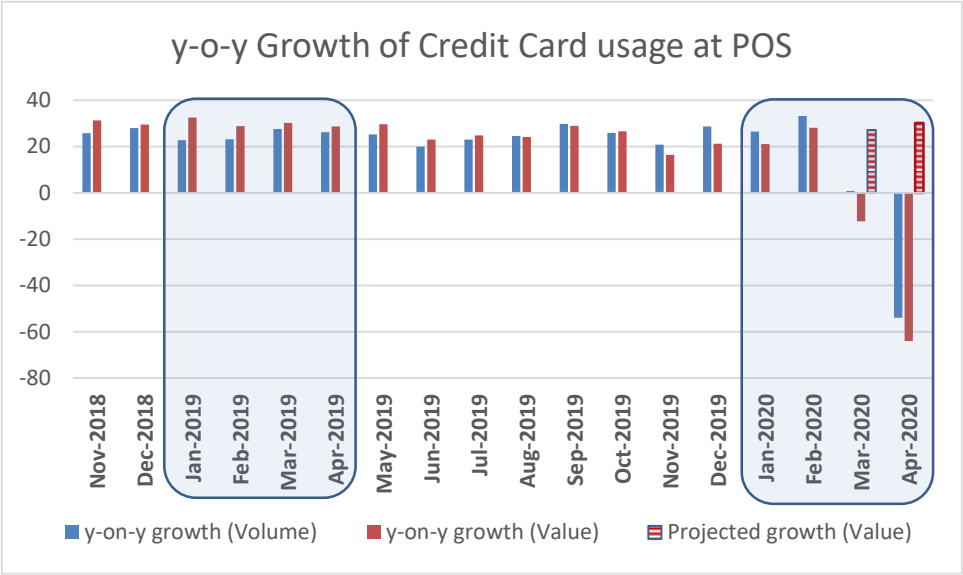


Source: RBI data

Chart 7: Credit card transactions

The period November 2018 through April 2020 are associated to the variable x, taking values 1,2,…,16,17,18, respectively. Let y denote the y-o-y growth in transaction values at POS. The fitted curve for credit card is

$$y = 29.264 + 1.499x - 0.351x^2 + 0.015x^3 \text{ with } R^2 = 0.565$$



Source: RBI data and authors' computation

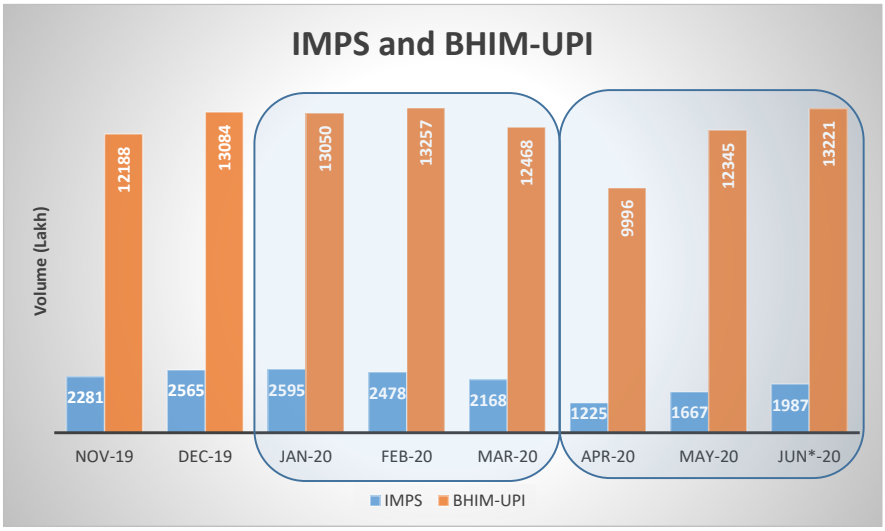
Chart 8: y-o-y growth patterns for credit card usage at POS

The y-o-y growth of credit card transactions (in value terms) for March 2020 over its trend estimate had been (–)39%, while for April 2020, it had been (–)94%. Since the third-degree polynomial fit has a relatively lower value of R^2, being conservative, we also look into the y-o-y growth (in value terms) for March 2020 over March 2019, which is (–)42% while for April 2020 over April 2019 is (–)93%.

3.3. IMPS and BHIM-UPI transactions

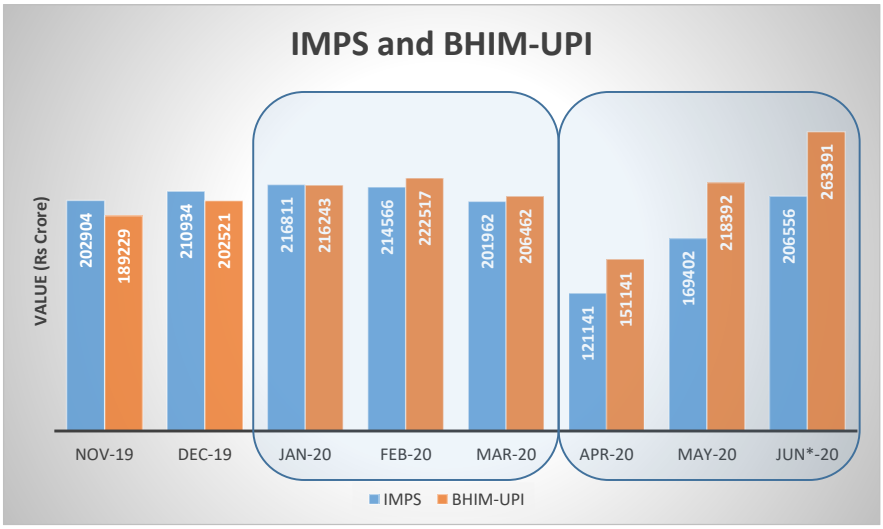
IMPS and BHIM-UPI based digital transactions allow real time account to account money transfers, be it person-to-person or person-to-merchant. Primarily based on mobile Apps, its ease of use has made it a well-accepted mode of retail payments and money transfer. NPCI has provided the transaction data till May 2020, while RBI, since June 2020, has started disseminating daily data for IMPS and BHIM-UPI transactions. Accordingly, as earlier, we use the daily data till June 23, 2020, to project the June figures. The estimated June 2020 IMPS volume and value are 1987 lakh and Rs. 2.07 lakh crore respectively. Similarly, for BHIM-UPI the volume and value estimates are 13221 lakh and Rs. 2.63 lakh crore respectively.

Tables 9 and 10 shows a marginal decrease in volume and values of such transactions in April 2020. However, omnipresence and proliferation of BHIM-UPI app has come very handy in the days of COVID (distancing from physical contacts) for day-to-day usage. This has led to a significant increase in BHIM-UPI transactions for the months of May and June 2020. IMPS has also shown recovery of transaction levels in May and June over April 2020.



Source: NPCI/RBI data and authors' computation

Chart 9: IMPS and BHIM-UPI transaction volumes



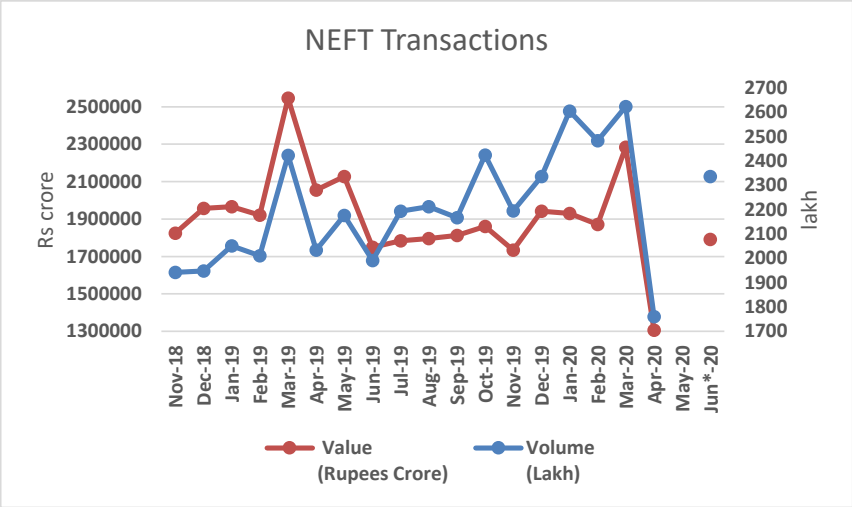
Source: NPCI/RBI data and authors' computation

Chart 10: IMPS and BHIM-UPI transaction values

3.4. NEFT transactions

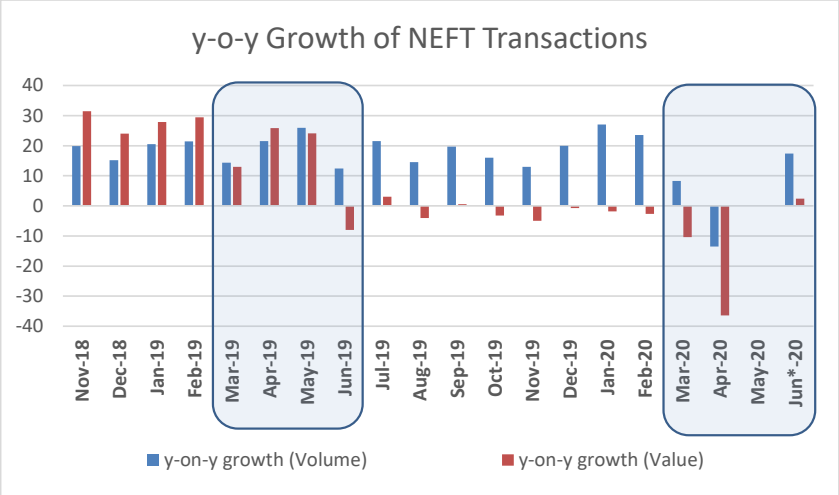
NEFT transactions are an important indicator of economic activity. In addition to the monthly data till April 2020, RBI's daily data for the period June 1-23, 2020 is used to project the June month's NEFT transactions. Following proportions, the estimated June 2020 volume is 2337 lakh and the value is Rs. 17.92 lakh crore. Trends in the NEFT transaction are provided in Chart 11 for the period November 2018 through April 2010 and June 2020. NEFT transactions usually increase in the month of March, being the financial year-end. The same is seen in March 2020 but to a lesser extent than in March 2019. However, for both volume and value, we see a significant drop in NEFT transactions in April 2020. In April 2020, there had been Rs. 13.06 lakh crore of NEFT transactions unlike a monthly average of Rs. 20.20 lakh crore during Q4 of FY20 (*i.e.*, a drop of over Rs. 7 lakh crore). However, in June 2020, we see a significant turnaround with increased NEFT transactions, being Rs. 17.92 lakh crore in value terms.

The impact of lockdown on NEFT transactions are better judged based on growth curves. In Chart 12, we provide y-o-y growths for the period November 2018 through April 2020. We see that the y-o-y growth of NEFT transactions (in value terms) for March 2020 over March 2019 had been (–)23%, for April 2020 over April 2019 had been (–)62%, while for June 2020 over June 2019 is estimated to be 10%. This indicates a revival of some economic activities, shown at least in terms of NEFT transactions.



Source: RBI data and authors' computation

Chart 11: NEFT transactions



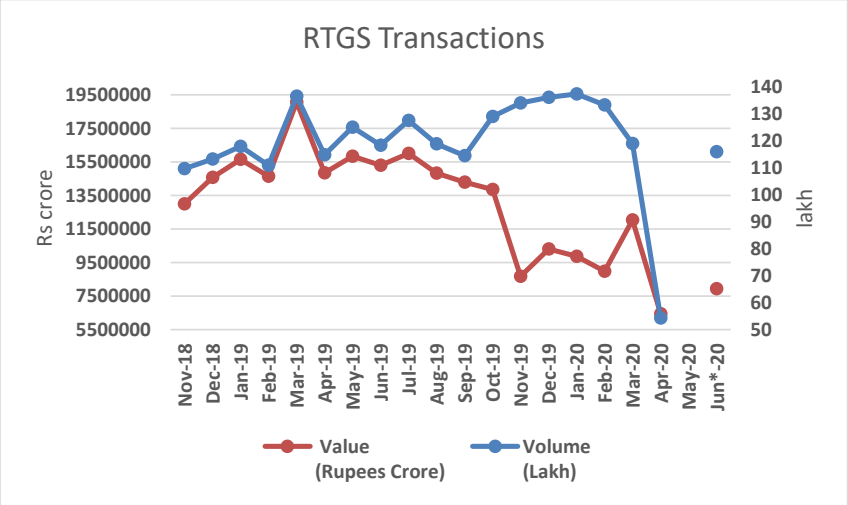
Source: RBI data and authors' computation

Chart 12: y-o-y growth of NEFT transactions

3.5. RTGS Transactions

RTGS transactions constitute RTGS customer transactions and RTGS interbank transactions. We focus on the RTGS transactions, being an indicator of economic activity. As earlier, the daily data for RTGS transactions till June 23, 2020 is used to project the June month's figure. The estimated June 2020 volume is 116 lakh and the value is Rs. 79.36 lakh crore for the RTGS transactions. Trends in the RTGS transaction for the period November 2018 through April 2020 and June 2020 is provided in Chart 13. For both volume and value, we see a significant drop in RTGS transactions in April 2020. In April 2020, there had been Rs. 64.44

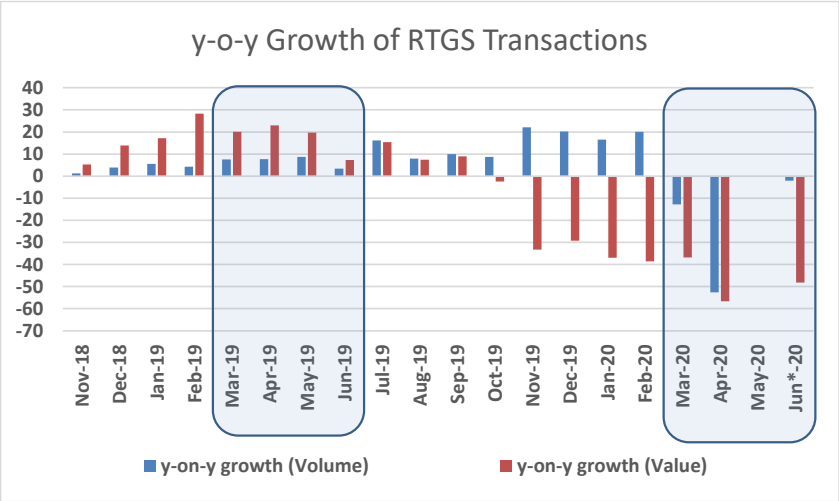
lakh crore of RTGS transactions unlike a monthly average of Rs. 103.06 lakh crore during Q4 of FY20 (*i.e.*, a drop of Rs. 38.62 lakh crore). However, in June 2020, we see a recovery with increased RTGS transactions, being Rs. 79.36 lakh crore in value terms.



Source: RBI data and authors' computation

Chart 13: RTGS transactions

In Chart 14, we provide y-o-y growths for the period November 2018 through April 2020 and that of June 2020. The y-o-y growth of RTGS transactions (in value terms) for March 2020 over March 2019 had been (–)57%, for April 2020 over April 2019 had been (–)80%, while for June 2020 over June 2019 is estimated to be (–)55%. One may note that such contracted growth are consistent and persisted since November 2019. It is remarked that from November 2019, RBI had introduced a new format for data dissemination.



Source: RBI data and authors' computation

Chart 14: y-o-y growth of RTGS transactions

4. Concluding Remarks

In India, COVID-19 cases started increasing exponentially from mid-March 2020. The subsequent lockdowns and thereafter slow unlocking had been witnessed during April-May-June 2020. Based on transaction data during Q4 of FY20 and Q1 of FY21, we relate the economic impact of COVID induced lockdown and subsequent relaxations in lockdown.

Although the economy is likely to suffer in FY21, we have begun to see some form of cautious and calibrated opening up of economic activity, as we see payment transactions picking up starting June 2020. This is clearly reflected in the country's ATM usage for cash withdrawal and a few other forms of digital payments. However, return to normalcy or near normalcy would depend on how quickly the COVID curve flattens out. The positive effect of COVID can be seen in form of increased levels of BHIM-UPI usage, as more and more people learn to use this mobile app-based, easy to use, digital mode of payment. Moreover, money transaction using BHIM-UPI does not cost the users of this payment mode.

Acknowledgement

Authors are thankful to a referee who provided some valuable comments that helped in improving the presentation of the paper.

References

Worldometer Coronavirus. https://www.worldometers.info/coronavirus/country/india/
myGOV #IndiaFightsCorona COVID-19. https://www.mygov.in/covid-19
RBI publishes daily data of select payment systems. Press Release. June 4, 2020. https://www.rbi.org.in/scripts/FS_PressRelease.aspx?prid=49901&fn=9
Reserve Bank of India Bulletin - June 2020. RBI. June 10, 2020. https://www.rbi.org.in/Scripts/BS_ViewBulletin.aspx
Retail Payments Statistics on NPCI Platforms. https://www.npci.org.in/statistics
Willis Towers Watson India COVID-19 Readiness- Survey Results and Key Insights April 2020. https://www.willistowerswatson.com/en-IN/Insights/2020/04/india-covid-19-readiness-pulse-survey-thank-you

Statistics and Applications ISSN 2452-7395 (online)
Volume 18, No. 1, 2020 (New Series), pp 253–268

Identification of Geographic Specific SARS-Cov-2 Mutations by Random Forest Classification and Variable Selection Methods

Manoj Kandpal and Ramana V Davuluri[1]
Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.
[1]*Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.*

Received: 06 June 2020; Revised: 28 June 2020; Accepted: 30 June 2020

Abstract

RNA viral genomes have very high mutations rates. As infection spreads in the host populations, different viral lineages emerge acquiring independent mutations that can lead to varied infection and death rates in different parts of the world. By application of Random Forest classification and feature selection methods, we developed an analysis pipeline for identification of geographic specific mutations and classification of different viral lineages, focusing on the missense-variants that alter the function of the encoded proteins. We applied the pipeline on publicly available SARS-CoV-2 datasets and demonstrated that the analysis pipeline accurately identified country or region-specific viral lineages and specific mutations that discriminate different lineages. The results presented here can help designing country-specific diagnostic strategies and prioritizing the mutations for functional interpretation and experimental validations.

Key words: Random forest; Feature selection; Classification; SARS-CoV-2; Coronavirus.

1. Introduction

In December 2019, researchers identified a novel coronavirus that first infected and caused coronavirus disease (COVID) in patients in Wuhan, China (Lu et al. 2020b). The virus, initially named as 2019-nCoV, was officially renamed as SARS-CoV-2 by the International Committee on Taxonomy of Viruses to indicate that it was very closely related to the SARS (Severe Acute Respiratory Syndrome Coronavirus). It infected 6,265,496 confirmed cases and caused 375,526 deaths globally as of June 1, 2020 (https://coronavirus.jhu.edu/). SARS-CoV-2 is an enveloped single-stranded RNA virus. It infects a human host by breaking into the host's cell and acquires mutations during replications in the cell. As it spreads from person to person, the accumulated mutations in the viral genomes can lead to different viral lineages. One particular type of mutations, called missense mutations, alter the amino acids encoded by the RNA sequences. For example, some missense mutations alter a protein to give growth advantage for the virus – allowing virus entry into a host cell, and others can lead to changes in the target region of a drug or antibody that acts against the virus protein (Zhao et al. 2018; Holland et al. 2020). Therefore, computational methods to prioritize specific mutations from a large set of passenger mutations and classify different lineages is of great importance for the ongoing COVID research.

Corresponding Author: Ramana V. Davuluri
Email: Ramana.Davuluri@stonybrookmedicine.edu

We developed a computational pipeline for constructing a tree based Random Forest classifier to discriminate SARS-CoV-2 lineages from different geographic regions and identify important mutations, using the rich source of existing mutational profiles and associated genomic annotations and geographic information. Here, we attempt to classify viral lineages from four geographic locations – 1) USA-New York; 2) China; 3) Europe-Spain and Italy; and 4) India. We prepared a dataset by processing publicly available mutational profiles that were curated by analyzing 20,746 SARS-CoV-2 genome sequences. These genome sequences were sequenced from infected patient samples in different countries. We systematically trained and evaluated Random Forest (RF) classifiers on subset of this dataset, using both cross validation and testing on independent test set, and selected the best performing RF classifier for the final algorithm.

2. Data Description

Working around the world in different countries, teams of scientists are racing to understand the virus's genetic sequences, develop treatments and vaccine candidates, and to accurately forecast future outbreaks. In this unprecedented effort, more than 30,000 SARS-CoV-2 genomes have been sequenced and submitted to public data repositories since the outbreak in December, 2019 (Colson et al. 2020; Lu et al. 2020a; Yadav et al. 2020). By aligning these genomic sequences to a reference SARS-CoV-2 genome, numerous mutation sites are identified and stored in public databases. We downloaded the following data files from 2019 Novel Coronavirus Resource at China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/release_genome).

1. **VCF file** from https://bigd.big.ac.cn/ncov/variation/statistics?lang=en. File name "2019-nCoV_total.vcf". VCF (Variant Call Format) file contains meta-information lines, header lines, and then data lines (rows) each containing information about a mutation in the genome. The columns contain genotype information on samples for each position. The downloaded file contains 10,261 non-header rows (each corresponding to specific mutation in the genome) and 20,755 columns, of which first 9 columns are mutation information and the rest of the columns contain genotype information for 20,746 virus samples. Supplement Table 1 provides an example of top-ranking mutations, and their genotype information for two samples (columns 10 and 11).

2. **Variant Annotation file** from https://bigd.big.ac.cn/ncov/variation/annotation. File name "Variation Annotation.xls". This file contains the genomic annotations of the identified mutations, such as a) genomic position, b) gene name or region in which the mutation is located, c) Number of viruses with the mutation, d) Annotation type – missense, synonymous or intergenic variant, etc., e) Mutation type – SNP, insertion or deletion, etc., and f) Protein position and amino acid change, etc.

In particular, we focused our analysis of missense variants – those genomic variants that alter the encoded amino acid sequences; because study of proteins is key to understanding the viral spread and successful development of vaccines and neutralizing antibodies. We choose four countries/regions based on the wide variations in infection and death rates. The four regions are –1) USA-NY, the epicenter in the United States; 2) China, where the pandemic originated; 3) Spain and Italy, two epicenters in Europe; and 4) India, where the world's biggest coronavirus lockdown measures were strictly implemented.

3. Methodology and Computational Framework

This Section describes the methodology and computational processing used in this analysis. We applied advanced tree-based ensemble learning algorithm – Random Forests (Breiman 2001) for building the classification model for discriminating the virus lineages of four geographical locations. Since RF results its output in a 'black-box' model, we applied Classification and Regression Tress (CART) methodology on selected feature sets due to its key advantage in terms of interpretability (James et al. 2013).

Random forest: Fandom forest is a collection of tree structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$, where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x*. For a given ensemble of classifiers, with the training set drawn at random from the distribution of the random vector *Y,X*, the margin function is defined as,

$$mg(\boldsymbol{X}, Y) = av_k I(h_k(\boldsymbol{X}) = Y) - max_{j \neq Y}\ av_k I(h_k(\boldsymbol{X}) = j),$$

where $I(.)$ is the indicator function. The confidence in the classification is directly proportional to the margin, as the margin measures the extent to which the average number of votes at $\boldsymbol{X,Y}$ for the right class exceeds the average vote for any other class. Each tree is constructed using a different bootstrap sample from the original data where about one-third (33%) of the cases are left out of the bootstrap sample and not used in the construction of the *k-th* tree. These left out samples, usually called "out-of-bag" data, is used to get a running unbiased estimate of the classification error as trees are added to the forest. Thus, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, let *j* to be the class that got most of the votes every time case/sample *n* was out-of-bag. The proportion of times that *j* is not equal to the true class of *n* averaged over all cases is the out-of-bag error estimate. At each node, only a subset of the possible predictors are used, primarily for reducing the correlation between trees and improving the accuracy of classification (Breiman and Cutler 2001).

CART and rpart: CART is a tree-based algorithm that works via recursive partitioning of the training set in order to obtain multiple subsets that are closest (or as homogeneous as possible) to a given target class (Breiman 1984). At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable.

We describe the computational pipeline (Supplementary Figure 1) in the following two steps:

Step 1 (*Data processing step*): We wrote a Perl program to scan the two downloaded files for extracting the geographic information (from sample IDs) and selecting only the missense mutations with minor allele frequency greater than a certain cut-off. For each mutation site, major allele is the one with the highest count and minor allele is the one with the second highest count. In other words, *Minor Allele Frequency* (MAF) is the frequency at which the second most common allele occurs in a given population. We included only those missense variants with minor allele count greater than 10. This step prepares the data in tab-separated tabular form for statistical analysis in R programing environment.

Step 2 (*Variable Selection and Classification Steps*): Prior to building the final classification model using RF, we applied a variable selection algorithm (Diaz-Uriarte 2007) to select a

small set of important non-redundant mutations. Feature selection was done using an RF based algorithm, that uses hybrid approach of selecting the virus mutations (predictor variables) based on importance spectrum (similar to scree plot) and backward variable elimination (for the selection of small sets of non-redundant variables) by changing parameters related to trees and iteration. Using 80,000 initial trees and a fractional variable drop of 10%, we finally selected 42 mutations as the most discriminative variables between the four geographic regions (classes) and created the RF classifier for viral lineage prediction with high accuracy. The accuracy of the classifier was determined by RF's cross-validation analysis (out-of-bag approach) and testing on independently set aside dataset. We split the data into 80:20 ratio and classification model generated from the training set (80%) was applied to the test set (20%) to strengthen the accuracy claims. We then developed CART model for better visualization of mutation differences among classes. For developing CART model, we used the important features as selected by RF. Analyses were done using custom scripts in R and libraries including, *randomForest* and *varSelRF*. Recursive Partitioning and Regression Trees (rpart), an R implementation of the CART algorithm, is used for developing CART model in this study. *randomForest* library provides an R interface to the Fortran programs (available at https://www.stat.berkeley.edu/~breiman/RandomForests/). *varSelRF* library is used for selecting a small set of mutations while preserving Random Forest classification accuracy.

4. Results

We downloaded mutation profile of 10,261 mutations in 20,746 SARS-CoV-2 samples as described in Section 3. After selecting only missense variants that show variation in at least 10 samples, we retained 588 missense mutations. Further, we selected a total of 2,927 samples that correspond to four geographic locations. Data was divided into training (2,341) and testing (586) sets based on number of samples, maintaining the class ratios. In Table 1, we list the top 50 mutations observed among all the sequenced viral samples. USA-NY samples showed highest mutation rate, suggesting that coronavirus was probably circulating undetected in the NY area for quite some time. Additionally, we found that the top four most common mutations showed significantly higher mutation rates in USA-NY samples than the other geographical regions, including rest of the world samples. For example, two of the mutations – one at genomic location 23403 (A mutated to G) and the other at 25563 (G mutated to T) – alter amino acids QHD43416.1:D614G and QHD43417.1:Q57H, respectively, in the S and ORF3a proteins of the SARS-Cov-2 virus. For the virus to break into a human cell (host cell), the S protein of the virus binds to ACE2 (angiotensin converting enzyme 2) protein on the human cell surface. The D614G mutation in S protein might change the protein structure so that it binds to the target enzyme (ACE2) with different affinity than the other lineage proteins (Amin et al. 2020). Similarly, the Q57H mutation in ORF3a protein might change important functional domains linked to virulence, infectivity, ion channel formation, and virus release (Issa et al. 2020). We speculate that this may partly explain why the infection rate is much higher in USA-NY area than other geographic locations.

Next, we built a four-class classification model for discriminating the SARS-CoV-2 samples grouped according to the four geographic locations, by specifying the country/region as factor variable $\boldsymbol{y}$ and mutation profile as predictor variable matrix $\boldsymbol{X}$ (a binary matrix, where 1 and 0 denote presence and absence of the mutation respectively). The accuracy of the finally selected classifier based on cross-validation analysis (out-of-bag approach) is 85%. Table 2 shows the confusion matrix for the final model and Figure 1 shows corresponding AUC. The robustness of developed model was then confirmed on the test data with high accuracy. Table 3 shows evaluation metrics of the model on training and test data. USA-NY and China classes

have shown the best classification accuracy, followed by Italy-Spain class. The least accuracy was observed for India class, which could be due to small sample size of that class. However, we speculate that the misclassification of 26.8% and 10.3% of India class into Italy-Spain and USA classes, respectively, might be due the viral samples from the infected patients who traveled back to India from those geographic regions and not due to local spread of the virus.

Class 4 (India) and Class 2 (China) showed better Specificity and PPV and lower FDR than the other two classes. DOR, ratio of the odds of positivity was also higher for Indian and Chinese lineages than the other two geographical regions. DOR depends significantly on the sensitivity and specificity of a test. A test with high specificity and sensitivity with low rate of false positives and false negatives has high DOR. A diagnostic odds ratio of 1 is similar to an AUC of 0.5 and represents an uninformative test. Higher values for both reflect usefulness of the classification model. Youden index analysis is useful in finding the optimal cutoff value. The value provides the best tradeoff between sensitivity and specificity and is highest for Class 2. F-score, which combines precision with recall is a good measure of goodness of model and shows high value for the current model predictions. Similarly, greater the value of positive likelihood ratio (PLR) for a particular test, the more likely a positive test result be a true positive. A good classifier should have high PLR and low Negative Likelihood Ratio (NLR). Matthews correlation coefficient (MCC), initially developed for binary classifier, considers true and false positives and negatives and is regarded as a balanced measure, which can be used even if the classes are of very different sizes. Optimized precision, a novel metrics used to negate the detrimental effects of using Precision (P) for performance evaluation of unbalanced data, also shows high values for all classes. The evaluation results from training model clearly indicate that the algorithm effectively distinguished the samples from the four regions based on the mutation profile of SARS-CoV-2. Results from testing the classifiers on test data agree with the cross-validation results and support the accurate predictive performance of the classification model.

The results from CART analysis are shown in Table 4 and Figure 2. Although, the CART model is not as accurate as the random forest model, it provided a better visualization of the associations between viral samples/lineages in different geographic regions and the mutation patterns. For example, the final CART model (Figure 3) found that the presence of mutations at genomic locations 1059_C and 17747_C, but not at genomic locations 14408_C, 13730_C, 9477_T and 11083_G classified majority of USA-NY samples from the rest in one branch of the tree. Similarly, mutation at genomic location 13730_C, but not at 1059_C, 14408_C classified majority of India samples in one branch of the tree. Interestingly, most of the missense mutations in the model alter the amino acids encoded by gene *orf1ab*. This gene encodes 16 non-structural poly-proteins (Nsp1-Nsp16) of viral RNA synthesis complex (Kirchdoerfer and Ward 2019). We expect that these results will help prioritization of select mutations, and study of their effect on SARS-Cov-2 and Human protein-protein interactions through focused experimental validations.

Table 1: Top 50 Mutations in the four geographic regions and the rest of the world. Column 2 – Genomic location of the mutation and the reference allele; Column 3 – Gene location of the missense mutation; Columns 4-8 – Percentage of samples mutation observed in each geographic location; Column 9 – Range (difference of highest and lowest mutation percentages among the four geographic locations).

			Percentage of mutated samples/isolates in (Total number of samples in the parentheses)					
Mutation rank	Genomic location of the mutation (Ref Allele)	Gene name	Rest of the World (17816)	USA-NY (1243)	China (656)	Italy, Spain (534)	India (494)	Range
1	23403 (A)	S	74.45	**93.24**	5.34	65.92	62.35	87.91
2	14408 (C)	orf1ab	74.25	**93.24**	3.66	64.42	62.35	89.58
3	25563 (G)	ORF3a	28.37	**86.32**	0.91	1.69	23.08	85.41
4	1059 (C)	orf1ab	21.57	**79.00**	0.91	1.31	0.61	78.40
5	28881 (G)	N	24.36	4.02	1.52	17.04	12.15	15.52
6	28882 (G)	N	24.33	4.10	1.22	16.67	12.15	15.45
7	28883 (G)	N	24.29	4.10	1.22	16.85	12.15	15.63
8	28144 (T)	ORF8	10.22	2.65	31.10	27.53	4.25	28.44
9	11083 (G)	orf1ab	9.65	4.67	5.95	6.55	**34.01**	29.34
10	17858 (A)	orf1ab	6.93	1.21	0.00	0.00	0.20	1.21
11	17747 (C)	orf1ab	6.87	1.21	0.00	0.00	0.20	1.21
12	26144 (G)	ORF3a	5.96	3.78	5.03	4.87	0.61	4.42
13	27964 (C)	ORF8	3.09	0.56	0.00	0.00	0.00	0.56
14	2558 (C)	orf1ab	2.23	0.48	0.15	0.00	0.61	0.61
15	28854 (C)	N	1.82	0.00	1.37	0.00	**14.17**	14.17
16	13730 (C)	orf1ab	1.36	0.08	0.46	0.19	**30.16**	30.08
17	28311 (C)	N	1.40	0.00	0.46	0.19	**28.95**	28.95
18	27046 (C)	M	2.09	0.16	0.00	1.12	0.00	1.12
19	2480 (A)	orf1ab	2.07	0.48	0.15	0.00	0.40	0.48
20	6312 (C)	orf1ab	1.25	0.00	0.46	0.19	**28.14**	28.14
21	11916 (C)	orf1ab	0.74	17.62	0.00	0.00	0.00	17.62
22	10097 (G)	orf1ab	1.96	0.08	0.00	0.00	0.00	0.08
23	25429 (G)	ORF3a	1.80	0.08	0.00	0.19	0.20	0.20
24	28077 (G)	ORF8	1.55	0.72	1.52	0.00	0.81	1.52
25	1440 (G)	orf1ab	1.48	0.00	0.00	0.00	0.00	0.00
26	2891 (G)	orf1ab	1.46	0.00	0.00	0.00	0.00	0.00
27	26530 (A)	M	1.32	0.24	0.00	1.12	0.61	1.12
28	18998 (C)	orf1ab	0.34	14.32	0.00	0.00	0.00	14.32
29	3177 (C)	orf1ab	1.27	0.40	0.15	0.00	0.00	0.40
30	490 (T)	orf1ab	1.22	0.40	0.15	0.00	0.00	0.40
31	28863 (C)	N	0.63	0.24	0.15	19.66	0.20	19.51
32	1397 (G)	orf1ab	0.98	0.16	2.90	0.94	3.85	3.69
33	9477 (T)	orf1ab	0.59	0.24	0.15	20.04	0.20	19.89
34	18736 (T)	orf1ab	1.19	0.40	0.00	0.00	0.00	0.40
35	25979 (G)	ORF3a	0.61	0.16	0.15	19.10	0.20	18.95
36	11109 (C)	orf1ab	1.20	0.00	0.00	0.00	0.00	0.00

37	6310 (C)	orf1ab	0.82	0.00	0.00	0.00	9.51	9.51
38	4002 (C)	orf1ab	1.03	0.08	0.00	0.00	0.00	0.08
39	28836 (C)	N	0.95	0.16	0.00	0.00	0.00	0.16
40	13862 (C)	orf1ab	0.94	0.00	0.00	0.00	0.40	0.40
41	24368 (G)	S	0.88	0.00	0.00	0.00	0.00	0.00
42	21575 (C)	S	0.69	1.13	0.00	0.37	0.61	1.13
43	28878 (G)	N	0.55	0.72	1.22	0.75	3.85	3.12
44	16289 (C)	orf1ab	0.72	0.16	0.00	0.19	0.00	0.19
45	25688 (C)	ORF3a	0.68	0.32	0.00	0.19	0.00	0.32
46	10323 (A)	orf1ab	0.68	0.08	0.30	0.19	0.00	0.30
47	10798 (C)	orf1ab	0.70	0.00	0.00	0.00	0.00	0.00
48	25350 (C)	S	0.66	0.00	0.00	0.00	0.20	0.20
49	28580 (G)	N	0.66	0.00	0.00	0.19	0.00	0.19
50	1302 (C)	orf1ab	0.66	0.00	0.00	0.00	0.00	0.00

While the Random Forest classification method used here is considered as a "black box" method, with no interpretable classification model, the method provides useful information, such as variable importance. One of the measures of variable importance in Random Forest method is the mean decrease in accuracy, calculated using the out-of-bag sample. The difference between the prediction accuracy on the untouched out-of-bag sample and that on the out-of-bag sample permuted on one predictor variable is averaged over all trees in the forest and normalized by the standard error. This gives the mean decrease in accuracy of that particular predictor variable which has been permuted. Figure 3 shows the list of feature variables ranked according to mean decrease in accuracy of classification. It is interesting to note that the mutations in genes orf1ab, ORF3A and S genes rank among the most discriminative variables from mean decrease in accuracy graph (Figure 3).

Table 2: Confusion matrix of the prediction results of fitted model on Training set

		True Class – Number of real samples in each class				
		USA-NY (Class 1)	China (Class 2)	Italy, Spain (Class 3)	India (Class 4)	Total
Predicted Class – Number of predicted samples in each class	USA-NY	907 (91.25%)	15 (1.51%)	72 (7.24%)	0 (0.00%)	994
	China	9 (1.71%)	497 (94.67%)	16 (3.05%)	3 (0.57%)	525
	Italy, Spain	20 (4.68%)	42 (9.84%)	362 (84.78%)	3 (0.70%)	427
	India	41 (10.38%)	23 (5.82%)	106 (26.84%)	225 (56.96%)	395
	Total	977	577	556	231	

Table 3: Evaluation metrics of the four-class classification model based on cross-validation and independent test data. Bold numbers represent best performance among classes in training and test set. NPV – Negative Predictive Value; FDR – False Detection Rate; FNR – False Negative Rate; DRO – Diagnostic Odds Ratio; PLR – Positive Likelihood Ratio; NLR – Negative Likelihood Ratio; MCC – Matthews correlation coefficient.

Metric	Cross validation				Test data			
	USA (Class 1)	China (Class 2)	Europe (Class 3)	India (Class 4)	USA (Class 1)	China (Class 2)	Europe (Class 3)	India (Class 4)
Balanced accuracy	0.93	**0.95**	0.87	0.78	0.91	**0.94**	0.83	0.82
Sensitivity or Recall	0.91	**0.95**	0.85	0.57	0.91	**0.95**	0.74	0.65
Specificity	0.94	0.95	0.89	**1.00**	0.92	0.94	0.91	**0.99**
PPV or Precision	0.93	0.86	0.65	**0.97**	0.90	0.83	0.66	**0.94**
NPV	0.93	**0.98**	0.96	0.91	0.92	**0.98**	0.94	0.92
FDR	0.07	0.14	0.35	**0.03**	0.10	0.17	0.34	**0.06**
FNR	0.09	**0.05**	0.15	0.43	0.09	**0.05**	0.26	0.35
False Omi-ssion Rate	0.07	**0.02**	0.04	0.09	0.08	**0.02**	0.06	0.08
False Positive Rate	0.06	0.05	0.11	**0.00**	0.08	0.06	0.09	**0.01**
DRO	161.44	331.48	46.76	**389.56**	109.32	**261.46**	29.20	196.11
Youden's Index	0.85	**0.90**	0.74	0.57	0.83	**0.88**	0.65	0.64
Geometric Mean	0.93	**0.95**	0.87	0.75	0.91	**0.94**	0.82	0.80
F-score (beta 0.5)	**0.93**	0.88	0.68	0.85	**0.90**	0.85	0.68	0.86
F-score (beta 1)	**0.92**	0.90	0.74	0.72	**0.91**	0.89	0.70	0.77
F-score (beta 2)	0.92	**0.93**	0.80	0.62	**0.91**	0.92	0.72	0.69
PLR	15.04	18.63	7.97	**168.23**	11.01	14.92	8.38	**69.98**
NLR	0.09	**0.06**	0.17	0.43	0.10	**0.06**	0.29	0.36
MCC	0.85	**0.87**	0.67	0.71	0.82	**0.85**	0.62	0.74
Markedness	0.85	**0.87**	0.67	0.71	0.82	**0.85**	0.62	0.74
Optimization Precision	0.84	**0.85**	0.82	0.58	**0.84**	**0.84**	0.74	0.63

Table 4: Confusion matrix of the prediction results using CART on training set

		True Class – Number of real samples in each class				
		USA-NY (Class 1)	China (Class 2)	Italy, Spain (Class 3)	India (Class 4)	Total
Predicted Class – Number of predicted samples in each class	USA-NY	914 (91.95%)	9 (0.91%)	65 (6.54%)	6 (0.60%)	994
	China	10 (1.90%)	489 (93.14%)	16 (3.05%)	10 (1.90%)	525
	Italy, Spain	22 (5.15%)	38 (8.90%)	363 (85.01%)	4 (0.94%)	427
	India	41 (10.38%)	10 (2.53%)	115 (29.11%)	229 (57.97%)	395
	Total	987	546	559	249	

5. Conclusions

In theory, accumulation of mutations could make a virus more infectious or deadly, or vice versa, but the vast majority of mutations do not affect a virus's performance. While some mutations lead to more virulent and lethal strains, other mutations make the virus less infectious and less lethal in the populations. Computational methods that effectively integrate genomic profiles to identify and prioritize important genomic features and classify different groups of samples are valuable tools for Bioinformatics researchers. SARS-CoV-2 related research is rapidly evolving with numerous publications. Phylogenetic methods have been applied to SARS-CoV-2 genome sequences to construct the phylogenetic trees (clusters of closely related lineages) and predict future global hot spots of disease transmission and surge (Forster et al. 2020). Similarly, analysis pipelines are being developed for analysis of SARS-CoV-2 genomes to facilitate identification of novel mutations (Pachetti et al. 2020) and for functional annotations of mutations in specific gene regions, for example, nonsynonymous mutations in the ORF3a protein (Issa et al. 2020). Here, we have developed a complementary computational pipeline based on Random Forest based classification methods to identify a subset of missense mutations that can classify groups of virus lineages. It was previously reported, based on analysis of 220 genomic sequences, that the mutations located at positions 2891, 3036, 14408, 23403 and 28881 positions were predominantly observed in Europe, whereas those located at positions 17746, 17857 and 18060 were exclusively present in North America (Pachetti et al. 2020). However, we found that the top-ranking mutations located at positions 14408 and 23403 were most frequent in USA-NY samples than the rest of the geographical regions. We believe that this contradictory result could be due to much bigger sample size and small geographic regions in our analysis. Our findings suggest that the virus is evolving locally, and presence of small geographic region-specific strains that could be accurately classified by different mutational patterns.

Random Forest based algorithms have been successfully applied in various genomic analysis studies. For example, we have earlier used an integrative modeling approach that combines CART (Breiman 1984) and Random Forest to classify different estrogen receptor alpha responsive promoters (Cheng et al. 2006) and SMAD target promoters (Qin et al. 2009) with reasonably good classification accuracy and reduced instability (Qin et al. 2009). Although the main goal in classification is to build a model with minimal misclassification error in cross-validation, in these applications we are equally interested in identifying

biologically important features, such as genomic mutations or single nucleotide polymorphisms, for future experimental prioritization. The computational pipeline presented here will help the discovery of geographic specific SARS-CoV-2 mutations for further computational modeling and experimental validations and help in the interpretation of their functional effects.

Acknowledgements

This work was supported by the National Library of Medicine of the NIH [R01LM011297 to RD]. We thank the reviewer and editor for their suggestions and thoughtful comments, which substantially helped the revised version.

References

Amin, M., Sorour, M. K. and Kasry, A. (2020). Comparing the Binding Interactions in the Receptor Binding Domains of SARS-CoV-2 and SARS-CoV. *Journal of Physical Chemistry Letters*, **11**, 4897-4900. doi:10.1021/acs.jpclett.0c01064.

Breiman, L. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.

Breiman, L. and Cutler, A. (2001). Random Forests . https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Cheng, A. S., Jin, V. X., Fan, M., Smith, L. T. et al. (2006). Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Molecular Cell*, **21**, 393-404.

Colson, P., Lagier, J. C., Baudoin, J. P., Bou, Khalil J., La Scola, B. and Raoult, D. (2020). Ultrarapid diagnosis, microscope imaging, genome sequencing, and culture isolation of SARS-CoV-2. *European Journal of Clinical Microbiology and Infectious Diseases*, 1-3. doi:10.1007/s10096-020-03869-w.

Diaz-Uriarte, R. (2007). GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* **8**, 328.

Forster, P., Forster, L., Renfrew, C. and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences USA*, **117**, 9241-9243.

Holland, L. A., Kaelin, E. A., Maqsood, R., Estifanos, B., Wu, L. I., Varsani, A., Halden, R. U., Hogue, B. G., Scotch, M. and Lim, E. S. (2020). An 81 nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (Jan-Mar 2020). *Journal of Virology*. doi:10.1128/JVI.00711-20.

Issa, E., Merhi, G., Panossian, B., Salloum, T. and Tokajian, S. (2020). SARS-CoV-2 and ORF3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems*, **5**.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*: *With Applications in R*. Springer, New York.

Kirchdoerfer, R. N. and Ward, A. B. (2019). Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications*, **10**, 2342.

Lu, I. N., Muller, C. P. and He, F. Q. (2020a). Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies. *Virus Research*, **283**, 197963.

Lu, R., Zhao, X., Li, J., Niu, P. et al. (2020b). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, **395**, 565-574.

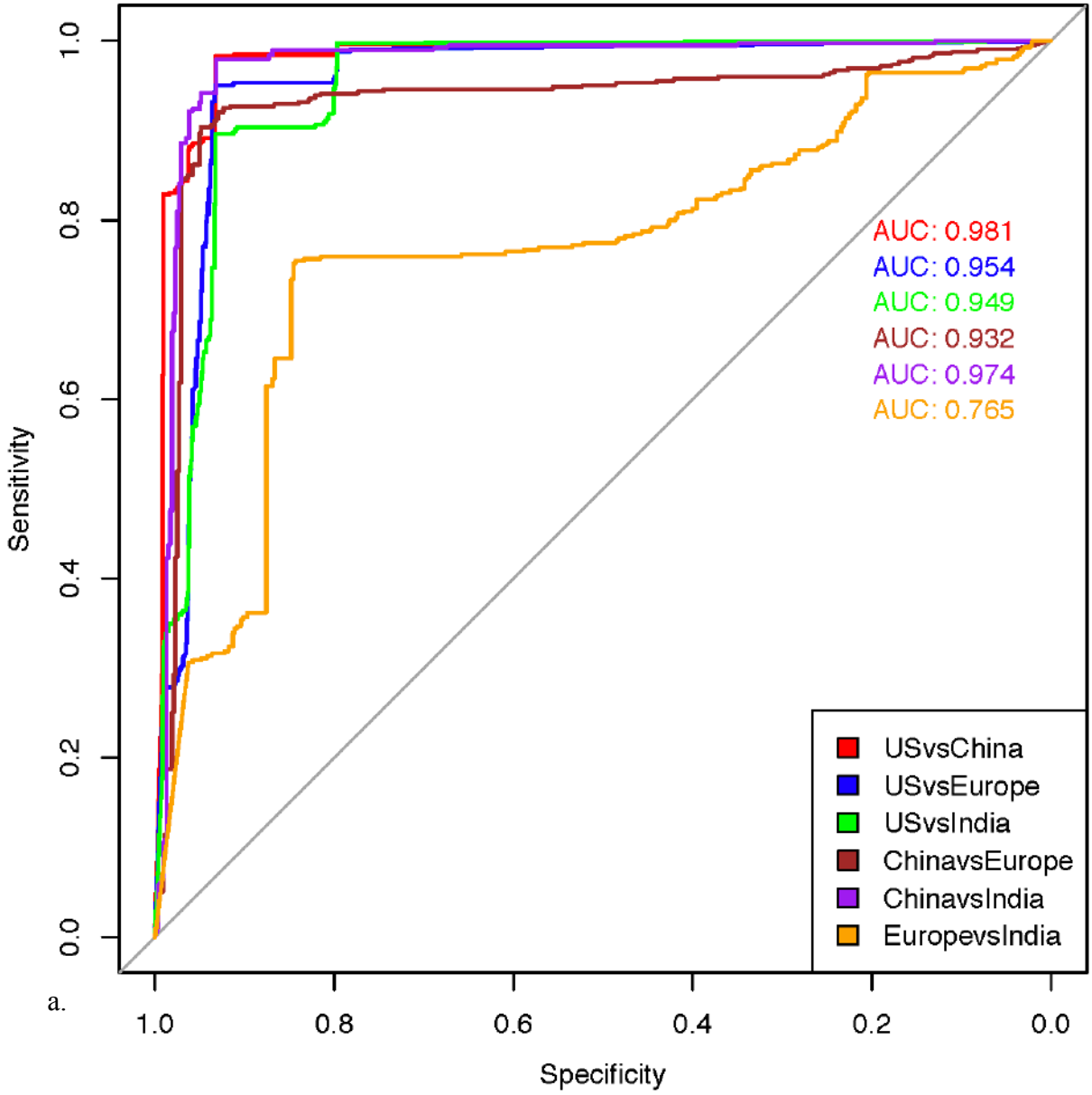
Pachetti, M., Marini, B., Benedetti, F., Giudici, F. et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*, **18**, 179.

Qin, H., Chan, M. W., Liyanarachchi, S., Balch, C. et al. (2009). An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. *BMC System Biology*, **3**, 73.

Yadav, P.D., Potdar, V.A., Choudhary, M. L., Nyayanit, D. A. et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian Journal of Medical Research*, **151**, 200-209.

Zhao, X., Sehgal, M., Hou, Z., Cheng, J. et al. (2018). Identification of residues controlling restriction versus enhancing activities of IFITM proteins on entry of human coronaviruses. *Journal of Virology*, 92.

Figure 1: ROC curve between classes for (a) training set USA-NY (Class 1); China (Class 2); Italy, Spain (Class 3); India (Class 4)



a.

Figure 2: Model features and their importance

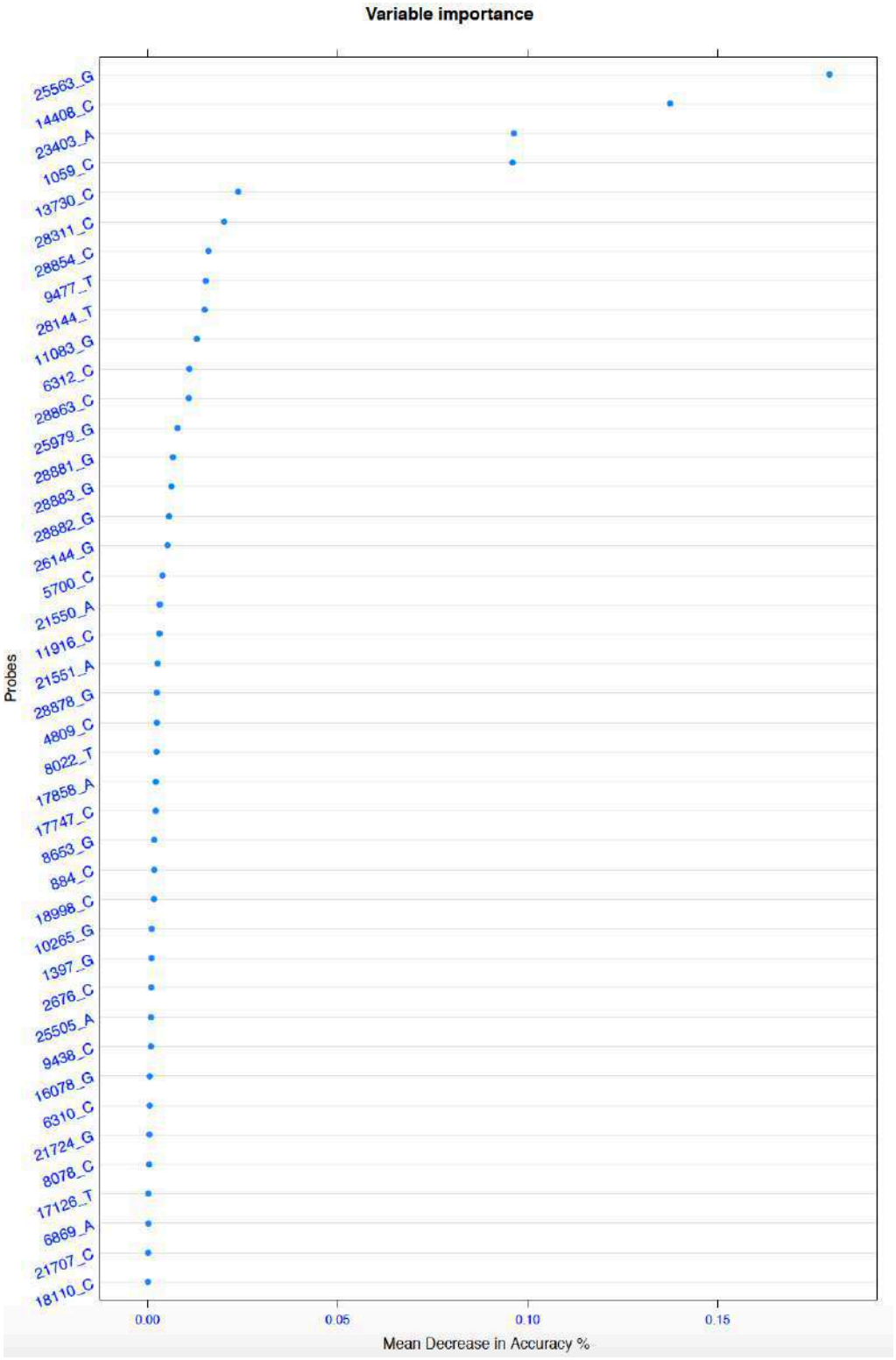
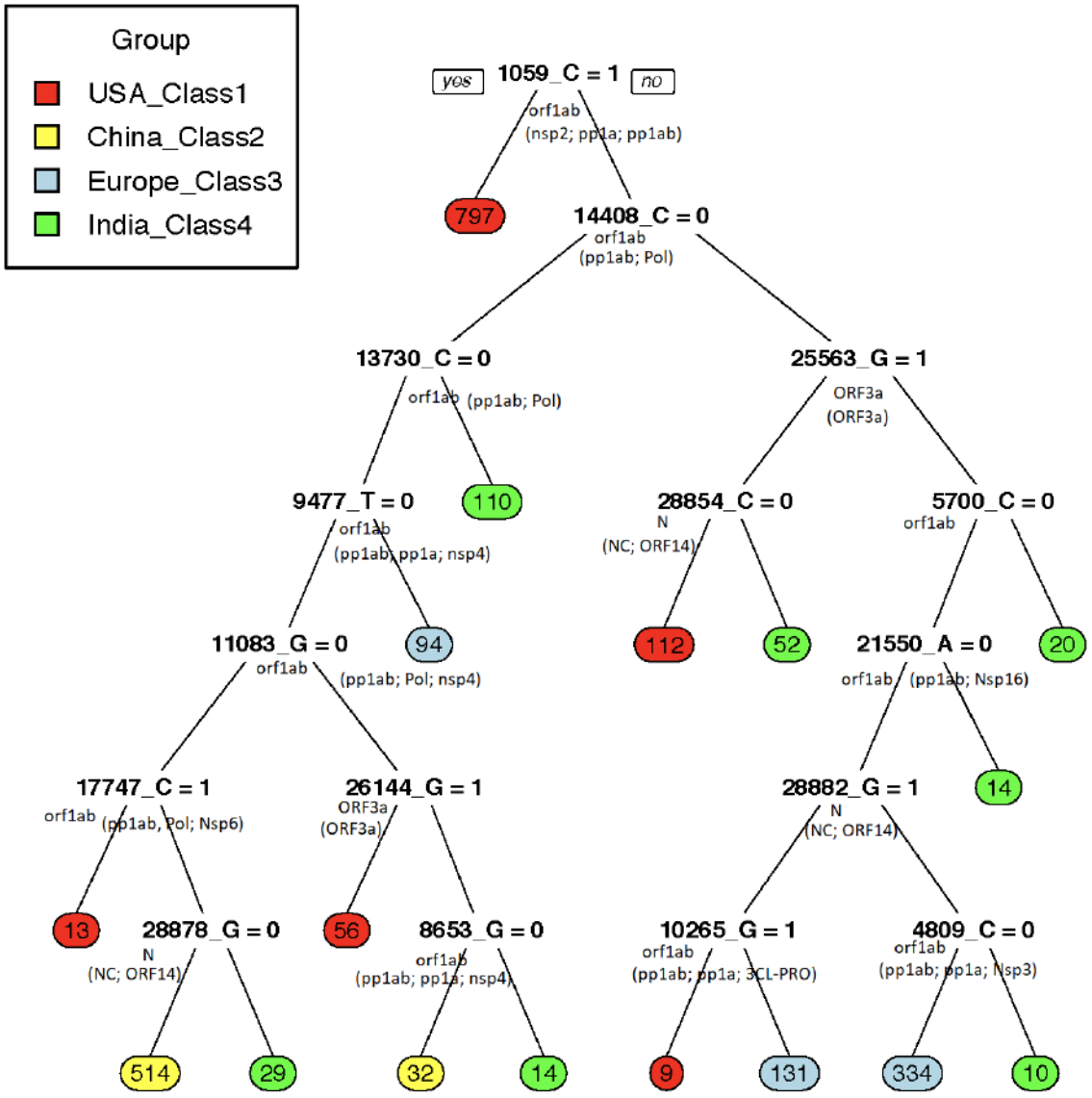


Figure 3: Pruned Tree representation of CART model, generated using 42 features selected by Random forest feature selection method. The gene name and UniProt Protein Products or Polypeptide Chains (in parentheses) in which the mutation is located is mentioned at the bottom of each mutation in the tree.

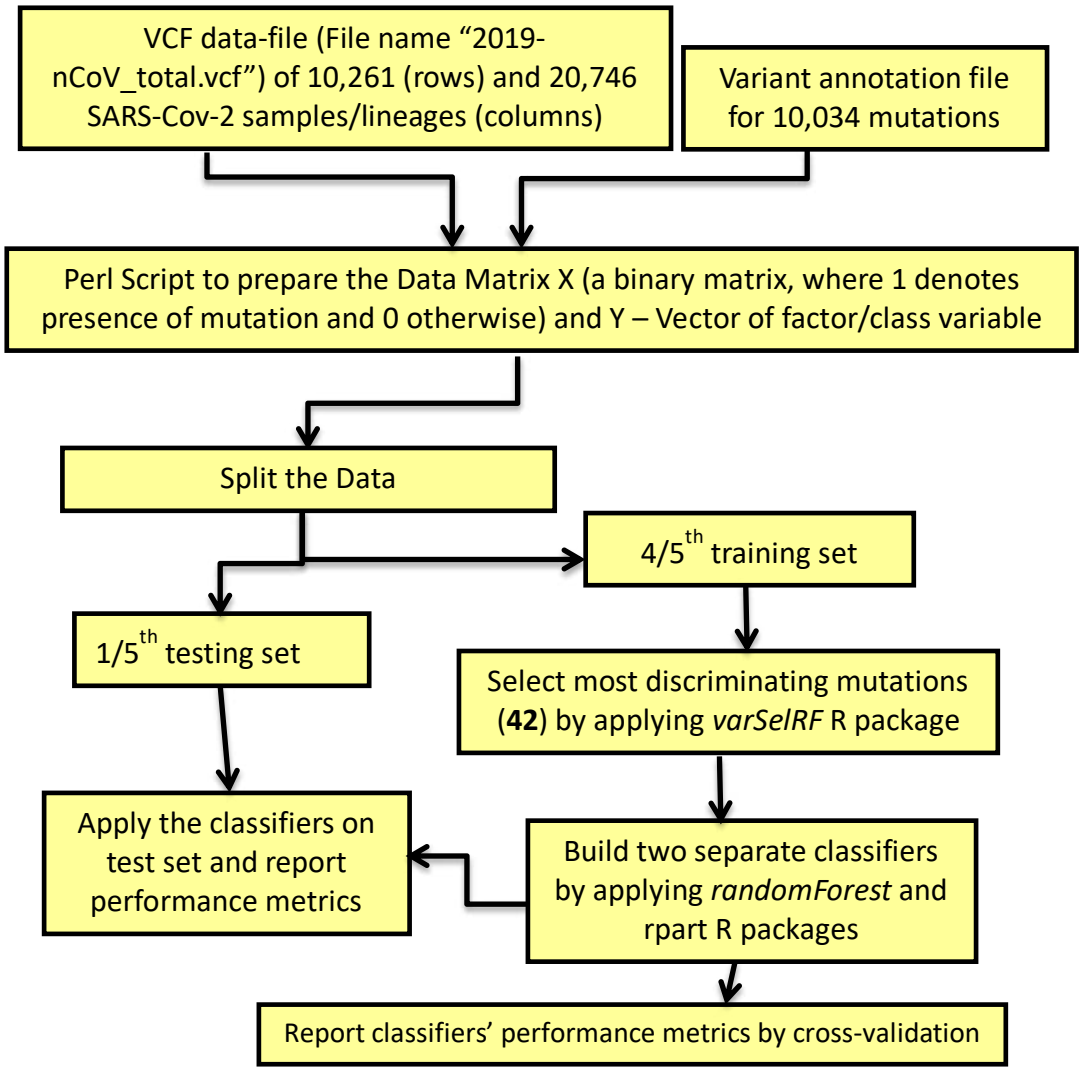


Supplement Table 1: List of top ranking (top 10) mutations. First four rows are header lines; fifth line is for column headings. Columns 1 to 9 provide mutation information, such as the chromosome (CHROM), genomic position (POS), unique identifier (ID), reference allele (REF), alternative alleles (ALT) identified in different lineages, sequence quality score (QUAL), filtering out (FILT) criteria for low quality mutations, any information (INFO) and format of the mutation, GT – Genotype. Genotype data are given for two samples, one for USA and the other from India. Missing information is denoted by period "." symbol. If more than one alternative alleles exist, those are comma-separated in ALT column. The nucleotide symbols in REF and ALT columns are: A – Adenine; C – Cytosine; G – Guanine; T – Thymine; R – G or A (purine); Y – C or T (pyrimidine); K – G or T; M – A or C; S – G or C; W – A or T; B – G or T or C; D – G or A or T; H – A or C or T; V – G or C or A.

##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##reference=file:///xtdisk/apod/licp/Virus/ref/2019-nCoV.fa
##contig=<ID=2019-nCoV,length=29903>

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	2019-nCoV/USA-AZ1/2020	SARS-CoV-2/human/IND/GMCKN318/2020
2019-nCoV	23403	2019-nCoV_23403	A	R,G	.	.	.	GT	0	2
2019-nCoV	14408	2019-nCoV_14408	C	A,T,Y	.	.	.	GT	0	2
2019-nCoV	3037	2019-nCoV_3037	C	T,Y	.	.	.	GT	0	1
2019-nCoV	241	2019-nCoV_241	C	T,Y	.	.	.	GT	0	1
2019-nCoV	25563	2019-nCoV_25563	G	T,C,R,K	.	.	.	GT	0	0
2019-nCoV	1059	2019-nCoV_1059	C	T,Y	.	.	.	GT	0	0
2019-nCoV	28881	2019-nCoV_28881	G	A,T,R	.	.	.	GT	0	0
2019-nCoV	28882	2019-nCoV_28882	G	A,T,R	.	.	.	GT	0	0
2019-nCoV	28883	2019-nCoV_28883	G	A,S,C	.	.	.	GT	0	0
2019-nCoV	8782	2019-nCoV_8782	C	T,Y	.	.	.	GT	1	0

Supplementary Figure 1: Flowchart of the computational frame-work



Statistics and Applications ISSN 2452-7395 (online)
Volume 18, No. 1, 2020 (New Series), pp 269–280

A Comprehensive Modeling Framework for Estimation and Prediction of COVID19 in India

P. Venkatesan
Former Scientist-F, Department of Statistics, ICMR National Institute for Research in Tuberculosis, Chennai-600 031

Received: 15 June 2020; Revised: 01 July 2020; Accepted: 03 July 2020

Abstract

There has been substantial interest worldwide in understanding the current status of Corona Virus Disease (COVID-19) epidemic and prediction of the future path through the pandemic. Many groups are attempting to provide the description of spread and modeling the transmission processes for short and long term projections. Since the epidemic is in its early stage, there is paucity of data for predicting the future course of the disease. The back-calculation approach is one of the methods used in such a situation. The back-calculation reconstructs the past pattern of the infection and predicts the future number of cases with the present infection curve. Lack of information about incubation distribution, effect of intervention on incubation period and errors in reporting the cases lead to uncertainties associated with modeling. This paper attempts to formulate the problem of estimating future COVID-19 cases as estimation of parameters in a multinomial likelihood with unknown sample size by EM algorithm. Illustrations are provided using reported cases in India and discussed.

Key words: COVID-19; Back-calculation; EM algorithm; Incubation period; Infection density.

1. Introduction

The extensive world-wide spread of COVID-19, which started in late 2019 in China, has become the first modern pandemic in less than six months (Korean Society for Infectious Diseases, 2020; Li *et al*., 2020; Liu *et al*., 2020, Sun and Vibound 2020). Given the need to develop a better understanding of the levels and trends in the epidemic and the limited information on which to base these estimates, the use of modeling approaches can make a valuable contribution has seen in other epidemics (*e.g.* Solomon *et al.*, 2003, Ravanan and Venkatesan 2008). The goal of any modeling exercise is to extract as much information as possible from the available data in order to provide an accurate representation of both the knowledge and uncertainty about the epidemic.

A range of different types of models have been developed and applied to the estimation of epidemics in variety of settings. (One major tradition in modeling infectious diseases like HIV and COVID-19 epidemic is the use of back-calculation of back projection techniques which provides statistical solutions convolution equations relating the number of cases diagnosed over time and incubation period distributions (Anderson, 1988; Venkatesan, 2006; Liu *et al*., 2020; Nishiura, 2020) The objective of this paper is to highlight areas in which further methodological developments are needed given currently available data sources. In general, epidemic modeling is categorized in to four broad categories, but not mutually exclusive ones.

Corresponding Author: P. Venkatesan
E-mail: venkaticmr@gmail.com

(a) *Deterministic models*: In this type of modeling the parameters such as number of susceptible, infected and disease individuals are assumed to deterministic. These models are described by a system of differential or integral equations. The progression of the epidemic is studied using these equations. Many such models were developed in the past for diseases such as HIV/AIDS (Anderson *et al*., 1989; Hyman and Stanley, 1988; Anderson and May, 1992).

(b) *Stochastic models*: Stochastic models assume that some of the key parameters are random variables. It is assumed that is a continuous time stochastic process. The stochastic models are considered to be more realistic than deterministic models and with some special assumptions the results of the deterministic models can be approximated through stochastic models. Several studies showed that stochastic models gave a better interpretation of epidemics than deterministic models (Tan, 2000; Mode *et al*., 1988; Isham, 1991).

(c) *Statistical models*: The statistical models are based on epidemiology of the disease and survey/surveillance data. These models make full use of the available data compared to deterministic and stochastic models. In this type of modeling the disease mechanism and prior information are usually not considered. The back-calculation approach for projection of disease epidemics can be categorized in to this type of modeling (Jewel *et al.*, 1992; Bacchetti *et al*., 1993; Venkatesan, 2006; Ravanan and Venkatesan, 2008; Egan and Hall, 2015).

(d) *State-Space models***:** The state-space models have been introduced by Wu and Tan (1995) for modeling HIV/AIDS epidemic, which takes advantages of both stochastic and statistical models. The state-space models were originally proposed for engineering control and communication. This model was also used for projections and detailed description is given by Tan (2000).

2. Back-Calculation Methodology

Brookmeyer and Gail (1986, 1988) introduced back calculation method for short-term projection of HIV/AIDS epidemic. This method uses a form of infection curve, either parametric or non-parametric, for the number of past HIV infections or equivalently a density function for infections as noted by Ding (1995, 1996). The time between infection and the diagnosis of disease is known as incubation time and it is modelled by a known distribution. Many distributions are used for the incubation curve depending on the length (Lawless and Sun, 1992; Ravanan and Venkatesan, 2008; Venkatesan *et al*., 2012). The next section presents some of the useful distributions for modeling the COVID-19 infection curve. The formulation of back-calculation for discrete and continuous cases are considered here.

2.1. Discrete back-calculation formulation

The number of reported COVID-19 cases is available during the calendar time T_0 to T_L. Here T_0 denotes the start of the epidemic and T_L denotes the time up to which the data on reported COVID-19 cases are available. The back-calculation method to reconstruct the COVID-19 infection and projection of future COVID-19 cases can be described in the following sections.

Let X_j denotes the number of COVID-19 cases reported in the interval $[T_{j-1}, T_j), j = 1,2,..., L$. Let $N = \sum_{i=1}^{L} X_i$, then $(X_1, X_2, ... X_L)$ can be assumed to follow multinomial distribution $(N, p_1, p_2, ..., p_L)$. Here p_j denotes the probability that a person infected at time T_0 is diagnosed with COVID-19 in the j^{th} interval. This probability p_j may be estimated using the equation

$$p_j = \frac{1}{N} \sum_{i=1}^{j} I_j [F(j+1-i) - F(j-i)] \tag{1}$$

where $F(t)$ denotes the discrete time formulation of incubation period distribution and I_j denotes the number of infected individuals at the beginning of the j^{th} interval.

Let $$f_{j-I} = F(j+1-i) - F(j-i) \tag{2}$$

then equation (1) can be modified as

$$p_j = \frac{1}{N} \sum_{i=1}^{j} I_j f_{j-i} \tag{3}$$

If p_j's values are known then Np_j denotes the expected number of COVID-19 cases in the interval $[T_{j-1}, T_j)$. Estimation of p_j is done by using various approaches. Let us consider the multinomial likelihood method of estimation of p_j, where a form of infection curve is given by $I_j = I_j(\theta_1, \theta_2, \ldots, \theta_k)$ which is assumed to be known except the k parameters. Therefore, p_j is a function of k unknown parameters, assuming $F(t)$ is completely specified. The unknown parameters p_j can be obtained using the multinomial likelihood as

$$L = \frac{N!}{x_1! x_2! ... x_L!} p_1^{x1} p_2^{x2} ... p_L^{xL} \tag{4}$$

Then $$\log(L) = N! - \sum_{i=1}^{L} \log x_i! + \sum_{i=1}^{L} x_i \log p_i \tag{5}$$

Fisher's scoring algorithm can be used to estimate the unknown parameters θ_i's and hence p_j can be estimated. The above formulation has been used by Taylor (1989). Future COVID-19 cases in the k^{th} time point following T_L can be obtained using the equation

$$\hat{X}_{T_{L+k}} = \sum_{j=k+1}^{T_L} I_j [F_{T_{L+1+k-j}} - F_{T_{L+k-j}}]$$

$$= \sum_{j=k+1}^{T_L} I_j f_{T_{L+k-j}} \tag{6}$$

where $\hat{X}_{T_{L+K}}$ is the minimum number of COVID-19 cases in the interval T_{L+K}.

2.2. Continuous time formulation of back-calculation

In the discrete time formulation the incubation time was treated as a discrete random variable. If the incubation time is treated as a continuous random variable, then the probability of infection in the j^{th} interval given in equation (1) can be rewritten as

$$p_j = \frac{1}{N} \int_{T_0}^{T_j} I(\theta,t)[F(T_j - t) - F(T_{j-1} - t)]\ dt \tag{7}$$

Now $I(\theta,t)$ is assumed to be a smooth function of t. Brookmeyer and Gail (1986) modified (7) by assuming $D(\theta,t)$ to be the density function of infection times of N individuals. Therefore $\int_{T_0}^{T_L} D(\theta,t)\ dt = 1$ and the equation (7) can be written as

$$p_j = \int_{T_0}^{T_j} D(\theta,t)[[F(T_j - t) - F(T_{j-1} - t)]\ dt \tag{8}$$

Thus a model for infection curve is $I(\theta,t)$ and a model for infection density are related by $I(\theta,t) = ND(\theta,t)$ where N = $\int_{T_0}^{T_L} I(\theta,t)dt$.

Hence we now work in the formulation of p_j as given in equation (8). The parameter in p_j can be estimated using the Fisher's scoring algorithm assuming a multinomial likelihood. Brookmeyer and Gail (1988) formulated the problem of estimation of future cases in short interval of time as the problem of estimation of parameters in multinomial likelihood with unknown sample size and the method as explained in in the next section.

2.3. EM algorithm approach

The Expectation-Maximization (EM) algorithm was first proposed by Dempster *et al*. (1977) for the analysis of incomplete data. The algorithm is formulated as follows:

X_{L+1} denote the number of individuals infected before the time T_L who have not become COVID-19 cases by time T_L. The problem is to estimate the total number of infections before the time T_L. This number N is the minimum size of the COVID-19 epidemic, because even if the infections after the time T_L could be prevented, the cumulative number of COVID-19 cases would eventually reach N. The minimum size is the sum of all cases already diagnosed, called $n = \sum_{i=1}^{L} X_i$ and all the susceptible individuals infected before T_L but not yet diagnosed, called $X_{L+1}=N-n$. It can be noted that, in this formulation both N and X_{L+1} are unknown. Therefore an estimate of the minimum cumulative incidence of COVID-19 that can be anticipated in some future time point T_{L+1} is

$$n + \hat{N}\hat{P}_{L+1} = n + \hat{N}\int_{T_0}^{T_L} I(\theta,t)[F(T_{L+1}-t) - F(T_L - t)]dt \tag{9}$$

where $\hat{N}$ is the estimate of N and $\hat{P}_{L+1}$ is the probability of becoming COVID -19 in the future interval $[T_L, T_{L+1})$.

Assuming $I(\theta,t)$ as step function, Brookmeyer and Gail (1988) gave the following EM algorithm for the estimation of the parameter.

Suppose $I(\theta,t) = \theta_i$ for t in $[T_{r_{i-1}}, T_{r_i})$ $i = 1,2,...L.T_{r_i}$ denote the time point defining the i^{th} step. Let X_{ij} denote the number of COVID-19 cases who were infected in the i^{th} step $[T_{r_{i-1}}, T_{r_i})$ and diagnosed in the j^{th} interval $[T_{r_{i-1}}, T_{r_i})$. Note that $T_{r_{i-1}} \le T_j$, since X_{ij} is not defined if $T_{r_{i-1}} > T_j$.

For a fixed N,

$$\hat{X}_{ij} = X_j(\hat{P}_{ij}/\hat{P}_j) \tag{10}$$

where $\hat{P}_{ij}$ is the estimated probability that individual infected in the i^{th} step is diagnosed as COVID-19 in the j^{th} interval. These estimates are obtained using the current estimate of the θ values at the m^{th} iteration *i.e.*, $\hat{\theta}^{(m)}$, then the updated estimates are obtained using the equation

$$\hat{\theta}_i^{(m)} = \sum_{j=1}^{L+1} \hat{X}_{ij} / N\delta_i \tag{11}$$

where δ_i is the width of the i^{th} step. The numeration in equation (11) is an estimate of the number of individuals N infected during the i^{th} step. Further detail of the algorithm for a step function $I(\theta,t)$ is given by Brookmeyer and Gail (1988).

3. Statistical Models for Incubation Period

The incubation period models are similar to survival models based on non-negative random variables and can be fitted using either parametric or semi-parametric approach. A detailed description can be found in Lawless (2011). Here we restrict our attention to only parametric models for incubation period as described in our earlier work (Ravanan and Venkatesan, 2008). The common distributions used for the incubation distribution are given in Table 1 and the infection densities used for prevalence are given in Table 2. Two other important distributions used are the staging model mode and change point model which are described below:

3.1. Staging model

Under staging models the incubation period is considered to be comprised of stages (Brookmeyer and Liao 1990). Different models for these two stages can be assumed. Let $h_1(t)$ and $h_2(t)$ denote the hazard functions of the two stages. The convolution equation for the incubation period comprising of these two stages

$$F(t) = \int_0^t f_1(u)\, F_2(t-u)du \tag{12}$$

where $$f_1(u) = h_1(u)\exp\{-\int_0^u h_1(s)ds\} \tag{13}$$

and $$F_2(u) = 1-\exp\{-\int_0^u h_2(s)ds\} \tag{14}$$

Suitable changes should be made in the above formulations to account for calendar time of infection

Table 1: Incubation period distributions

Model	Distribution Function
Weibull	$F(t) = 1-e^{-(\lambda t)^{\alpha}}$ $\quad \lambda>0, \alpha>0,\ t>0$
Gamma	$F(t) = \frac{1}{\sigma\Gamma(k)}\int_0^t \left(\frac{x}{\sigma}\right)^{k-1} \exp(-\frac{x}{\sigma})\ dx \quad t>0,\ \sigma>0,\ k>0$
Lognormal	$F(t) = \Phi\frac{(\log t-\mu)}{\sigma},\ \Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx$ $t>0,\ -\infty<\mu<\infty,\ \sigma>0$
Log-logistic	$F(t) = 1-[1+(\lambda t)^{\upsilon}]^{-1} \quad \lambda>0,\ \upsilon>0,\ t>0$
Gen. Exponential	$F(t) = 1-[1-\exp\{-t\lambda\}]^{\upsilon} \quad t>0,\ \lambda>0,\ \upsilon>0$
Gen. Log-logistic	$G(t) = \frac{1}{\beta(m_1,m_2)}\int_0^{H(t)} x^{m_1-1}(1-x)^{m_2-1}dx;\ t>0, m_1>0, m_2>0$ $H(t) = \frac{1}{1+\exp\{-[\log(t)-\mu]/\tau\}} \quad t>0,\ -\infty<\mu<\infty,\ \tau>0$
Gen. Gamma	$F(t) = \int_0^t\left[\sum_{j=1}^{k} B_{1k}(j)\exp(-\lambda_j x)\right]dx$ $B_{1k}(j) = (\prod_{i=1}^{k}\lambda_i)/\prod_{\substack{i=1\\ i\neq j}}^{k}(\lambda_i-\lambda_j),\quad \lambda_j = 1/\sigma_j$
Mixed Weibull	$F(t) = pF_1(t)+(1-p)F_2(t) \quad 0<p<1$ $F_1(t) = 1-e^{-(\lambda_1 t)^{\alpha 1}} \quad \lambda_1>0, \alpha_1>0,\ t>0$ $F_2(t) = 1-e^{-(\lambda_2 t)^{\alpha 2}} \quad \lambda_2>0, \alpha_2>0,\ t>0$

3.2. Change point model

Estimates of the population parameters are obtained in the case of subpopulations are exponentially distributed and sampling is censored at a predetermined test termination time was first introduced by Mendenhall and Hader (1958). The change point model considered here is briefly presented below.

Suppose the hazard before and after the change point is constant, then $h(t)$ is given by

$$h(t) = \begin{cases} \alpha & t \le \tau \\ \beta & t > \tau \end{cases} \tag{15}$$

The distribution function is given by

$$F(t) = \begin{cases} 1 - e^{-\alpha t} & t \le \tau \\ 1 - e^{-\alpha\tau}\, e^{-\beta(t-\tau)} & t > \tau \end{cases} \tag{16}$$

The median of the incubation period is given by

$$M = \begin{cases} \log 2 / \alpha & M \le \tau \\ (2e^{-\tau\alpha} - 1)/(e^{\tau(\beta-\alpha)}) & M > \tau \end{cases} \tag{17}$$

Table 2: Infection densities

Model	Infection Density
Logistic Prevalence	$D_1(\theta,t) = \frac{1}{k_1} I_1(\theta,t) = \frac{1}{k_1} \frac{\theta_1\theta_3 e^{\theta_2+\theta_3 t}}{(1+e^{\theta_2+\theta_3 t})^2}$
Logistic Incidence	$D_2(\theta,t) = \frac{1}{k_2} I_2(\theta,t) = \frac{1}{k_2} \frac{\theta_1 e^{\theta_2+\theta_3 t}}{1+e^{\theta_2+\theta_3 t}}$
Double Exponential incidence	$D_3(\theta,t) = \frac{1}{k_3} I_3(\theta,t) = \frac{1}{k_3} \theta_1 e^{-e^{\theta_2+\theta_3 t}}$
Log-logistic incidence	$D_4(\theta,t) = \frac{1}{k_4} I_4(\theta,t) = \frac{1}{k_4} \theta_1\theta_2(\theta_1 t)^{\theta_2 - 1}/[1+(\theta_1 t)^{\theta_2}$
Exponential incidence	$D_5(\theta,t) = \frac{1}{k_5} I_5(\theta,t) = \frac{1}{k_5} \theta_1 e^{\theta_2 t}$
Root exponential incidence	$D_6(\theta,t) = \frac{1}{k_6} I_6(\theta,t) = \frac{1}{k_6} \theta_1 e^{\theta_2 t^{1/4}}$ $k_i = \int_{T_0}^{T_L} I_i(\theta,t)\, dt \quad ; \quad i = 1,2,\ldots,6.$

4. An Illustration Using Indian Data

The basic data required for back-calculation methodology is the number of COVID-19 cases over a period of time (Brookmeyer and Gail, 1986, 1988, 1990; Ding, 1995, 1996). The Ministry of Health publishes daily updates of the reported COVID-19 cases for the past few

months. The updates of the recent days also suffer reporting delays and under reporting and therefore pooled weekly reported cases may be more reliable. In this illustration only weekly reported cases were considered. The period is from 1st March 2020 to 30th May 2020 (13 weeks) (https://www.coronatracker.com/country/india/). It is also reported that level of under reporting may vary from 50-90%. For this work, it is assumed that the level of under reporting is around 90% in early March 2020 and gradually decreased to 50% exponentially in the end of May 2020. The exponential decay model

$$P(t) = 0.90\, e^{-0.05t} \tag{18}$$

gives a better approximation of the above assumption. The upward adjustments for the weeks are carried out. Table 3 gives the actual number of reported and adjusted COVID-19 cases along with 3-week moving averages. The reported cases are smoothed using a three week moving averages as a first step. The linear and quadratic models are fitted to find the best fit liner model for the trend for the moving average cases which serves as a bench mark for comparisons. The results are given in Table 4. From the table it is seen that the quadratic trend model seems to be a better fit for moving averages and the corresponding model is

$$\text{Cases} = 1114.9 - 1205.7\, Time + 373.2\, Time^2 \tag{19}$$

Table 3: Weekly confirmed COVID-19 cases in India

Week	Weekly Confirmed COVI-19 Cases	3week Moving Average	Cumulative COVID-19 Cases	Adjusted weekly Cases
March 1-7	31	-	31	63
March 8-14	50	104	81	93
March 15-21	231	294	312	419
March 22-29	601	995	913	1067
March29-April 4	2154	2407	3067	3741
April 5-11	4467	4628	7534	7581
April 12-18	7263	7460	14797	12105
April 19-25	10650	10319	25447	17404
April 26-May 2	13044	15193	38491	20913
May 3-9	21886	20924	60377	34446
May 10-16	27842	30168	88219	43040
May17-23	40775	39797	128994	61948
May24-30	50404	-	179398	75300

Table 4: Trend lines based on the moving averages

Trend	Variable	B	Se(B)	Z	Sig	R^2
Linear	Constant	–8742.3	2500.8	–3.496	0.01	
	Time	3333.0	368.7	9.039	8.24e-05	0.889
Quadratic	Constant	879.0	1314.0	1.505	0.163	
	Time	–1107.6	519.2	–4.655	0.0009	0.984
	$Time^2$	370.1	49.9	7.413	7.54e-05	

4.1. Estimation of parameters

Based on the availability of data, the starting point of the epidemic T_0 is taken as March 2020. The incubation distributions discussed in the previous section are used in this section to illustrate projection of COVID-19 in India. The estimates of minimum size of the epidemic and future COVID-19 cases are obtained assuming a median incubation period of two weeks. For the incubation period models Weibull, gamma, log-logistic, log-normal and generalized exponential distribution prior estimates of their parameters are obtained methods described in Venkatesan (2006). All these models have two parameters and one parameter is fixed based on the estimates reported (Table 1). The other parameter was determined such that the median distribution period is known. The parameters of the generalized log-logistic, generalized gamma, mixed Weibull and change point models are not available. The parameters of these models are decided based on the simulation study as described in our earlier work (Ravanan and Venkatesan, 2009). For the infection density, the exponential, root exponential, double exponential, logistic and log-logistic are commonly used (Table 2). In this work only logistic density incidence based projections are given for illustrative purpose. The projections based on logistic infection density under various incubation period distributions are presented in Table 5.

Table 5: Projection of COVID-19 prevalence under logistic infection density and total expected confirmed cases (Median incubation = 2 weeks)

Incubation period Model	***Projection of COVID-19 cases ('000)***		
	Up to June 6	**Up to June 13**	**Up to June 20**
Weibull	240.5	318.8	408.1
Gamma	242.8	321.9	411.8
Log-logistic	248.5	327.7	419.2
Log-normal	244.8	327.0	419.7
Gen. Exponential	249.6	328.7	420.9
Gen. Log-logistic	248.0	326.9	418.4
Gen. Gamma	249.3	327.1	417.8
Mixed Weibull	241.5	322.1	414.4
Change point	247.7	326.1	418.1
Quadratic	246.9	324.8	413.8
Observed	236.2	321.6	411.8

From Table 5 we see that the projections obtained for the next three weeks under different models do not differ widely. This may be due to the behaviour of the epidemic in the early stage. However, the projections based on Weibull, gamma, lognormal and mixed Weibull are close the observed cases. The quadratic model also gives results close to the observed cases. The projections based on exponentially adjusted cases for under reporting resulted higher cases and are not reported here. We also considered the other infection densities for incidence given

in Table 2. They resulted in higher cases than the logistic prevalence. Hence only the results pertaining to the logistic prevalence infection density are given to illustrate the use of the models. After obtaining sufficient data in the infection curve, the comparisons will provide valid estimates.

5. Discussion

There has been research showing that on average, each infected person spreads the infection to more than two persons. Therefore the majority of the population is at risk of infection if no intervention measures were undertaken. The true size of the COVID- 19 epidemic remains unknown, as a significant proportion of infected individuals only exhibit mild symptoms or are even asymptomatic. Timely assessment of the evolving epidemic size is crucial for resource allocation and plan strategies. In this article, we used the back-calculation algorithm to obtain a lower bound estimate of the numbers of COVID-19 infected confirmed cases in India using the available data. Since the data source is limited and suffers from under reporting, under diagnosis and delay in reporting, adjustments are needed before making any modeling and projections.

One of the critical issues in infectious disease epidemiology is that the time of infection event is seldom directly observable. For this reason, the time of infection needs to be statistically estimated, employing a back-calculation method. It is observed that the short-term projection of three weeks do not vary much across various incubation period distributions. Further the estimates vary widely for different infection densities. The projected COVID-19 cases for three weeks under Weibull, gamma, lognormal and mixed Weibull are similar and close to the confirmed cases. One reason could be that they are related models and the do not differ in the initial stages. We also considered projections under the logistic, exponential double exponential and root exponential infection densities with varying median incubation periods. But the estimates vary significantly particularly under exponential infection density. Since the infection curve is at its early sage, we have reported only the logistic infection density estimates as an illustration. Once sufficient size data is available, the comparisons are reliable. This paper provides a methodology based on the back-calculation for short-term projections which ae widely used in diseases like HIV/AIDS.

References

Anderson, R. M. (1988). The role of mathematical models in the study of HIV transmission and the epidemiology of AIDS. *AIDS*, **1**, 241-246.

Anderson, R. M., Blythe, S. P., Gupta, S. and Konings, E. (1989). The transmission dynamics of the human immunodeficiency virus type I in the male homosexual community in the United Kingdom: The influence of changes in sexual behaviour. *Philosophical Transactions of Royal Soc*iety, **B325**, 45-98.

Anderson, R. M. and May, R. M. (1992). Understanding the AIDS epidemic. *Scientific Amer*ican, **266**, 58-66.

Bacchetti, P., Segal, M. and Jewell, N. P. (1993). Back-calculation of HIV infection rates, *Statistical Science*, **8**, 82-119.

Brookmeyer, R. and Gail, M. H. (1986). Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. *Lancet*, **2**, 1320-1322.

Brookmeyer, R. and Gail, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association*, **83**, 301-308.

Brookmeyer, R. and Liao, J. (1990). Statistical modeling of the AIDS spread for forecasting health care need. *Biometrics*, **46**, 1151-1163.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* **B39,** 1-3.

Ding, Y. (1995). Computing back-calculation estimates of AIDS epidemic. *Statistics in Medicine*, **14**, 1505-1512.

Ding, Y. (1996). On the asymptotic normality of multinomial population size estimates with application to the back-calculation epidemic of AIDS. *Biometrika*, **83**, 695-699.

Egan, J. R. and Hall, I. M. (2015). A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. *Journal of the Royal Society Interface*, **12**, 20150096, 1-14.

https://www.coronatracker.com/country/india/ (Accessed as on 30/05/2020)

Hyman, J. M. and Stanley, E. A. (1988). Using mathematical models to understand the AIDS epidemic. *Mathematical Biosciences*, **90**, 415-474.

Isham, V. (1991). Assessing the variability of stochastic epidemic. *Mathematical Biosciences*, **107**, 209-224.

Jewell, N. P., Dietz, K. and Farewell, V. T. (1992). *AIDS Epidemiology: Methodological issues*. Birkhauser, Basel.

Korean Society of Infectious Diseases/Korean Society of Pediatric Infectious Diseases/ Korean Society of Epidemiology/Korean Societyfor Antimicrobial Therapy/Korean Society for Healthcare-associated Infection Control and Prevention/Korea Centers for Disease Control and Prevention, (2020) Report on the Epidemiological Features of Coronavirus Disease 2019 (COVID-19) Outbreak in the Republic of Korea from January 19 to March 2, 2020. *Journal of Korean Medical Sciences*, **35**, e112.

Lawless, J. F. (2011). *Statistical Models and Method for Life Time Data*. John-Wiley.

Lawless, J. F. and Sun, J. (1992). A comprehensive back-calculation framework for the estimation and prediction of AIDS cases. In: *AIDS Epidemiology: Methodological Issues*, (Eds. Jewell, N. P., Dietz, K. and Farewell, V. T.), 81-104.

Li, Y., Wang, B., Peng, R., Zhou, C., Zhan, Y., Liu, Z., Jiang, X. and Zhao, B. (2020). Mathematical Modeling and Epidemic Prediction of COVID-19 and Its Significance to Epidemic Prevention and Control Measures. *Annals of Infectious Disease and Epidemiology*, **5** (**1052**), 1-9.

Liu, Y., Qin, J., Fan, Y., Zhou, Y., Follmann, D. A. and Huang, C. Y. (2020). Infection Density and Epidemic Size of COVID-19 in China outside the Hubei province. https://doi.org/10.1101/2020.04.23.20074708

Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504–520.

Mode, C. J., Gollwitzer, H. E. and Hermann, N. (1988). A methodological study of a stochastic model of an AIDS epidemic. *Mathematical Biosciences*, **92**, 201-229.

Nishiura, H. (2020). Back-calculating the Incidence of Infection with COVID-19 on the Diamond Princess. *Journal of Clinical Medicine*, **9**, 1-4.

Ravanan, R. and Venkatesan, P. (2008). Some new approaches for modeling the incubation period of HIV/AIDS epidemic. *International Journal of Computer, Mathematical Sciences and Applications,* **2**, 223-237.

Ravanan, R. and Venkatesan, P. (2009). A simulation study on uncertainties associated with Back-calculation methodology. *International Journal on Information Sciences and Computing,* **3**, 47-52.

Solomon, P. J. and Wilson, S. R. (1990). Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics*, **46**, 1165-1170.

Sun, K., Chen, J. and Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *Lancet Digit Health*, **2**, e201-e208.

Tan, W. Y. (2000). *Stochastic Modeling of AIDS Epidemiology and HIV Pathogenesis*. World Scientific Publication, Singapore.

Venkatesan, P. (2006). A Comprehensive back-calculation frame work for estimation and projection of HIV/AIDS in India. *Journal of Communicable Diseases*, **38**, 40-56.

Venkatesan, P., Ramamurthy, D. and Sundaram, N. (2012). HIV/AIDS projection for Tamil Nadu using back calculation method. *Indian Journal of Science and Technology*, **5**, 3157-3162.

Wu, H. and Tan, W. Y. (1995). Modeling the HIV epidemic: A state space approach. In: *American Statistical Association 1995 Proceeding of the Epidemiology Section*, ASA, Alexdria, VA, 66-71.

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 281–294

Reducing the Number of Tests for COVID-19 Infection via Group Testing Methodologies

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

Received: 25 June 2020; Revised: 09 July 2020; Accepted: 12 July 2020

Abstract

Total economic shutdown being detrimental to a nation's prosperity, most governments are reopening businesses and schools with the requirement of frequent and mass-scale testing to determine each person's status of COVID-19 infection. Obviously, the costs add up quickly and impose a heavy economic toll. As a way out of this dilemma, employers and administrators should consider seriously the application of group testing methodologies.

Group testing methods check samples in batches, rather than individually, for the presence of a disease. If the group tests positive, then the group is investigated further to identify who all are positive. On the other hand, if the group tests negative, not just once but also a second or a third time, then everyone within the group is cleared for activity. With a carefully chosen protocol, group testing costs can be 30-80% lower than those of individual testing, with the savings being higher when prevalence of the disease is lower.

Key words: Design of Experiments; Sensitivity; Specificity; False positive; False negative.

AMS Subject Classifications: 62K05

1. Introduction

Originating in Wuhan Province of China in November 2019, the novel corona virus has inflicted the SARS COVID-19 pandemic across the globe. In an immediate attempt to curtail the spread of the virus in the absence of a vaccine, many governments imposed lock-downs on their respective jurisdictions. However, the economies of many states, provinces and countries have been severely damaged because of lock-downs and stay-at-home orders. To prevent a total collapse of the economy, many governments are forced to reopen businesses and schools, notwithstanding the risk of spreading the disease. Therefore, it has become imperative to isolate people who have the disease and quarantine people who test positive for the virus. Consequently, frequent and mass-scale testing for COVID-19 infection has become a necessary precondition for restarting the economy. See, for example, The White House, *et al.* (2020) report.

The cost of testing has become burdensome on the payers (individuals, employers, administrators, insurance companies and governments). Also, testing capacity is often limited. In the midst of this dire situation, the celebrated group testing methodologies can offer a valuable relief without compromising safety.

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

Even though little in this paper is theoretically a new finding, except perhaps the extension to imperfect tests, the importance and necessity of the day prompt us to review this methodology anew. A recent paper by Aprahamian *et al.* (2017) mentions that it is precisely a lack of understanding of how an optimal pooling scheme should be designed to maximize classification accuracy under a budget constraint that hampers screening efforts. Perhaps a wide-spread familiarity with the techniques will not only reduce the cost of administering the tests, but also put suspicious minds at rest knowing that safety is not compromised in an attempt to reduce cost. In fact, Conger (2020) reports how pooling patient samples for COVID-19 testing helped Stanford researchers track the early spread of the virus in the Bay Area prior to the last week of February, 2020.

Section 2 gives a brief history of the group testing methodologies (GTMs). Section 3 determines the optimal group size for a perfect test and computes percentage savings for various choices of group sizes. Section 4 studies the more realistic case of imperfect tests, evaluating sensitivity and specificity of a group's sample as functions of those same quantities for an individual's sample and prevalence of disease, and determines the optimal size and cost savings. Section 5 gives some advancements in Group Testing Designs (GTDs) and GTMs that shed light on identifying a few defective items intermixed with many good items. Section 6 draws some practical implications of GTMs in the COVID-19 context.

All figures are drawn using the freeware R; and the codes are given in the Annexure.

2. A Brief History of Group Testing Methodology Used in Medicine

We borrow the history of group testing from Ding-Zhu and Hwang (1993). GTMs have had a humble start in Dorfman (1943). During the World War II, the United States Public Health Service and the Selective Service carried out a large-scale project to isolate all syphilitic men called up for induction. Testing an individual for syphilis involved drawing a blood sample and then analyzing the sample to determine the presence or absence of syphilis. At the time, performing this test was expensive, and testing every soldier individually would have been very expensive and inefficient.

Here is how the methodology works: Suppose that there are N soldiers. Testing each individual separately requires N tests, which is a reasonable approach if a large proportion of the people are infected. However, if a small proportion of men are infected, there is a much more efficient testing scheme: Split the soldiers into groups, and in each group combine the blood samples together. If one or more of the soldiers in this group has syphilis, then the test will be positive; and each member of the group has to be tested individually to find which soldier(s) are syphilitic. On the other hand, if the test is negative all members of the group are declared free of syphilis using only one test.

Sterrett (1957) improved Dorfman's procedure: Perform individual testing on a positive group sequentially; stop as soon as a defective is identified; then test the remaining items in the group, as a smaller group, since it is likely that none of them are defective. The first thorough treatment of group testing was given by Sobel and Groll (1959) who described five new procedures, including when prevalence is unknown. For the most optimal procedure, they provided an explicit formula for the expected number of tests it would use. Ungar (1960) proved that the range of prevalence of disease (or proportion of defective items), for which there is a group testing plan with expected number of perfect tests less than the

numer of items, is $[0, (3-\sqrt{5})/2) \approx [0, .382)$. Hwang (1972) detects all defective members in a population by group testing. Sobel and Elashoff (1975) used group testing for estimation.

The above methods work under the simplistic assumption that testing is error-free. However, in reality, most often a diagnostic test is "imperfect" in the sense that there is some chance that the result of a test is erroneous—exhibiting either false positive (that is, the test comes out positive when the sample contains no defectives) or false negative (that is, the test is negative even though the sample contains defectives).

As with most diagnostic tests, COVID-19 tests are imperfect. In particular, Yang *et al.* (2020) suggests that the sensitivity of polymerase chain reaction (PCR) tests on samples collected by nasal swab is around 70%, implying that about 30% of infected patients will return a false negative test result. Much of this error is caused by factors related to sample collection (for example, the patient does not have high enough viral-load levels at the time of collection or the swab did not reach the right place) rather than failures of the PCR test itself. But the danger of a false negative in this situation is unacceptably high: If someone believes he or she is free of disease contrary to truth, they risk spreading the disease to other employees, customers or students—precisely what the testing is trying to prevent!

GTMs have seen a lot of advancements during the last 40 years. Applications abound in both industrial product testing and in medical diagnostic tests. We completely skip the industrial applications (referring readers to Wikipedia, n. d.). In Section 5 of this paper, we mention some important advancements on GTDs and GTMs. Subsection 5.1 illustrates use of one such advancement. Here, we highlight a few references on medical diagnostics. Keeler E. and Berwick D. (1976) presents models of how test performance is degraded by pooling, and of the financial savings that pooling allows. They demonstrate the method of computing optimal pool size on a screening test and on a test for gastrin. Schisterman and Vexler (2008) examines the effect of different sampling strategies of biospecimens for exposure assessment that cannot be detected below a detection threshold. They compare use of pooled samples to a randomly selected sample from a cohort to evaluate the efficiency of parameter estimates.

To apply these GTMs, one must ensure that the following assumptions hold: (i) individual samples can be combined into groups, (ii) group tests have comparable accuracy to individual tests, and (iii) results of group tests can be correctly interpreted.

3. Optimal Group Size and Savings: Perfect Test

Consider first the simpler case of group testing with a perfect test. We explain the relationship among group sizes, number of tests and infection rates. Suppose that an employer splits up N workers into groups of equal size g and tests each group (with one or more group tests) in the first stage. In the second stage, all workers who belong to groups that tested positive (at least once) are subjected to individual testing, as in Dorfman (1943).

Extensions of the above simple protocol is not too hard: One can apply this framework to strategies with more than two stages. For example, lower-prevalence regions can afford to test larger pools and carry out multiple pooled testing rounds before beginning individualized testing. Moreover, employers can learn the actual prevalence over time based on the prevalence level revealed in earlier rounds of testing through statistical predictions. If

prevalence is higher or lower than anticipated, pool size can be adjusted in the next round.

Here, we also assume that workers are homogeneous in their risk of testing positive. But the method extends to heterogeneous population: The employer can simply stratify its workforce into subgroups by risk, and solve the test-minimization problem on each subpopulation separately using a smaller group size for a high risk stratum and a larger group size for a low risk stratum. We leave these extensions to the interested readers, and focus on a homogeneous population of employees tested in two stages—group and individual.

Suppose that the prevalence of active infection in the workforce is θ, and it is *known* to the employer. (Even if θ is unknown, the savings might decrease, but still there will be considerable savings as long as θ is estimated reasonably well.) Suppose also that N employees are split into N/g groups of size g each. Then the probability that a pool of g workers contains at least one infected member is $\theta_g = 1 - (1-\theta)^g$. The expected number of groups testing positive is given by $\theta_g\, N/g$. All g members of these COVID-19 positive pools have to be tested individually. Therefore, the expected number of workers receiving follow-up tests will be $\theta_g\, N$. Hence, the total number of tests will be

$$\frac{N}{g} + N\theta_g = N\left\{\frac{1}{g} + 1 - (1-\theta)^g\right\} \tag{1}$$

On the other hand, an individual testing protocol requires exactly N tests.

3.1. Minimize the number of tests

To compare the pooled testing protocol versus the individual testing, it suffices to minimize the number of test *per worker* less one, or to solve the minimization problem:

$$\min_{1\le g\le N}\left\{\frac{1}{g} - (1-\theta)^g\right\}$$

Using calculus, the argmin may be obtained by the solving

$$g^2(1-\theta)^g = [-\ln(1-\theta)]^{-1}$$

and then taking the largest integer no more than the solution, since g is a whole number (because people are indivisible). Alternatively, we can compare a contemplated $g \ge 3$ against the previous value $(g-1)$; and we prefer g over $(g-1)$ if and only if $g(g-1)\theta(1-\theta)^{g-1} < 1$. If this condition holds, we increase g by one and check this condition again; otherwise, we stop and declare the previous g as optimal. Hence, for a given prevalence θ, the optimal group size is given by

$$g^*(\theta) = \max\{g, 1 \le g \le N : i(i-1)\theta(1-\theta)^{i-1} < 1, \text{for all } i = 3, 4, \ldots, g\} \tag{2}$$

See the Annexure for the R codes to compute g^*. Indeed, the optimal g^* is a non-increasing step function of θ. For example, $g^*(0.05) = 5$, $g^*(0.02) = 8$, $g^*(0.01) = 11$, $g^*(0.005) = 16$, $g^*(0.002) = 24$, $g^*(0.001) = 32$. See more details in Figure 1. Of course, aligning with our intuition, we have $g^*(0+) = \lim_{\theta\to 0^+} g(\theta) = N$ and $g^*(1-) = \lim_{\theta\to 1^-} g(\theta) = 1$.

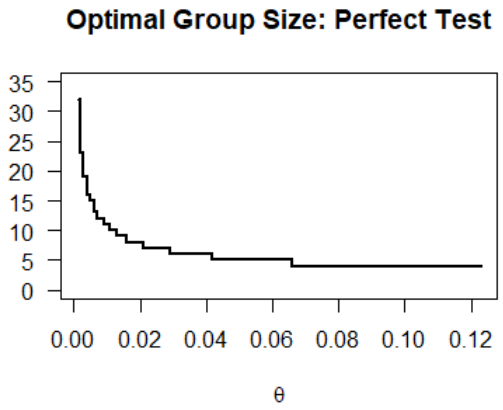


Figure 1: For a perfect test, the optimal group size $g^*(\theta)$

Alternatively, solving the dual problem, one can determine the range of values of θ for which a given group size g is optimal. For instance, $g = 4$ is optimal for $\theta \in (.0655, .1235)$, $g = 5$ is optimal for $\theta \in (.0415, 0.0655)$, etc. Again, more details are in Figure 1. Thus, oftentimes an imprecise knowledge of θ does not significantly alter the optimal choice of g.

3.2. Maximize the percentage savings

For a perfect test, instead of N individual tests, a group testing protocol with a fixed group size g requires fewer number of tests given by Eq. (1). Hence, a group testing protocol achieves a percentage saving over the individual testing protocol given by

$$\text{Savings (Perfect Test)} = \left((1-\theta)^g - \frac{1}{g}\right) \times 100\% \tag{3}$$

Clearly, savings are maximum when the optimal group size $g^*(\theta)$ is chosen; otherwise, savings are reduced. See Figure 2 (and the R codes in the Annexure).

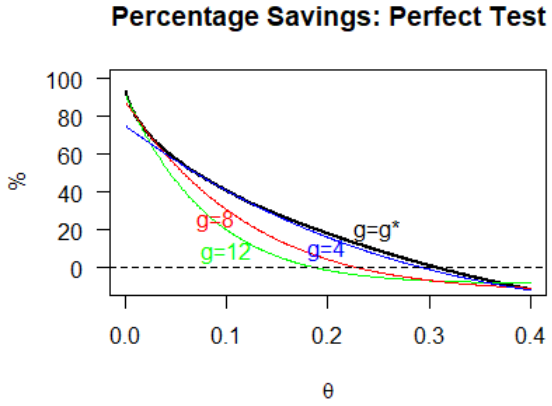


Figure 2: For a perfect test, the percentage savings for various group sizes: optimum and fixed

Caution: The group size must be chosen with careful consideration of the prevalence. For otherwise, savings may be *negative* when group size is chosen far from the optimum. For example, if prevalence is 12% then using the optimum group size $g^* = 4$, given by Eq. (2), the savings would be about 35%, using Eq. (3). However, if group size is (incorrectly) set at 32, then the savings would be -1.45%.

To illustrate the use of Figure 2, employers can reduce the number of tests by using groups of 4 workers, so long as prevalence is under 25 percent. Employers who know the prevalence with higher precision can further reduce the number of tests by choosing larger group sizes. But if employers underestimate the true prevalence rate and use larger group sizes than warranted, then the number of tests will increase.

4. Determining Group Size: Imperfect Test

The best COVID-19 test available to date is still imperfect. Let the sensitivity of the test be $\tau = \Pr\{+|D\}$, the probability that a person with the disease (D) will test positive. Then $(1 - \tau)$ is the probability of a false negative. Likewise, let the specificity of the test be $\eta = \Pr\{-|N\}$, the probability that a person with no disease (N) will test negative. Then $(1 - \eta)$ is the probability of a false positive. Assume that the sensitivity and the specificity remain the same whether we are testing a nasal swab of an individual or the combined swabs of the group (of any size). As mentioned earlier, PCR test for COVID-19 infection has sensitivity 0.70 and specificity 0.99.

4.1. Specificity of group testing

Suppose that a person has no COVID-19 infection. The group's test will be negative because of two disjoint and exhaustive cases: (1) all other $(g - 1)$ group members are also negative and the group tests negative (which is a true negative); (2) not all group members are negative, but the group's test comes out negative (which is a false negative). Writing $a = (1 - \theta)^{g-1}$, the probability that the group's test is negative, obtained by adding the probabilities of true negative and false negative, is

$$a\,\eta + (1 - a)(1 - \tau) = 1 - \tau + a(\eta + \tau - 1) \tag{4}$$

Similarly, the probability that the group's test will be positive, obtained by adding the probabilities of false positive and the true positive, is

$$a\,(1 - \eta) + (1 - a)\tau = \tau - a(\eta + \tau - 1) \tag{5}$$

which is one minus the probability that the group's test is negative, as it should be.

In this case of imperfect test, whenever the group tests negative, no individual tests are run: All individuals are declared negative. However, if the group tests positive, the individual without disease must get a negative result during the second-stage individual testing in order to be declared disease-free. Note that the specificity of the group testing is the same as the probability that the person without the disease will be declared negative after passing through both stages of the group testing protocol. As such, using Eqs. (4) and (5), the specificity of the group testing is given by

$$1 - \tau + a(\eta + \tau - 1) + [\tau - a(\eta + \tau - 1)]\eta = 1 - (1 - \eta)[(1 - a)\tau + a(1 - \tau)] > \eta \tag{6}$$

since the factor within square brackets on the right hand side of Eq. (6) is less than one. Thus, a group testing protocol increases specificity compared to individual testing. This reduces the probability of false positive results—a happy achievement. However, as we shall see in the next Subsection, it comes at the cost of increasing the probability of false negative results, which is highly risky in the COVID-19 application.

4.2. Sensitivity of group testing

A person with disease exhibits a positive result if and only if the group tests positive and so does the individual. Hence, The probability that a person with disease exhibits a positive result is τ^2, which is less than τ. This is a cause for concern, since the probability of false negative, $1-\tau^2$, is quite high.

A possible remedy is to run a follow-up group testing when the first group test is negative. Then the sensitivity of the group testing protocol, with two group-tests performed sequentially and if either group-test is positive then testing all group members individually, is the probability that a patient with disease shows a positive test result, and is given by

$$\tau^2 + (1-\tau)\tau^2 = \tau^2[2-\tau] > \tau^2 \tag{7}$$

For COVID-19 test, with $\tau = .7$, a single group test has sensitivity $\tau^2 = .49$; but a second follow-up group testing protocol has an overall group-test sensitivity $\tau^2(2-\tau) = .637$, which is tolerable. If a third follow-up group testing is used, then the sensitivity further rises to $\tau^2(3-3\tau+\tau^2) = .681$. Of course, while multiple follow-up group testing will increase sensitivity (though it will never exceed τ), it will also reduce savings. Therefore, let us settle on the protocol of *at most two group testings* before the group is found to be either positive (at least once) or declared to be negative. For this protocol, the sensitivity is given by the left hand side of Eq. (7). We leave it to the reader to study the group testing protocol that allows at most three group testings.

Suppose that a group has at least one COVID-19 patient, which happens with probability $\theta_g = 1-(1-\theta)^g$. Following the "two-group-testing protocol," the group's test will be positive with probability

$$1-(1-\tau)^2 = \tau(2-\tau) \tag{8}$$

Similarly, if a group has no COVID-19 patient, which happens with probability $(1-\theta)^g$, the probability that this group's test will be positive (falsely) is

$$1-\eta^2 \tag{9}$$

4.3. Minimize the number of tests

For the two-group-testing protocol, using Eqs. (8) and (9), the number of tests *per worker* is $2/g + [\theta_g\tau(2-\tau) + (1-\theta)^g(1-\eta^2)]$, which simplifies to

$$\frac{2}{g} + \tau(2-\tau) - (1-\theta)^g[\tau(2-\tau) + \eta^2 - 1] \tag{10}$$

Consequently, to determine the optimum group size we may drop the constant $\tau(2-\tau)$ from Eq. (10) and solve the minimization problem:

$$\min_{1\le g\le N}\left\{\frac{2}{g} - (1-\theta)^g[\eta^2 + \tau(2-\tau) - 1]\right\}$$

Analogously to the method of solution in case of a perfect test, here in the imperfect test case with two-group-testing protocol, given θ, the optimal group size $g^{\#}(\theta)$ is the largest g such that

$$i(i-1)\theta(1-\theta)^{i-1} < \frac{2}{\eta^2 + \tau(2-\tau) - 1}, \text{ for all } i = 3, 4, \ldots, g \tag{11}$$

The R codes to compute $g^{\#}$ are given in the Annexure. Note that the optimal $g^{\#}$ is a non-increasing step function of θ. For example, $g^{\#}(0.05) = 10$, $g^{\#}(0.02) = 16$, $g^{\#}(0.01) = 23$, $g^{\#}(0.005) = 33$. See more details in Figure 3, where the thick (black) curve shows $g^{\#}$ and the thinner (green) curve shows g^{*} already depicted in Figure 1 corresponding to the perfect test and repeated here for easy comparison. The optimal group size in the imperfect test case, with the two-group-testing protocol, is about 40–80% larger than the optimal group size in the perfect test case.

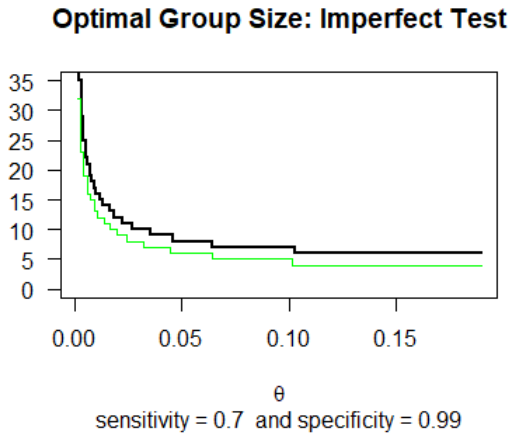


Figure 3: For an imperfect test, the optimal group size $g^{\#}(\theta)$ is shown by a thick, black curve. The thin, green curve shows g^{*} for the perfect test as in Figure 1

4.4. Maximize the percentage savings

For an imperfect test, with the two-group-testing protocol and fixed group size g, the percentage savings is given by

$$\text{Savings (ImperfectTest)} = \left([\eta^2 + \tau(2-\tau) - 1](1-\theta)^g - \frac{2}{g} \right) \times 100\% \tag{12}$$

Once again, the percentage savings is maximum when the optimal group size $g^{\#}(\theta)$ is chosen; otherwise, savings is reduced as the group size deviates from the optimum size. See details in Figure 4 (and the R codes in the Annexure).

As it happened in the perfect case, so also in the imperfect case, employers can reduce the number of tests by using groups of 8 workers, so long as prevalence is under 20 percent. Employers who know the prevalence more precisely can further reduce the number of tests by choosing larger group sizes. But if employers underestimate the true prevalence rate and choose larger group sizes, then the number of tests will increase.

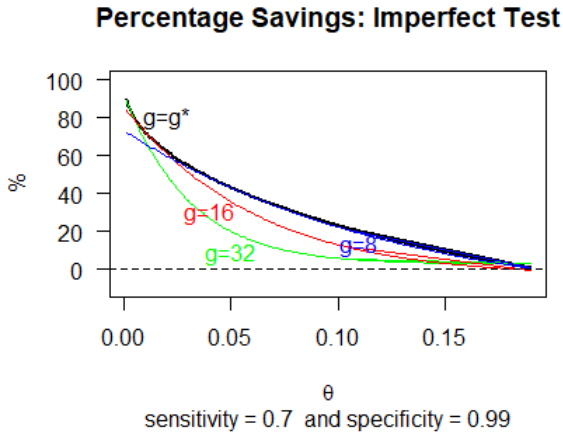


Figure 4: For an imperfect test, the percentage savings for various group sizes: optimum and fixed

5. Further Improvements in Group Testing Methodologies

As mentioned in Section 2, GTMs have undergone tremendous advancement during the last 40 years. We mention some of these results that focus on identifying a few defective items from among many good items, since that is the situation we find ourselves in when a group tests positive in the first-stage. We hope this short review of the GTD research landscape will inspire other researchers pursue this fascinating field of study.

Bush *et al.* (1984) introduced a new class of combinatorial designs with completeness property on t symbols, and used them in group testing to separate defective items from good ones using fewer number of tests than items. In their language: "if a large population of v items has exactly 1 bad item, it can be detected in b tests, where b is only a very tiny fraction of v." Weideman and Raghavarao (1987 a) carried out a systematic study of non-adaptive hypergeometric GTDs for identifying two defectives items from among n, obtaining bounds for n, given the number of tests. Weideman and Raghavarao (1987 b) extended the work to identifying *at most* two defective items.

Das and Roy Choudhury (1987) provides simple methods of forming a small number of groups out of a large number of individuals so that the group test results uniquely (and easily) determine all defective individuals. The methods consist of first encoding suitably all individuals and then forming the groups by using certain properties of the encoders. Hwang and Sos (1987) mentions that even though adaptive or sequential designs (which keep constructing new groups based on the results of previous groups) typically outperform non-adaptive combinatorial designs (which declare the groups at the outset), with the advent of parallel processing, the time-saving feature of non-adaptive designs remains attractive.

Whereas Ding-Zhu and Hwang (1993) develops the conventional disjunct search model, D'yachkov *et al.* (2001) discusses two non-standard models of nonadaptive combinatorial search for a small number of defective elements contained in a finite population in the presence of inhibitors. Hung and Swallow (1999) discusses the robustness of group testing for estimating proportions when the underlying assumptions of no testing errors and independent individuals are violated. Adhikari, Ghosh and Sinha (2001) considered a multi-component

system which can be tested even though the components cannot be tested separately. For a five-component system, they utilized Taguchi's L16 orthogonal array design to identify defective components and to estimate the proportion of defectives. They also designed a sequential experiment which reduced the number of tests from 16 to 12.

Du and Hwang (2006) surveys both sequential pooling designs and nonadaptive group testing. Rao, Rao and Sinha (2006) obtained improved bounds on the number of group tests necessary for both adaptive and nonadaptive GTDs. They established that when a group of 2^n items have *at most two* defective items, any nonadaptive GTD needs at least $2n$ group tests for identifying all defective items, and the optimal nonadaptive GTDs can be constructed using generalized Petersen graphs. In the same context, they presented an adaptive multistage GTD with a maximum of $2n$ group tests, and constructed a family of two-stage adaptive GTDs. Rao, Sinha and Rao (2013) gives a comprehensive review of both adaptive and non-adaptive GTDs, exposing hidden combinatorial and probabilistic challenges and offering a storehouse of unsolved problems.

We leave to the reader to apply these advanced GTDs to the COVID-19 detection problem. We describe below a situation where a group that tested positive in the first-stage is subjected to another group testing before conducting individual tests.

5.1. A second-stage group testing

According to our proposed two-group-testing protocol, if two successive group tests are both negative, we declare the entire group negative. But when the first or the second group-test turns out positive, then we test every member of the group individually. As mentioned in Section 3, if the group size is big (say, bigger than 10), it is more efficient to apply GTMs on this group, before moving on to individual testing, and harvest additional savings.

Suppose that a university estimated the prevalence of COVID-19 among its 25,000 students, faculty and staff members to be very low, say 0.1% or 0.001, and accordingly chose $g = 49$ (using Eq. (11)). The number of infected individuals in each group of 49 has an approximate Poisson(0.049) distribution. Hence, about 25 groups are likely to test positive; and within each group that tests positive, with a very high probability (about .9757) exactly one member is positive, with a low probability (about .0239) two members are positive, and with a negligible probability (about .0004) more than two members are positive.

Therefore, instead of testing individually every member of a positively tested group, as our proposed first-stage two-group-test protocol recommends, the employer can update the prevalence (for each positively tested group) to be $1/25 = .04$, or more conservatively to be $2/25 = .08$, and adopt a second-stage group testing with $g = 7$ (again using Eq. (11) or Figure 3). Then with a high probability (.9757) it will take 13 group tests to identify one positive subgroup, or with a low probability (.0239) it will take 12 group tests to identify two positive subgroups, or with a negligible probability (.0004) it will take at most 11 group tests to identify more than two positive subgroups.

It may not be worth applying the same logic to the one (or two or three) subgroups of size 7 that tested positive during the second-stage group testing because an individual's nasal swab specimen can be divided into about four or five sub-specimens while still retaining tractability of COVID-19 infection. It will be best to simply apply an individual test on every

member of the latest positively tested subgroup using the fifth sub-specimen, if any, or collect fresh new specimens from them. Thus, overall the 49 individual tests are reduced *on average* to about $.9757 * (13 + 7) + .0239 * (12 + 2 * 7) + .0004 * (11 + 3 * 7) = 20.15$ tests, achieving a 59% reduction on costs after first-stage testing.

To summarize, the first stage two-group-test protocol requires about $2 * \lceil 25000/49 \rceil - 25 = 997$ tests. Thereafter, (i) a second stage individual testing would involve $25 * 49 = 1225$ tests, for a total of 2222 tests. But (ii) a second stage two-group-test followed by a third stage individual testing would involve about $25 * 20.15 = 504$ tests. Thus, strategy (i) saves 91% compared to testing all 25000 people individually, and strategy (ii) additionally saves 721 tests, or 2.9%. Finally, strategy (ii), when compared to strategy (i), harvests a $721/2222 = 32.4\%$ saving across all tests in all stages combined.

6. Implications of Group Testing for COVID-19 Detection

Economic shutdown causes a tremendous loss in GDP (gross domestic product). On the other hand, reopening the economy necessitates frequent and mass-scale testing of all employees. Although the latter cost is relatively smaller than the former cost, if employers have to bear all the cost, then a cost effective way to carry out these tests is imperative. Group testing, with a follow-up test for the group if the first group test is negative (or when two group tests are negative), is a promising methodology to mitigate the challenge.

We have shown significant savings when prevalence is known with a high degree of certainty. We have also demonstrated that the quality of diagnostics made using the group testing protocol with two-group-tests is reasonably close to that of the individual testing protocol. The quality can be increased slightly (at the cost of reducing savings) by using a three-group-tests protocol before declaring the group negative. Also, there can be a second-stage group testing followed by individual testing to reap additional savings. Until testing sensitivity improves, employers should implement group testing together with other strategies, such as symptom monitoring and contact tracing.

Uncertainty about prevalence reduces savings only marginally if prevalence can be estimated well. As more information becomes available to estimate prevalence better, higher savings are anticipated. Also, as the sensitivity of the test improves—when researchers discover better tests—the optimal group sizes will rise yielding additional savings. Moreover, such better tests will reduce complexity of testing and discomfort to the employees. For instance, saliva-based tests may become a suitable alternative to nasal swabs. Let us look forward to those better days until a vaccine or a cure becomes available and affordable.

Acknowledgements

I thank Dr. Honglang Wang for suggesting some improvements on a first draft. From an anonymous reviewer, I received a treasure trove of references, which give a bird's eye view of the entire landscape of advancement in GTMs. I express my heart-felt thanks to this generous reviewer. Thanks are due to Professor Bikas K. Sinha for providing a hard-to-find reference on short notice. I am grateful to the Chair Editor for his patience, guidance, encouragement and instructions at multiple stages.

References

Adhikari, S., Ghosh, J. K. and Sinha, Bikas K. (2001). Group testing for identification of defective components: An application of orthogonal arrays. *Statistics and Applications*, **3(1-2)**, 89–96.

Aprahamian, H., Bish, E. K. and Bish, D. R. (2017). Adaptive risk-based pooling in public health screening. *Institute of Industrial and Systems Engineers (IISE) Transactions*, **50(9)**, 753–766. https://doi.org/10.1080/24725854.2018.1434333

Bush, K. A., Federer, W. T., Pesotan, H. and Raghavarao, D. (1984). New combinatorial designs and their applications to group testing. *Journal of Statistical Planning and Inference*, **10**, 335–343.

Conger, K. (2020). Testing pooled samples for COVID-19 helps Stanford researchers track early viral spread in Bay Area. *Stanford Medicine News Center*. https://med.stanford.edu/news/all-news/2020/04/testing-pooled-samples-to-track-early-spread-of-virus

Das, M. N. and Roy Choudhury, D. (1987). On problems of search using group testing. *Sankhya B*, **49**, 137–147.

Ding-Zhu, D. and Hwang, F. K. (1993). *Combinatorial Group Testing and Its Applications*. Singapore: World Scientific.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, **14(4)**, 436–440.

Du, D. and Hwang, F. (2006). *Pooling Designs and Nonadaptive Group Testing*. Boston: Twayne Publishers.

D'yachkov, A. G., Macula, A. J., Torney, D. C. and Vilenkin, P. A. (2001). Two models of nonadaptive group testing for designing screening experiments. In: Atkinson A. C., Hackl P., Müller W. G. (eds) *mODa 6 – Advances in Model-Oriented Design and Analysis*, pp 63–75. Contributions to Statistics. Physica, Heidelberg. https://doi.org/10.1007/978-3-642-57576-1_8

Hung, M. and Swallow, W. H. (1999). Robustness of group testing in the estimation of proportions. *Biometrics*, **55(1)**, 231–237. doi:10.1111/j.0006-341X.1999.00231.x

Hwang, F. K. (1972). A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, **67(339)**, 605–608. doi:10.2307/2284447.

Hwang, F. K. and Sos, V. T. (1987). Non-adaptive hypergeometric group testing. *Studia Scientiarum Mathematicarum Hungarica*, **22**, 257–263.

Keeler E. and Berwick D. (1976). Effects of pooled samples. *Health Laboratory Science*, **13(2)**, 121–128. https://www.ncbi.nlm.nih.gov/pubmed/1270261

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rao, P. S. S. N. V. P., Rao, S. B. and Sinha, Bikas K. (2006). Improved bounds for group testing designs. *Journal of Statistical Planning and Inference*, **136(1)**, 260–269.

Rao, S. B., Sinha, Bikas K. and Rao, Prasad. (2013). Group testing designs: A combinatorial marvel. *Statistics and Applications*, **11(1-2)**, 111–126.

Schisterman, E. F., and Vexler, A. (2008). To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatric and Perinatal Epidemiology*. https://doi.org/10.1111/j.1365-3016.2008.00956.x

Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal*, **38(5)**, 1179–1252.

doi:10.1002/j.1538-7305.1959.tb03914.x

Sobel, M. and Elashoff, R. M. (1975). Group testing with a new goal, estimation. *Biometrika*, **62(1)**, 181–193. doi:10.1093/biomet/62.1.181

Sterrett, A. (1957). On the detection of defective members of large populations. *The Annals of Mathematical Statistics*, **28 (4)**, 1033–1036. doi:10.1214/aoms/1177706807

The White House, Centers for Disease Control and Prevention, and Food and Drug Administration. (2020). *Testing Blueprint: Opening Up America Again.* https://www.whitehouse.gov/wp-content/uploads/2020/04/Testing-Blueprint.pdf

Ungar, P. (1960). The cutoff point for group testing. *Communications on Pure and Applied Mathematics*, **13**, 49–54.

Weideman, C. A. and Raghavarao, D. (1987 a). Some optimum nonadaptive hypergeometric group testing designs for identifying two defectives. *Journal of statistical Planning and Inference*, **16**, 55–61.

Weideman, C. A. and Raghavarao D. (1987 b). Nonadaptive hypergeometric group testing designs for identifying at most two defectives. *Communications in Statistics*, **16 (10)**, 2991–3006.

Wikipedia Contributors (no date). *Group Testing. Wikipedia, The Free Encyclopedia.* Retrieved 13 June 2020. https://en.wikipedia.org/wiki/Group_testing

Yang Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., Zhang, M., Wang, Z., Xing, L., Wei, J., Peng, L., Wong, G., Zheng, H., Liao, M., Feng, K., Li, J., Yang, Q., Zhao, J., Zhang, Z., Liu, L. and Liu, Y. (2020). Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. *MedRxiv*. https://doi.org/10.1101/20 20.02.11.20021493

ANNEXURE

Here we document the R codes used to draw Figures 1–4.

```
### Fig 1. Optimal group size (Perfect test)
gs=function(the){b=3
  for (i in 4:40){
    if (i*(i-1)*(1-the)^(i-1)*the<1){b=i}
    else{break} }
b}

theta=seq(0.001, .123, .001)
gsp=rep(3,123)
for (j in 1:123){gsp[j]=gs(j/1000)}
plot(theta, gsp, type='s', las=1, lwd=2, ylim=c(0,35),
  ylab='', xlab=expression(theta),
  main='Optimal Group Size: Perfect Test')

### Fig 2. Savings (Perfect test)
thet=c(theta,seq(.124,.400,.001))
save=100*c((1-theta)^gsp-1/gsp, (1-seq(.124,.400,.001))^3-1/3)
plot(thet, save, type='l', las=1, lwd=2, ylim=c(-10,100),
```

```
   ylab='%', xlab=expression(theta),
   main='Percentage Savings: Perfect Test')

lines(thet, 100*((1-thet)^12-1/12), col='green')
lines(thet, 100*((1-thet)^8-1/8), col='red')
lines(thet, 100*((1-thet)^4-1/4), col='blue')
abline(c(0, .400),c(0,0), lty=2)
text(0.1, 8,'g=12', col='green'); text(0.09, 25,'g=8', col='red')
text(0.20, 10,'g=4', col='blue'); text(0.25, 20,'g=g*', col='black')

### Fig 3. Optimal group size (Imperfect test)
tau=.70; eta=.99;
(mult=eta^2+tau*(2-tau)-1)
gs=function(the){b=3
  for (i in 6:80){
    if (i*(i-1)*(1-the)^(i-1)*the<2/mult){b=i}
    else{break} }
b}

theta=seq(0.001, .190, .001)
gst=rep(6,190)
for (j in 1:190){gst[j]=gs(j/1000)}
plot(theta, gst, type='s', las=1, lwd=2, ylim=c(0,35),
  ylab='', xlab=expression(theta),
  main='Optimal Group Size: Imperfect Test',
  sub=paste('sensitivity =', tau, ' and specificity =', eta) )
par(new=TRUE)
plot(theta[1:123], gsp, type='s', las=1, ylim=c(0,35),
   xlab='', ylab='', xaxt='n', yaxt='n', lty=1, col='green')


### Fig 4. Savings (Imperfect test)
thet=c(theta,seq(.104,.190,.001))
save=100*c(1-tau*(2-tau)+(1-theta)^gst*mult-2/gst,
   1-tau*(2-tau)+(1-seq(.104,.190,.001))^6*mult-1/3)
plot(thet, save, type='l', las=1, lwd=2, ylim=c(-10,100),
   ylab='%', xlab=expression(theta),
   main='Percentage Savings: Imperfect Test',
   sub=paste('sensitivity =', tau, ' and specificity =', eta) )

lines(thet, 100*(1-tau*(2-tau)+(1-thet)^32*mult-1/16), col='green')
lines(thet, 100*(1-tau*(2-tau)+(1-thet)^16*mult-1/8), col='red')
lines(thet, 100*(1-tau*(2-tau)+(1-thet)^8*mult-1/4), col='blue')
abline(c(0, .200),c(0,0), lty=2)
text(0.05, 8,'g=32', col='green'); text(0.04, 30,'g=16', col='red')
text(0.11, 13,'g=8', col='blue');  text(0.02, 80,'g=g*', col='black')
```

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 295-306

A Mixture of Regressions Model of COVID-19 Death Rates and Population Comorbidities

M. Maleki[1], G. J. McLachlan[2], R. Gurewitsch[3], M. Aruru[4,5], S. Pyne[3,5,6]

[1]*Faculty of Mathematics and Statistics, Department of Statistics, University of Isfahan, Isfahan, Iran*
[2]*Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia*
[3]*Public Health Dynamics Laboratory, University of Pittsburgh, Pittsburgh, USA*
[4]*Program Evaluation and Research Unit, University of Pittsburgh School of Pharmacy, Pittsburgh, USA*
[5]*Health Analytics Network, PA, USA*
[6]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA*

Received: 08 July 2020; Revised: 22 July 2020; Accepted: 24 July 2020

Abstract

As the COVID-19 pandemic spread worldwide, it has become clearer that prevalence of certain comorbidities in a given population could make it more vulnerable to serious outcomes of that disease, including fatality. Indeed, it might be insightful from a health policy perspective to identify clusters of populations in terms of the associations between their prevalent comorbidities and the observed COVID-19 specific death rates. In this study, we described a mixture of polynomial time series (MoPTS) model to simultaneously identify (a) three clusters of 86 U.S. cities in terms of their dynamic death rates, and (b) the different associations of those rates with 5 key comorbidities among the populations in the clusters. We also described an EM algorithm for efficient maximum likelihood estimation of the model parameters.

Keywords: Mixture of regressions; EM; Death rate; Comorbidities; COVID-19.

1. Introduction

COVID-19 is an acute, respiratory disease due to novel coronavirus SARS-CoV-2, and similar to previous diseases such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), it can lead to respiratory failure and death (CDCa, 2019). In the absence of any proven medical treatment for COVID-19, and in the face of acute shortage of critical care capacity, including ventilators, during peaks of incidence, as was witnessed in some countries such as Italy, it is of critical importance for any local administration to evaluate and stratify the risk levels in terms of its population comorbidities.

As per current understanding (CDCb, 2020), individuals at-risk for severe illness from COVID-19 include people who are or have: (a) 65 years and older, (b) living in nursing home or long-term care facilities, (c) chronic lung disease or moderate to severe asthma, (d) serious

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

heart conditions, (e) immuno-compromised, (f) severe obesity, (g) diabetes, (h) chronic kidney disease undergoing dialysis, (i) hemoglobin disorders, and (j) liver disease.

Early retrospective studies from China, Italy, and the U.S. revealed the prevalence of comorbidities in exacerbation of disease resulting in poor outcomes. Retrospective analysis of 1,590 confirmed COVID-19 cases between December 11, 2019 and January 31, 2020 across 31 provinces/autonomous regions in China indicated that 25.1% of the cases had at least one comorbidity. Specific prevalence included hypertension (16.9%), other cardiovascular diseases (53.7%) and diabetes (8.2%), whereas asthma, COPD, kidney diseases, immunodeficiency were below 1% (Guan *et al*., 2020). In a larger study across China, 72,314 patient records until February 11, 2020 were examined of which 44,672 were confirmed cases and further evaluated for comorbidities. (Feng *et al*., 2020) More than 50% of the records were missing data on comorbidities. Of available data on confirmed cases, comorbidities included hypertension (12.8%), diabetes (5.3%), and less than 4% for cardiovascular disorders, chronic respiratory disease, and cancers.

The case-fatality ratio (CFR) is a useful indicator of survival prognosis among critically ill patients. During the peak of COVID-19 infections in China, the highest CFR was 14.8% among patients over age 80 years. CFR for patients with comorbid conditions was 10.5% for those with cardiovascular disease, 7.3% for diabetes, 6.3% for chronic respiratory disease, 6% for hypertension and 5.6% for cancer, when compared to 0.9% for patients with no comorbid conditions. CFR for critical cases was 49% indicting the need to mitigate risks earlier in the infection. A meta-analysis of 6 COVID-19 studies indicated hypertension, diabetes, COPD, cardiovascular disease, and cerebrovascular disease as major risk factors for COVID-19 patients (Wang *et al*., 2020). Another meta-analysis of pooled data from 8 studies in China, indicated a higher prevalence of hypertension and diabetes among patients with severe cases of COVID-19 disease (Yang *et al*., 2020).

As the pandemic spread, Italy was next to experience large caseloads and CFR. COVID-19 cases peaked in March to 3000-6500 new cases and 350-900 deaths daily. Comorbidities observed in deceased patients as of April 29, 2020 included hypertension (69.2%), diabetes (31.8%), cardiovascular conditions (22%), COPD (16.9%), obesity (11.6%) among others (Statista, 2020). In China and Italy, hypertension, diabetes, cardiovascular conditions, chronic respiratory disease accounted for the top comorbidities associated with severe outcomes (Chen et al., 2020).

The first COVID-19 case in the U.S. was reported from Washington State on January 31, 2020. Since then, the virus spread to the east coast and the rate of infections in New York exceeded every other state. As of April 30, 2020, New York had more than 30% of all the U.S. cases. A retrospective analysis of 5,700 records of patients hospitalized with COVID-19 in the New York City area revealed that while 6.1% of the patients had no comorbidities, and 6.3% had 1 comorbidity, as high as 88% had more than one comorbidity. This study further indicated that older patients, men, and those with pre-existing hypertension and/or diabetes were highly prevalent among those hospitalized for COVID-19 (Richardson et al., 2020). This pattern was similar to observations from China and Italy. In yet another observational cohort study at two New York hospitals, as of April 28, 2020, 39% of patients who were critically ill with COVID-19 had died, and similar to studies in China and Italy, hypertension and cardiopulmonary comorbidities were found to be associated with increased mortality (Cummings *et al*., 2020).

In this study, we modeled COVID-19 death rates and key comorbidities for $N = 86$ U.S. cities. We used a mixture of regressions modeling approach to simultaneously identify the

clusters of cities in terms of their dynamic death rates, which had different associations with the prevalence of 5 key comorbidities among the populations in those clusters. We provided an EM algorithm for efficient maximum likelihood estimation of the model parameters. While the use of mixture of regressions models is well-known, [*e.g.*, Jones and McLachlan (1992) and references in McLachlan and Peel (2000)], we think that the combination of finite mixtures, linear regression and polynomial modeling of time series makes our approach ideally suited for the present problem. We described our mixture of polynomial time series (MoPTS) model and the datasets in the next section, followed by the modeling results and discussion.

2. Materials and Methods

2a. Datasets

Comorbidities Data: The Centers for Disease Control and Prevention (CDC) has conducted each year, since 1984, the Behavioral Risk Factor Surveillance System (BRFSS), a cross-sectional representative telephone survey of U.S. adults (18 years or above) regarding their health conditions and behaviors. BRFSS now collects data from all 50 U.S. states and the districts of Columbia, Guam, and Puerto Rico. Both landline and cellular phone numbers are sampled in the survey and aggregated results are presented each year (CDCc, 2020). Health risk questions range from infectious diseases to chronic diseases and behaviors. The fixed core of the survey consists of a standard set of questions used by all states to include demographics and current health behaviors, *e.g.* tobacco use. In this study, we focus on $p = 5$ comorbidities collected by BRFSS and reported at the county level: diabetes, obesity, coronary heart disease (CHD), hypertension, and chronic obstructive pulmonary disease (COPD).

COVID-19 Data: Based on cumulative COVID-19 deaths data from the Johns Hopkins Coronavirus Resource Center (COVID-19 U.S. cases, 2020), we compiled time series data on daily deaths from the disease for the U.S. counties by their 5-digit FIPS code or county name. Since a single county may consist of multiple cities, we include the list of all city labels within each aggregate group to represent a greater metropolitan area. While we denote such metropolitan areas as "cities" for convenience, we identify them by their underlying county FIPS codes. In this study, we used the data for $N = 86$ cities across the U.S. which had at least 100 reported deaths from COVID-19 by May 31, 2020. We excluded New York City from our clustering as it is an outlier in terms of uniquely high incidence and extreme population heterogeneity, which could skew the model with its own singleton cluster. For each included city j $(j = 1, \dots, N)$, the daily number of deaths is given as a time series of $n_j = 93$ time points, over the period starting from 29 February, 2020, up to 31 May, 2020. For compatibility, the COVID-19-specific death rate was standardized as the number of deaths per 100,000 people using population data of the cities from the U.S. Census Bureau (U.S. Census Bureau).

2b. Mixture of Regressions Model

In this study, we used a mixture of polynomial time series (MoPTS) model to cluster the above-mentioned $(N = 86)$ cities in terms of their dynamic death rates into an optimal number (g) of clusters (identified by the mixture model components $C_1, \dots C_g$), and investigate the differences among the estimated regression coefficients of 5 known comorbidities (given as 5 static city-specific covariates) across those clusters.

Let Y_{jt} be the random variable representing death rate at time t $(t = 1, \dots n_j)$ for city j $(j = 1, \dots, N)$. It is assumed that the mean $\mu_i(t)$ of each time series Y_{jt} in the i^{th} cluster C_i follows

$Y_{jt} - \mu_i(t) \overset{iid}{\sim} N(0, \sigma_i^2)$ in C_i $(i = 1, \dots, g)$ for all values of $t = 1, \dots n_j$ and $j = 1, \dots, N$. Further, let $\boldsymbol{x}_j$ be a p-dimensional vector that describes static values of p comorbidities for the city j $(j = 1, \dots, N)$. Also, let Z_j is a component-indicator random variable to determine that city j belongs to the component (or cluster) C_i with probability $Pr(Z_j = i) = \pi_i$ such that $\pi_i > 0$ and $\sum_{i=1}^{g} \pi_i = 1$ for $i = 1, \dots, g$ and $j = 1, \dots, N$. Thus, our mixture of regressions on time series modeled with order D polynomial and covariates has the form:

$$Y_{jt} = \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \sum_{d=1}^{D} \beta_{id} t^d + \varepsilon_{jt}; \;\; \varepsilon_{jt} \sim N(0, \sigma_i^2) \text{ with probability } \pi_i; i = 1, \dots, g \qquad (1)$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ip})^\top$ are regression coefficients corresponding to the p static covariates, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{iD})^\top$ are coefficients corresponding to the polynomial time series.

Let $\boldsymbol{Y}_j = \left(Y_{j0}, Y_{j1}, \dots, Y_{jn_j}\right)^\top$ arise from a g-component mixture of order D polynomial time series with covariates model, denoted by MoPTS(D, g), if it can be characterized by the conditional density function

$$f(y_{jt} | \boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i) = \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2), \;\; j = 1, \dots, N; t = 1, \dots, n_j, \qquad (2)$$

where $\boldsymbol{t}_D = (t, t^2, \dots, t^D)^\top$ and $\boldsymbol{\theta}_i = (\boldsymbol{\alpha}_i^\top, \boldsymbol{\beta}_i^\top, \sigma_i^2)^\top$.

Under the characterization (2), we can further characterize $\boldsymbol{Y}_j$ via the joint density function

$$f(\boldsymbol{y}_j | \boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i) = \sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2), \;\; j = 1, \dots, N. \qquad (3)$$

Using the characterization (3), we can write the log-likelihood of an IID sample $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_N$ as

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{N} \log \left(\sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2) \right), \qquad (4)$$

where the parameter $\boldsymbol{\theta} = \left(\pi_1, \dots, \pi_{g-1}, \boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_g^\top\right)^\top$ is traditionally estimated by maximization of $\ell(\boldsymbol{\theta})$ given by (4). However, in the absence of an analytical solution, we used an EM–type algorithm (Dempster *et al*., 1977; McLachlan and Peel 2000). Using log-likelihood criteria, we chose $D = 3$ and $g = 3$, *i.e*., our final fit produced a 3-component normal mixture regression model with the dependence of the mean on time modeled by a cubic polynomial.

2c. The observed information matrix

In this section, the observed (Fisher) information matrix of MoPTS, defined as $\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$, where $\ell(\boldsymbol{\theta}) = \sum_{j=1}^{N} \ell_j(\boldsymbol{\theta})$, $\ell_j(\boldsymbol{\Theta}) = \log\left(\sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^\top \boldsymbol{x}_j + \boldsymbol{\beta}_i^\top \boldsymbol{t}_D, \sigma_i^2)\right)$, is obtained.

It is well known that, under mild regularity conditions, the covariance matrix of the ML estimates $\widehat{\boldsymbol{\Theta}}$ can be approximated by the inverse of $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y})$. Thus, following Basford *et al*. (1997), we approximated $\mathbf{J}(\boldsymbol{\Theta}|\boldsymbol{y})$ as

$$\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{y}) \approx \sum_{j=1}^{N} \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^{\top}, \tag{5}$$

where $\hat{\boldsymbol{s}}_j = \frac{\partial \ell_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, and consider now the vector $\hat{\boldsymbol{s}}_j$ which is partitioned into components corresponding to all the parameters in $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{s}}_j = \left(\hat{s}_{j,\pi_1}, \ldots, \hat{s}_{j,\pi_{g-1}}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\alpha}_1}^{\top}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\alpha}_g}^{\top}, \hat{\boldsymbol{s}}_{j,\boldsymbol{\beta}_1}^{\top}, \ldots, \hat{\boldsymbol{s}}_{j,\boldsymbol{\beta}_g}^{\top}, \hat{s}_{j,\sigma_1^2}, \ldots, \hat{s}_{j,\sigma_g^2}\right)^{\top},$$

where its coordinate elements for $i = 1, \ldots, g$ are given by

$$\hat{s}_{j,\pi_i} = \frac{\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i) - \phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_g)}{f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta})}, \quad \hat{\boldsymbol{s}}_{j,\boldsymbol{\Delta}_i} = \frac{\pi_i D_{\boldsymbol{\Delta}_i}\left(\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right)}{f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta})},$$

where $f(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i \phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)$ and $\boldsymbol{\Delta}_i = \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ and σ_i^2, for which $\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i) = \sum_{i=1}^{g} \pi_i \prod_{t=1}^{n_j} \phi(y_{jt}; \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j + \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D, \sigma_i^2)$ and $D_{\boldsymbol{\Delta}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \partial\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)/\partial\boldsymbol{\Delta}_i$. Thus,

$$D_{\boldsymbol{\alpha}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} s_{xy} \exp\left(-\frac{1}{2}s_{yy}\right);$$

$$D_{\boldsymbol{\beta}_i}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} s_{ty} \exp\left(-\frac{1}{2}s_{yy}\right);$$

$$D_{\sigma_i^2}\left[\phi_{n_j}(\boldsymbol{y}_j|\boldsymbol{x}_j, \boldsymbol{\theta}_i)\right] = \left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^{n_j} \frac{1}{2}\sigma_i^{2n_j-2}\left\{s_{yy} - n_j \exp\left(-\frac{1}{2}s_{yy}\right)\right\};$$

where $s_{yy} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)^2$, $s_{xy} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}\boldsymbol{x}_j(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)$ and $s_{ty} = \frac{1}{\sigma_i^2}\sum_{t=1}^{n_j}\boldsymbol{t}_D(y_{jt} - \boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j - \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D)$.

2d. ML estimation

In this section, we develop an efficient EM algorithm for maximum likelihood (ML) estimation of the parameters of the MoPTS model, using an incomplete-data framework. To compute this procedure, we use the following hierarchical representation:

$$Y_{jt}|\boldsymbol{x}_j, Z_{ji} = 1 \overset{iid}{\sim} N(\boldsymbol{\alpha}_i^{\top}\boldsymbol{x}_j + \boldsymbol{\beta}_i^{\top}\boldsymbol{t}_D, \sigma_i^2), \quad i = 1, \ldots, g.$$

$$Pr(Z_{jt} = i) = \pi_i \ (i = 1, \ldots, g; \ j = 1, \ldots, N; t = 1, \ldots, n_j), \tag{6}$$

Let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, , \dots, \boldsymbol{y}_N^\top)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1^\top, , \dots, \boldsymbol{x}_N^\top)^\top$ and $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \dots, \boldsymbol{z}_N^\top)^\top$ for which $\boldsymbol{z}_j = (z_{j1}, \dots, z_{jg})^\top$; $j = 1, \dots, N$, so considering the complete data $\boldsymbol{y}_c = (\boldsymbol{y}^\top, \boldsymbol{x}^\top, \boldsymbol{z}^\top)^\top$ and using the hierarchical representation in (6) of the MoPTS(D, g) model, the complete data likelihood, ignoring the constant term, is given by

$$\begin{aligned}\mathcal{P}_c(\boldsymbol{\theta}|\boldsymbol{y}_c) &= \sum_{j=1}^{N}\sum_{i=1}^{g}\sum_{t=1}^{n_j} I_{[i]}(z_{jt})\{\log\pi_i + \log f(y_{jt}|\boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i)\} \\ &= \sum_{i=1}^{g}\log\pi_i \sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt}) - \sum_{i=1}^{g}\log\sigma_i \sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}\sum_{j=1}^{N}\sum_{t=1}^{n_j} I_{[i]}(z_{jt})(y_{jt} - \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j - \boldsymbol{\beta}_i^\top\boldsymbol{t}_D)^2 \end{aligned} \quad (7)$$

where $I_{[i]}(z_{jt})$ is 1 if y_{jt} belongs to the ith component and zero otherwise, and $f(\cdot\,|\boldsymbol{x}_j, Z_j = i, \boldsymbol{\theta}_i)$ was defined in (2).

Starting from some initial value $\boldsymbol{\theta}^{(0)}$, the conditional expectation of (7), given the observed data, was computed using $\boldsymbol{\theta}^{(k)}$ for $\boldsymbol{\theta}$, can be written as

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{g}\log\pi_i \sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)}) - \sum_{i=1}^{g}\log\sigma_i \sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{g}\frac{1}{\sigma_i^2}\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\boldsymbol{\theta}^{(k)})(y_{jt} - \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j - \boldsymbol{\beta}_i^\top\boldsymbol{t}_D)^2, \end{aligned} \quad (8)$$

where

$$\gamma_{jit}(\boldsymbol{\theta}) = \frac{\pi_i\phi(y_{jt}; \boldsymbol{\alpha}_i^\top\boldsymbol{x}_j + \boldsymbol{\beta}_i^\top\boldsymbol{t}_D, \sigma_i^2)}{\sum_{s=1}^{g}\pi_i\phi(y_{jt}; \boldsymbol{\alpha}_s^\top\boldsymbol{x}_j + \boldsymbol{\beta}_s^\top\boldsymbol{t}_D, \sigma_s^2)}. \quad (9)$$

The posterior probability is the conditional probability that y_{jt} belongs to the i^{th} component given $\boldsymbol{x}_j$ and $\boldsymbol{y}_{j(t)}$ for $i = 1, \dots, g$; $j = 1, \dots, N$ and $t = 1, \dots, n_j$. To perform the M-step, under the restriction $\sum_{i=1}^{g}\pi_i = 1$, by constructing the Lagrangian $\Lambda(\boldsymbol{\theta}, \lambda) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \lambda(\sum_{i=1}^{g}\pi_i - 1)$, we maximize (8) in the EM algorithm by solving the equation corresponding to the first-order condition $\nabla\Lambda(\boldsymbol{\theta}, \lambda) = \boldsymbol{0}$, where ∇ is the gradient operator, which yields the following updates

$$\hat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})}{N}, \quad (10)$$

$$\hat{\boldsymbol{\alpha}}_i^{(k+1)} = \left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{x}_j\boldsymbol{x}_j^\top\right]^{-1}\left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{x}_j\left(y_{jt} - \boldsymbol{t}_D^\top\hat{\boldsymbol{\beta}}_i^{(k)}\right)\right], \quad (11)$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)} = \left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{t}_D\boldsymbol{t}_D^\top\right]^{-1}\left[\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}(\hat{\boldsymbol{\theta}}^{(k)})\boldsymbol{t}_D\left(y_{jt} - \boldsymbol{x}_j^\top\hat{\boldsymbol{\alpha}}_i^{(k+1)}\right)\right], \quad (12)$$

$$\hat{\sigma}_i^{2(k+1)} = \frac{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)\left(y_{jt}-\boldsymbol{x}_j^{\top}\hat{\boldsymbol{\alpha}}_i^{(k+1)}-\boldsymbol{t}_D^{\top}\hat{\boldsymbol{\beta}}_i^{(k+1)}\right)^2}{\sum_{j=1}^{N}\sum_{t=1}^{n_j}\gamma_{jit}\left(\hat{\boldsymbol{\theta}}^{(k)}\right)}. \quad (13)$$

The steps of the above algorithm are iterated until a suitable convergence rule is satisfied, e.g., $\left|\ell\left(\hat{\boldsymbol{\theta}}^{(k+1)}\right)/\ell\left(\hat{\boldsymbol{\theta}}^{(k)}\right)-1\right| \leq \varepsilon$ for a pre-decided tolerance ε.

3. Results and Discussion

Based on the log-likelihood model selection criteria, our optimal 3-component MoPTS model with the dependence of the mean death rate on time was modeled by a cubic polynomial. Thus, we identified 3 clusters of U.S. cities specified by the different associations between their death rates and comorbidities. Figure 1 depicts the time series of each cluster along with the fitted cubic polynomial models. The differences in the death rates (y-axes) across the 3 clusters can be noted. Table 1 lists the counties and states corresponding to these 86 cities, and their cluster memberships. Table 2 shows the parameter coefficients of the MoPTS $(D = 3, g = 3)$ model as estimated by the EM algorithm, along with the corresponding standard deviations.

Cluster 1 contains 21 cities concentrated mostly in the northeastern part of the U.S. (Figure 2a) with death rates associated with diabetes (α_{11}) and hypertension (α_{14}). Cluster 2 consists of 29 cities (Figure 2b) that are more spatially distributed than those in cluster 1, and has relatively weaker association with CHD (α_{23}), diabetes (α_{21}) and COPD (α_{25}). Cluster 3 (Figure 2c), comprised of 36 cities, is the most geographically heterogeneous, which possibly explains why its association with the comorbidities hypertension (α_{34}) and COPD (α_{35}) are the weakest among the three clusters. Nonetheless, the results of our MoPTS model are in general agreement with the current understanding of the role of comorbidities in COVID-19 outcomes (CDCb, 2020).

While the geographical distribution of the clusters are no doubt represented by the populations therein and the underlying health and behavioral risk factors, environmental exposures, chronic conditions and comorbidities, it is important to avoid the risk of so-called ecological fallacy in attempting to infer individual disease outcomes based on data or results obtained at the level of large cities or counties. Indeed, the aim of our analysis is to provide insights into the existence of multiple patterns by which the current pandemic could affect the death rates in different metropolitan areas or counties in terms of their prevailing comorbidities. Such patterns could vary locally even within the same state. For instance, the disease dynamics as of 31 May 2020, at Allegheny county, which is located in western Pennsylvania and contains the city of Pittsburgh, was set apart (in cluster 3) from the other counties in the same state that lie more to the east and, indeed, share a different dynamic pattern (all of these are in cluster 2).

In the future, we aim to study possible interactions between different comorbidities as well as potential pathways leading from such comorbidities to various COVID-19 outcomes that were observed in diverse populations globally. With further availability of electronic health records, clinical evaluation and other microdata derived from monitoring the progression of the disease, more precise inferences can be drawn. Towards this, we could extend our present approach to time series models such as due to Hajrajabi and Maleki (2019) and Zarrin *et al*. (2019), and also the flexible class of distributions introduced by Hoseinzadeh *et al*. (2019), Moravveji *et al*. (2019) and Maleki *et al*. (2019).

As the pandemic progresses, we will have better understanding of the complex interplay of different comorbidities – acting either singly or in combination – in COVID-19 disease progression, especially of those leading to severe outcomes. Naturally, population differences between diverse geographies and societies might make such associations hard to generalize. Yet, certain putative risk factors, such as Type 2 diabetes or CHD, may be common among populations around the world. Toward this end, healthcare systems may want to develop pandemic resiliency with targeted communications and policies that are aimed at patients grouped by low to high risk categories based on their health, lifestyles, and environments. Such risk stratification would enable healthcare systems to prepare for effectively treating critical cases of the disease and minimizing fatality during both this pandemic as well as those in the future based on the known prevalence of comorbidities within the populations they serve.

References

Basford, K. E., Greenway, D. R., McLachlan, G. J. and Peel, D. (1997). Standard errors of fitted means under normal mixture models. *Computational Statistics*, **12**, 1-17.

CDCa. People who are at higher risk for severe illness. *Centers for Disease Control and Prevention.* https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html.

CDCb. Information for healthcare professionals: COVID-19 and underlying conditions. *Centers for Disease Control and Prevention.* https://www.cdc.gov/coronavirus/2019-ncov/hcp/underlying-conditions.html.

CDCc. Behavioral risk factor surveillance system. *Centers for Disease Control and Prevention*. https://www.cdc.gov/brfss/data_documentation/index.htm.

Chen, J., Lu, H., Melino, G., et al. (2020). COVID-19 infection: the China and Italy perspectives. *Cell Death and Disease*. **11(6)**, 438. doi:10.1038/s41419-020-2603-0.

COVID-19 United States cases by county. *Johns Hopkins University and Medicine*. https://coronavirus.jhu.edu/us-map.

Cummings. M. J., Baldwin, M. R., Abrams, D., et al. (2020). Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet,* **395(10239)**, 1763-1770. doi:10.1016/S0140-6736(20)31189-2.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B39**, 1–38.

Feng, Z., Li, Q., Zhang, Y., et al. (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *China Centers for Disease Control and Prevention Weekly*, **1**, 145-151. doi:10.3760/cma.j.issn.0254-6450.2020.02.003.

Guan, W. J., Liang, W. H., Zhao, Y., et al. (2020). Comorbidity and its impact on 1590 patients with Covid-19 in China: A Nationwide Analysis. *European Respiratory Journal*, **55(5)**, 2000547. doi:10.1183/13993003.00547-2020.

Hajrajabi, A. and Maleki, M. (2019). Nonlinear semiparametric autoregressive model with finite mixtures of scale mixtures of skew normal innovations. *Journal of Applied Statistics,* **46(11)**, 2010-2029.

Hoseinzadeh, A., Maleki, M., Khodadadi, Z. and Contreras-Reyes, J. E. (2019). The skew-reflected-Gompertz distribution for analyzing symmetric and asymmetric data. *Journal of Computational and Applied Mathematics,* **349**, 132-141.

Jones, P. N. and McLachlan, G. J. (1992). Fitting finite mixture models in a regression context. *Australian Journal of Statistics,* **34**, 233–240.

Maleki, M., Mahmoudi, M. R., Wraith, D., Pho, K. H. (2019a). Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Medicine and Infectious Disease,* https://doi.org/10.1016/j.tmaid.2020.101742.

Maleki, M., Contreras-Reyes, J. E., Mahmoudi, M. R. (2019b). Robust mixture modeling based on two-piece scale mixtures of normal family. *Axioms*, **8(2)**, 38.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. (ISBN: 978-0-471-00626-8).

Moravveji, M., Khodadadi, Z., Maleki, M. (2019). A Bayesian analysis of two-piece distributions based on the scale mixtures of normal family. *Iranian Journal of Science and Technology, Transactions A: Science*, **43(3)**, 991–1001.

Richardson, S., Hirsch, J. S., Narasimhan, M., et al. (2020). Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *Journal of the American Medical Association*, **10022**, 1-8. doi:10.1001/jama.2020.6775.

Statista. Italy: comorbidities in COVID-19 deceased patients 2020. https://www.statista.com/statistics/1110949/common-comorbidities-in-covid-19-deceased-patients-in-italy/.

U.S. Census Bureau. https://www.census.gov/quickfacts/.

Wang, B., Li, R., Lu, Z., Huang, Y. (2020). Does comorbidity increase the risk of patients with covid-19: Evidence from meta-analysis. *Aging*, **12(7)**, 6049-6057. doi:10.18632/AGING.103000.

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., et al. (2020). Prevalence of comorbidities and its effects in patients infected with SARS CoV-2: a systematic review and meta-analysis. *International Journal of Infectious Diseases,* **94**, 91-95. https://doi.org/10.1016/j.ijid.2020.03.017.

Zarrin, P., Maleki, M., Khodadadi, Z., Arellano-Valle, R. B. (2018). Time series process based on the unrestricted skew normal process. *Journal of Statistical Computation and Simulation*, **89** (**1**), 38-51.

APPENDIX

Table 1: Estimated parameters of the MoPTS $(\boldsymbol{D = 3, g = 3})$ with standard deviations.

$\boldsymbol{C_1}$	Est.	S.D.	$\boldsymbol{C_2}$	Est.	S.D.	$\boldsymbol{C_3}$	Est.	S.D.
$\boldsymbol{\pi_1}$	0.244186	4.235e-01	π_2	0.33721	3.683e-02	π_3	0.418604	5.012e-02
α_{11}	2.09397	8.311e-02	α_{21}	0.54520	1.321e-01	α_{31}	–0.02475	1.074e-03
$\boldsymbol{\alpha_{12}}$	–1.32063	3.242e-02	α_{22}	–0.13055	2.032e-02	α_{32}	–0.03225	2.034e-02
$\boldsymbol{\alpha_{13}}$	–2.66249	4.101e-01	α_{23}	0.65799	1.721e-02	α_{33}	–0.02447	2.984e-03
$\boldsymbol{\alpha_{14}}$	1.62983	3.422e-03	α_{24}	–0.14953	2.032e-03	α_{34}	0.06790	3.352e-03
$\boldsymbol{\alpha_{15}}$	–0.29323	3.012e-02	α_{25}	0.52714	2.857e-02	α_{35}	0.03763	2.405e-02
$\boldsymbol{\beta_{1,1}}$	–1.85318	6.037e-02	$\beta_{2,1}$	–0.73512	9.063e-02	$\beta_{3,1}$	–0.08560	7.342e-03
$\boldsymbol{\beta_{1,2}}$	0.14396	3.755e-02	$\beta_{2,2}$	0.048204	4.234e-03	$\beta_{3,2}$	0.00883	8.311e-03
$\boldsymbol{\beta_{1,3}}$	–0.00136	4.937e-05	$\beta_{2,3}$	–0.00039	9.043e-04	$\beta_{3,3}$	–0.00006	1.003e-05
$\boldsymbol{\sigma_1^2}$	574.38301	2.223e01	σ_2^2	75.21687	0.389e01	σ_3^2	16.30303	0.232e01

Table 2: Three clusters of 86 U.S. cities identified by the 3 components of the MoPTS model. The county and the state (including District of Columbia) of each city are shown.

Cluster 1 (21 members)	**Cluster 2** (29 members)	**Cluster 3** (36 members)
Fairfield, Connecticut	Arapahoe, Colorado	Jefferson, Alabama
Hartford, Connecticut	Denver, Colorado	Mobile, Alabama
New Haven, Connecticut	Weld, Colorado	Maricopa, Arizona
Dougherty, Georgia	Washington, District of Columbia	Pima, Arizona
Jefferson, Louisiana	Cook, Illinois	Los Angeles, California
Orleans, Louisiana	DuPage, Illinois	Orange, California
Essex, Massachusetts	Lake, Illinois	Riverside, California
Hampden, Massachusetts	Will, Illinois	San Bernardino, California
Middlesex, Massachusetts	Lake, Indiana	San Diego, California
Norfolk, Massachusetts	Marion, Indiana	Santa Clara, California
Plymouth, Massachusetts	Caddo, Louisiana	Adams, Colorado
Suffolk, Massachusetts	East Baton Rouge, Louisiana	Jefferson, Colorado
Macomb, Michigan	Baltimore City, Maryland	New Castle, Delaware
Oakland, Michigan	Bristol, Massachusetts	Broward, Florida
Wayne, Michigan	Worcester, Massachusetts	Lee, Florida
Essex, New Jersey	Genesee, Michigan	Miami Dade, Florida
Hudson, New Jersey	Hennepin, Minnesota	Palm Beach, Florida
Mercer, New Jersey	St. Louis City, Missouri	DeKalb, Georgia
Passaic, New Jersey	Camden, New Jersey	Fulton, Georgia
Union, New Jersey	Erie, New York	Kane, Illinois
Westchester, New York	Lucas, Ohio	Polk, Iowa
	Mahoning, Ohio	Jefferson, Kentucky
	Berks, Pennsylvania	Ramsey, Minnesota
	Lackawanna, Pennsylvania	Clark, Nevada
	Lehigh, Pennsylvania	Hillsborough, New Hampshire
	Northampton, Pennsylvania	Onondaga, New York
	Philadelphia, Pennsylvania	Cuyahoga, Ohio
	King, Washington	Franklin, Ohio
	Milwaukee, Wisconsin	Hamilton, Ohio
		Summit, Ohio
		Allegheny, Pennsylvania
		Shelby, Tennessee
		Dallas, Texas
		Harris, Texas
		Tarrant, Texas
		Snohomish, Washington

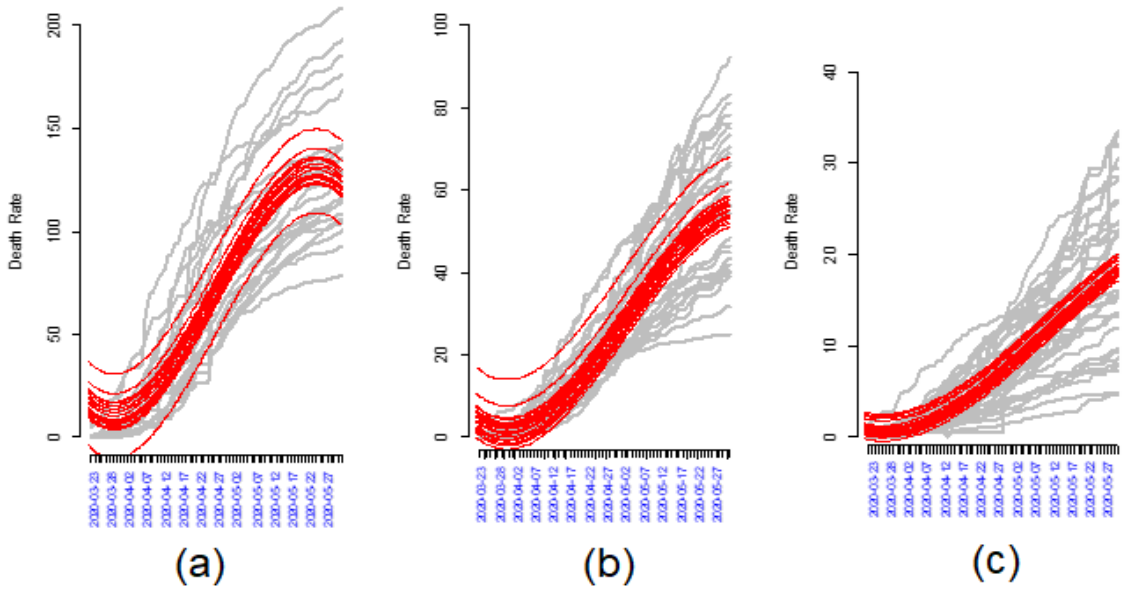


Figure 1: Time series plots (grey) of 86 U.S. cities belonging to clusters 1 (a), 2 (b) and 3 (c) based on $g = 3$ mixture components, superimposed with the fitted MoPTS model (red) for each city. The x-axis denotes time and y-axis COVID-19 death rate. The clusters show marked differences in their respective ranges of death rates (y-axes) over time.

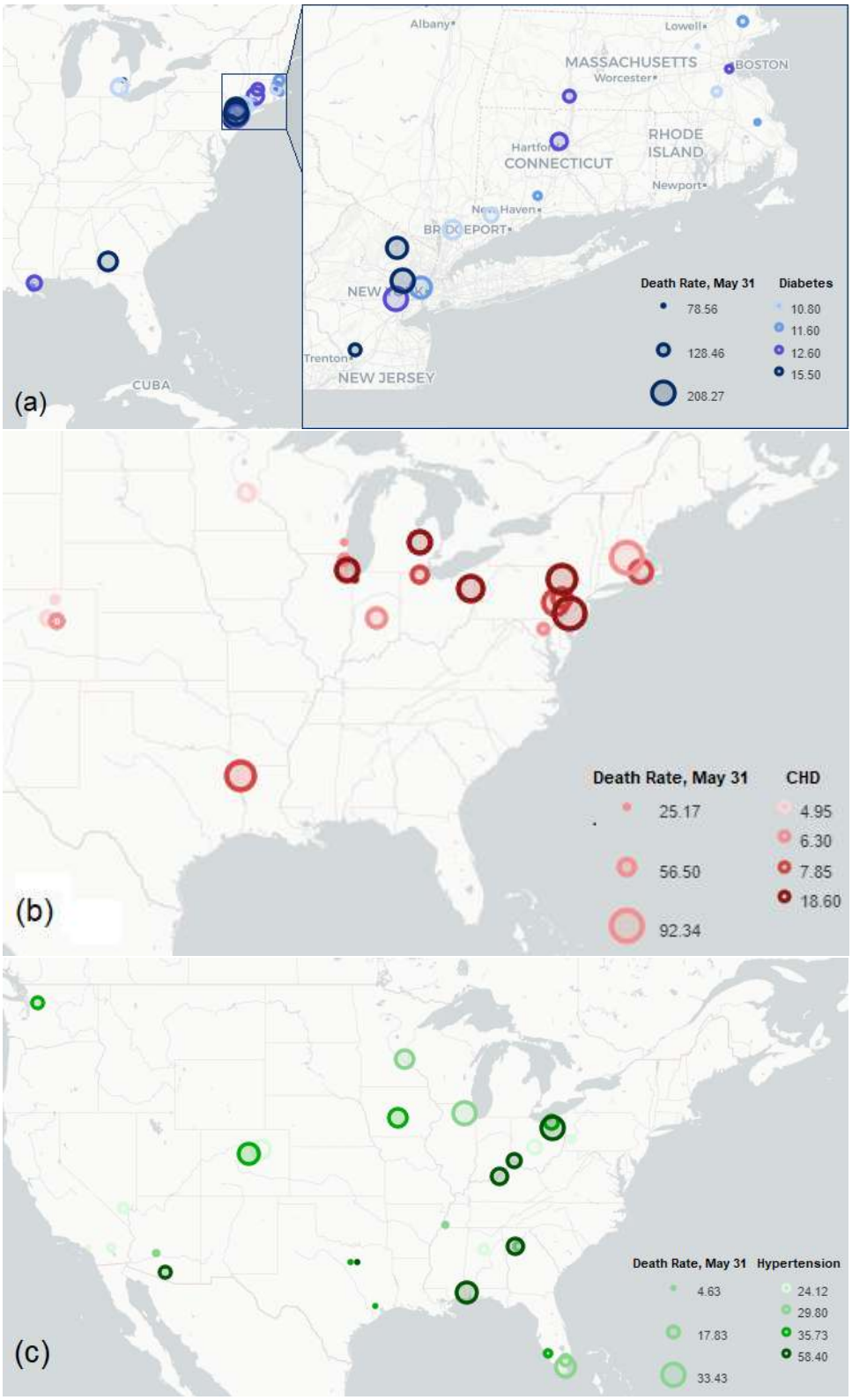


Figure 2. Geographical distribution of the U.S. cities in clusters 1 (a), 2 (b), and 3 (c) based on MoPTS model. On a map, a bubble's location shows the latitude and longitude of a city *C*, its size is proportional to the COVID-19-specific death rate at *C* as on May 31, 2020, while its shading is proportional to the prevalence of the leading comorbidity in the cluster to which *C* belongs.

Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 307–318

Change Point Modeling of Covid-19 Data in the United States

Sheng Zhang[1], Ziyue Xu[2] and Hanxiang Peng[1]
[1]*Indiana University-Purdue University Indianapolis*
[2]*University High School of Indiana*

Received: 06 July 2020; Revised: 26 July 2020; Accepted: 28 July 2020

Abstract

To simultaneously model the change point and the possibly nonlinear relationship in the Covid-19 data of the US, a continuous second-order *free* knot spline model was proposed. Using the least squares method, the change point of the daily new cases against the total confirmed cases up to the previous day was estimated to be 04 April 2020. Before the point, the daily new cases were proportional to the total cases with a ratio of 0.287, suggesting that each patient had 28.7% chance to infect another person every day. After the point, however, such ratio was no longer maintained and the daily new cases were decreasing slowly. At the individual state level, it was found that most states had change points. Before its change point for each state, the daily new cases were still proportional to the total cases. And all the ratios were about the same except for New York State in which the ratio was much higher (probably due to its high population density and heavy usage of public transportation). But after the points, different states had different patterns. One interesting observation was that the change point of one state was about 3 weeks lagged behind the state declaration of emergency. This might suggest that there was a lag period, which could help identify possible causes for the second wave. In the end, consistency and asymptotic normality of the estimates were briefly discussed where the criterion functions are continuous but not differentiable (irregular).

Key words: Asymptotic normality; Change point; Consistency; Covid-19; Free knot; Irregular criterion function.

1. Introduction

The first case of Novel Coronavirus disease 2019 (Covid-19) was reported in Wuhan, China on 17 November 2019. This disease was caused by SARS-CoV-2 virus, and in about 6 months, it has spread throughout the whole world, infected 15.5 million people, and killed more than 635,000 (https://covid19.who.int/). In the United States, there are 4 million confirmed cases, and 143,000 deaths by 25 July 2020. Many states have ordered their residents to stay at home and keep social distancing to slowdown the rapid spread of the virus, so that the health care system will not be overwhelmed. The trend of daily new cases in the US appeared to be flattened in the early April. Here, we first fitted the data with the change point model (Bai, 1997; Julious, 2000) to identify the possible date for the trend change.

The first case in US was reported on 21 January 2020 in Washington State. By the end of February, several more confirmed cases were recorded there. By the end of March, the number of confirmed cases quickly went up to about 6,000. On 29 February 2020, the Governor declared the state emergency. A few weeks later, the daily new cases stabled and slowly started

Corresponding Author: Hanxiang Peng
Email: hanxpeng@iupui.edu

decreasing. Similar patterns could also be observed in other states. By fitting the state data with a change point model, we found that the change point was correlated with the date when the state emergency was declared. Thus, we believed that one main possible cause for the change point could be the declaration of state emergency. Once people started to protect themselves more carefully, the effect of the protection would be noticeable after 2-3 weeks. Knowledge about this delay period would help us identify the causes if the trend changes again.

In this study, we used the data collected by New York Times Company. The data is stored at GitHub (http://github.com/nytimes/covid-19-data/blob/master/). It contains the number of cumulated cases at the county level, state level, and country level, starting from 21 January 2020. We downloaded the data up to 18 June 2020 for this study.

2. Change Point Model and Data Fitting Procedure

First, we fitted the data at the county-level. Displayed in Figure 1A is the plot of the number of daily new cases against the total number of cases up to the previous day. Noticeably, there is a change point between 28 March and 05 April 2020, around which an increasing relationship of the daily new cases against the total cases was progressed to decreasing. Specifically, at first, the number of daily new cases was drastically increasing with the total number of cases up the previous day. Then after some critical point, the increasing relationship turned to decreasing but at a slow rate. This seems to be no surprising. When Covid-19 broke out, a great number of people got infected within a short period of time. Meanwhile, measures such as social distancing and using of personal protective equipment were taken, the spreading was slowed down. Motivated by these plots, we chose to use a change point model to fit the data.

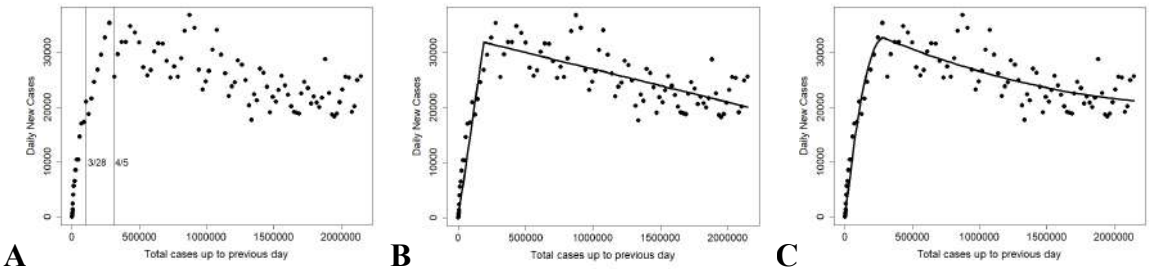


Figure 1: The scatter plot of the data. The *y*-axis is the daily new cases, and the *x*-axis is the total cases up to the previous day. **A**: the scatter plot; **B**: the scatter plot superimposed with the fitted linear model (1); **C**: the scatter plot superimposed with the fitted quadratic model (2).

In a linear change point model, the expected value $E(y_i)$ of the number y_i of daily new cases is expressed as a linear function of the total number x_i, of cases up to the previous day, *i.e.*,

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+, \quad i = 1, \dots, n \tag{1}$$

where δ is an unknown change point, and $x_+ := \max(x, 0)$ is the positive part of x. Here β_0 is the intercept, which is expected to be very close to zero (there should be almost no new case if there is no confirmed cases), β_1 is the rate of infection before the change point, which can be interpreted as how many persons will be infected by each patient every day; β_2 can be interpreted as the effectiveness of the protective measures taken. Treating δ as known, we

estimated the parameters beta by the least squares method. To estimate the change point δ, we searched all possible values of the change point and compared the corresponding sum of squared residuals (SSE). The estimate of change point is the one corresponding to the smallest SSE. The data in Figure 1A was fitted to the linear model with one change point, with the daily new cases as the response and the total number of cases up to the previous day as the predictor. Listed in Table 1 are all the possible change points with corresponding SSE values.

Table 1: The possible change points and corresponding SSE from linear model

Date	δ	SSE$\times 10^{-9}$
2020-03-28	102835	1.34
2020-03-29	123730	1.23
2020-03-30	142406	1.14
2020-03-31	163873	1.08
2020-04-01	**188425**	**1.05**
2020-04-02	215176	1.08
2020-04-03	244636	1.17
2020-04-04	277279	1.34
2020-04-05	312519	1.61

From the Table 1, the estimate of the change point is 01 April 2020, which is consistent with our observation. The results of all other parameter estimates are listed in Table 2. With this change point, the fitted equation is:

$$E(y) = 495 + 0.1662x - 0.1722(x - 188425)_+$$
$$= \begin{cases} 495 + 0.1662x, & x \le 188425, i.e., \text{before 01 April 2020} \\ 32942 - 0.006x, & x > 188425, i.e., \text{after 01 April 2020} \end{cases}$$

In this equation, $\hat{\beta}_0$ is not significantly different from zero, which was consistent with our intuition: β_0 should be very close to zero. $\hat{\beta}_1 = 0.1662$ indicated that before the change point, each patient had 16.6% chance to infect another healthy person every day. $\hat{\beta}_2 = -0.1772$ suggested that after the total number of confirmed cases reached to 188,425, the slope of the linear trend is $(\hat{\beta}_1 + \hat{\beta}_2) = -0.006$. This showed that the daily new cases were decreasing after 01 April 2020, but at a very slow rate.

Displayed in Figure 1B is the scatter plot of the data superimposed with the fitted lines, using 01 April 2020 as the change point. The simple linear model fit data well, except that there are some noticeable non-linear features for both before and after the change point. This motivated us to fit the data with a continuous quadratic change point model:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 \quad (2)$$

where β_1 is the initial rate when there is only a small amount of confirmed cases; β_{12} is the correction factor for the non-linear feature before the change point; β_2 and β_{22} indicate the effectiveness of the prevention measures after the change point. Our study exhibited in this model that the LSE is asymptotic normal. The estimation method is the same as described above, and the possible change points and their corresponding SSE are listed in Table 3.

In this model, the change date is 04 April 2020, and all other parameter estimates are listed in Table 2. The fitted equation is:

$$E(y) = 290 + 0.2247x - 3.898 \times 10^{-7}x^2 - 0.01845(x - 277279)_+ + 3.919 \times 10^{-7}(x - 277279)_+^2$$

$$= \begin{cases} 290 + 0.2247x - 3.898 \times 10^{-7}x^2, & \text{before 04 April 2020} \\ 35531 - 0.01106x + 2.028 \times 10^{-9}x^2, & \text{after 04 April 2020} \end{cases}$$

The superimposed plot is shown in Figure 1C. The quadratic model appeared to be a better fit to the data. To confirm this, we performed ANOVA test to test if the linear model is significant. The ANOVA test result in Table 4 indicated that the full model is appropriate. Another question that arises is - should we still pick 01 April 2020 as the change point as suggested from the linear model? The ANOVA test result in Table 4 suggested 04 April 2020. Possibly the linear model is somewhat oversimplified, as it ignores the curve features before and after the change point, which could lead to restrictions on selecting the change point due to its lack of flexibility. Thus, we would suggest 04 April 2020 as the change point for the US.

Table 2: Parameter estimates from model (1)-(3)

Model (1): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+$

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	495	335	1.474	0.143
$\hat{\beta}_1$	0.1662	0.0037	44.533	0.000
$\hat{\beta}_2$	−0.1722	0.0041	−41.685	0.000

Model (2): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	290	323	0.898	0.371
$\hat{\beta}_1$	0.2247	0.00146	15.356	0.000
$\hat{\beta}_{12}$	-3.898×10^{-7}	5.56×10^{-8}	−7.01	0.000
$\hat{\beta}_2$	−0.01845	0.01798	−1.025	0.307
$\hat{\beta}_{22}$	3.919×10^{-7}	5.53×10^{-8}	7.091	0.000

Model (3): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$ +weekly effect

The residual has AR (1) pattern

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	132	197	0.671	0.504
$\hat{\beta}_1$	0.2871	0.0161	17.825	0.000
$\hat{\beta}_{12}$	-5.143×10^{-7}	4.37×10^{-8}	−11.77	0.000
$\hat{\beta}_2$	-0.0128	0.011	−1.158	0.249
$\hat{\beta}_{22}$	5.17×10^{-7}	4.36×10^{-8}	11.87	0.000
Monday effect	−10130	1586	−6.385	0.000
Tuesday effect	−8960	1595	−5.618	0.000
Wednesday effect	−7848	1602	−4.9	0.000
Thursday effect	−4933	1611	−3.062	0.003
Friday effect	−3513	1557	−2.256	0.026
Saturday effect	−5379	1566	−3.44	0.001
Sunday effect	−10090	1577	−6.397	0.000

Table 3: The possible change points & corresponding SSE from the quadratic model (2)

Date	Value of δ	SSE$\times 10^{-9}$
2020-03-28	102835	1.2719
2020-03-29	123730	1.1982
2020-03-30	142406	1.1266
2020-03-31	163873	1.0520
2020-04-01	188425	0.9893
2020-04-02	215176	0.9446
2020-04-03	244636	0.9190
2020-04-04	**277279**	**0.9111**
2020-04-05	312519	0.9112
2020-04-06	337984	0.9163
2020-04-07	367599	0.9285
2020-04-08	399388	0.9454

In Figure 1A, one notices that besides the trend, the variation of daily new cases exhibited strong weekly effect: during the weekend, the number was small, and during the middle of a week, the number was high. Here, the plot of the residual after 18 March 2020 is shown in Figure 2A. The plot indicated that there was an oscillation pattern. The auto-correlation function (ACF) plot of the residual is shown in Figure 2B. From the ACF plot, the weekly effect was apparent: the residual was highly positive correlated on 7 days and 14 days.

Table 4: The ANOVA test results

Full model: Quadratic model (2); reduced model: linear model (1)

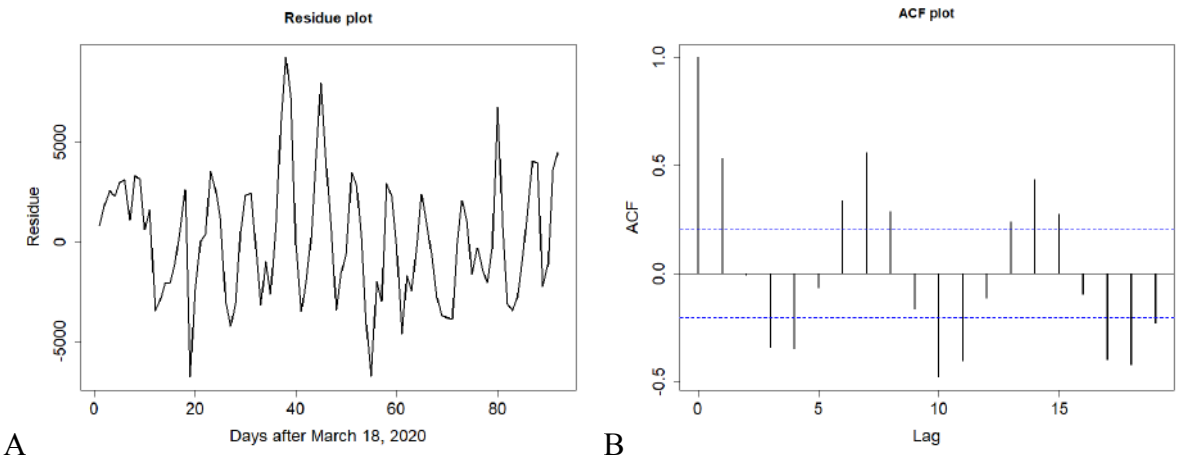
Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.9111×10^9	143			
Reduced	1.05×10^9	145	68849507	10.81	0.00043

Full model: Quadratic model (2); reduced model: change date is 01 April 2020

Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.9111×10^9	143			
Reduced	0.9893×10^9	144	78144937	12.26	0.00061

Full model: Quadratic model (3); reduced model: Quadratic model (4)

Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.3793×10^9	141			
Reduced	0.5202×10^9	136	140938174	10.10	0.0000



A B

Figure 2: The residual plot (A) and ACF plot (B)

Table 5: The possible change points and corresponding SSE from model (3)

Date	Value of δ	SSE$\times 10^{-8}$
2020-04-01	188425	4.348
2020-04-02	215176	4.102
2020-04-03	244636	3.899
2020-04-04	**277279**	**3.793**
2020-04-05	312519	3.801
2020-04-06	337984	3.885
2020-04-07	367599	4.023
2020-04-08	399388	4.187

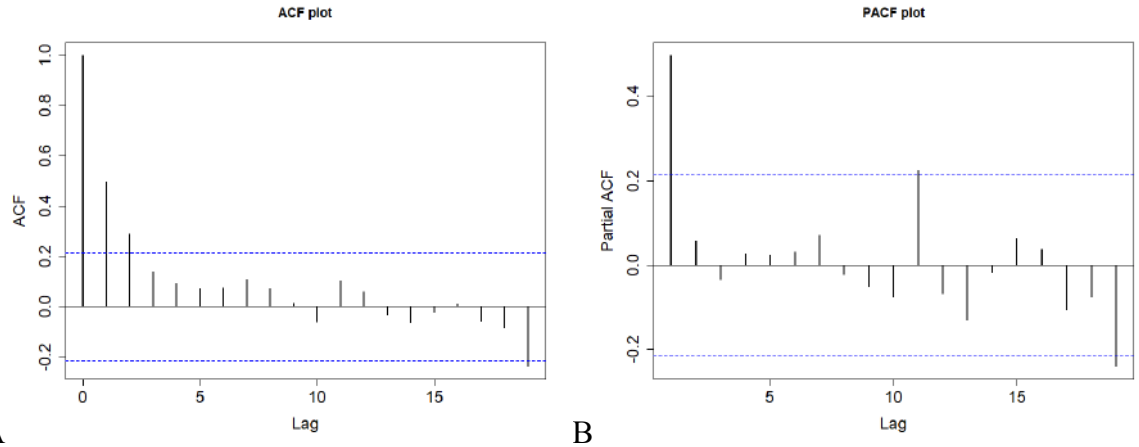
To address the weekly effect, we include the weekday-indicator in the model for the data collected after 27 March 2020:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 + \left(\sum_{j=1}^{7} \beta_{3j} \cdot \mathbf{1}\{\text{Weekday}_i = j\}\right) \cdot \mathbf{1}\{Date_i \geq 27\ March\ 2020\} \quad (3)$$

Another model for the weekly effect to use the periodic sine and cosine functions:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 + \left(\beta_s \sin\left(2\pi \cdot \frac{\text{Weekday}_i}{7}\right) + \beta_c \cos\left(2\pi \cdot \frac{\text{Weekday}_i}{7}\right)\right) \cdot \mathbf{1}\{Date_i \geq 27\ March\ 2020\} \quad (4)$$

It can be seen that that model (4) is a reduced model of model (3): $\beta_{3j} = \beta_s \sin\left(\frac{2j\pi}{7}\right) + \beta_c \cos\left(\frac{2j\pi}{7}\right), j = 1, \ldots, 7$. Thus, we can use the ANOVA to test if model (4) is significant. The ANOVA test result in Table 4 supported the full model (3). For the model (3), the possible change point and the corresponding SSE is listed in Table 5, and the result still showed that 04 April 2020 was the change point.



A B

Figure 3: The ACF plot (A) and PACF plot (B) for residue from model (3).

The data was fitted to the weekly effect model (3), and the ACF and partial ACF (PACF) plot of resulted residuals were shown in Figure 3. The PACF plot indicated that the residuals had auto-regression pattern {AR (1)}. The data was then fitted with the weekly-effect model with AR (1). The results are listed in Table 2 and shall be discussed in next section.

The data from individual state was fitted using the following model,

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 \quad (5)$$

Here we removed the second order term before the change point, and our motivation was that this model is more sensitive to the change point based on our theoretical study. The results are listed in Table 6 and shall be discussed in next section.

Table 6: The results for individual state data

Model (5): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$

The average delay-time between the date to declare state emergency and change point was 21.8 days with the standard deviation of 5.2 days.

State Name	Change point	$\hat{\beta}_1$	Date to declare state emergence[1]
Alabama	4/2/2020	0.133	3/13/2020
Arizona	3/28/2020	0.229	3/11/2020
California	3/30/2020	0.161	3/4/2020
Colorado	3/25/2020	0.256	3/10/2020
Connecticut	4/5/2020	0.151	3/10/2020
Delaware	4/5/2020	0.166	3/12/2020
D. C.	3/31/2020	0.175	3/11/2020
Florida	4/2/2020	0.156	3/1/2020
Illinois	3/25/2020	0.296	3/9/2020
Indiana	3/30/2020	0.220	3/6/2020
Iowa	4/8/2020	0.168	3/9/2020
Kansas	3/26/2020	0.289	3/9/2020
Kentucky	4/6/2020	0.119	3/9/2020
Louisiana	4/1/2020	0.276	3/11/2020

Maine	3/27/2020	0.106	3/15/2020
Maryland	4/2/2020	0.186	3/5/2020
Massachusetts	3/27/2020	0.328	3/10/2020
Michigan	3/31/2020	0.205	3/11/2020
Mississippi	4/2/2020	0.115	3/4/2020
Missouri	3/31/2020	0.185	3/13/2020
Nebraska	4/8/2020	0.156	3/13/2020
Nevada	3/28/2020	0.221	3/12/2020
New Hampshire	3/28/2020	0.173	3/13/2020
New Jersey	3/29/2020	0.237	3/9/2020
New Mexico	4/6/2020	0.120	3/11/2020
New York	3/22/2020	0.436	3/7/2020
North Carolina	3/26/2020	0.243	3/10/2020
Ohio	4/2/2020	0.144	3/9/2020
Pennsylvania	4/3/2020	0.187	3/6/2020
Rhode Island	4/8/2020	0.157	3/9/2020
South Carolina	3/31/2020	0.168	3/13/2020
South Dakota	4/8/2020	0.254	3/13/2020
Tennessee	3/30/2020	0.143	3/12/2020
Texas	4/5/2020	0.141	3/13/2020
Utah	3/27/2020	0.229	3/6/2020
Virginia	3/31/2020	0.189	3/12/2020
Washington	3/26/2020	0.159	2/29/2020

[1]: *the date of the declaration of state emergency is from wikipedia.org (https://en.wikipedia.org/wiki/U.S._state_and_local_government_response_to_the_COVID-19_pandemic)*

3. Results and Discussions

For the US data, from Table 2 the fitted equation is given by

$$E(y) = 132 + 0.287x - 5.143 \times 10^{-7}x^2 - 0.01278(x - 277279)_+ + 5.170 \times 10^{-7}(x - 277279)_+^2 + \text{weekly effect}$$

$$= \begin{cases} 132 + 0.287x - 5.143 \times 10^{-7}x^2 + \text{weekly effect}, & \text{before 04 April 2020} \\ 43424 - 0.01248x + 2.7 \times 10^{-9}x^2 + \text{weekly effect}, & \text{after 04 April 2020} \end{cases}$$

Here, $\hat{\beta}_1 = 0.287$ suggested that at the early stage when the total number of confirmed cases was small, each patient had 28.7% chance to infect another healthy person each day. Since a Covid-19 patient usually recovered within 2 weeks, R_0 value can be estimated by $0.287 \times 14 = 4.01$, which was consistent with the published results (median value 5.7 with 95% confidence interval: 3.8 - 8.9, Steven Sanche, *et. al.*, 2020). $\hat{\beta}_{12} < 0$ indicated that even before the change point, the rate was decreasing from 0.287. In fact, the rate at 04 April 2020 can be calculated as $0.287 - 5.143 \times 10^{-7} \times 277279 = 0.144$, which was only half of the original rate. In our study of the state data, we found that several states had their change points in late March. This could be the due to the reason that the rate was decreased to 0.144, as several states had already slowed down.

For the whole US data, the change point was 04 April 2020. Because the median incubation time of Covid-19 was 4-5 days, implying that what led to the change point should

have played the role at least one week before 04 April 2020. This seems to indicate that the change point could be resulted from the issuance of National Emergency on 13 March 2020. If it was true, it suggested that the effect of people's behavior would be reflected by the change point about 21 days later. The same lag effect was also observed at the state level.

Listed in Table 5 are the fitting results for the data from individual states. Washington State was the first with the outbreak of Covid-19. The scatter plot of the data superimposed with the fitted curve is shown in Figure 4A. Before the change point, the daily new cases were increasing. After 26 March 2020, however, the number started to decrease. But the number seemed to comeback recently. The estimate was $\hat{\beta}_1$= 0.154, indicating that the initial rate in Washington States was less than the average rate (0.287) of the US. The state emergency was declared on 29 February 2020, and the change point was on 26 March 2020, thus it showed about 25-day delay.

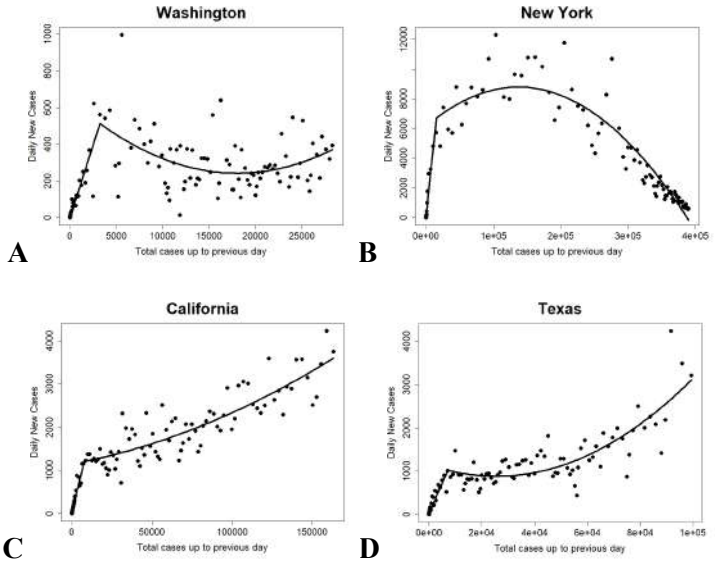


Figure 4: The scatter plot of state data. The *y*-axis is the number of daily new cases, and the *x*-axis is the total number of cases up to the previous day. **A**: Washington State. **B**: New York State. **C**: California State. **D**: Texas State.

New York State was a hot spot in March. The plot is shown in Figure 4B. The estimate $\hat{\beta}_1$ is equal to 0.436, which is the highest among all states, which could be due to its high population density and heavy public transportation. The state emergency was declared on 07 March 2020 and the change point was 22 March 2020, which lagged behind 15 days. After 22 March 2020, the daily new cases stayed with high value and then dropped down. This seemed to indicate that the Covid-19 appeared to be controlled.

The plot of data from California State is shown in Figure 4C. The state emergency was declared on 04 March 2020 and the change point was 30 March 2020, which lagged behind about 26 days. However, after the change point, the daily new cases were only slowing down and still kept increasing. To further control Covid-19, more efforts would be needed. The plot of data from Texas is shown in Figure 4D. The state emergency was declared on 13 March 2020 and the change point was 05 April 2020. For Texas, the lag time was 22 days.

As we discussed before, the estimate $\hat{\beta}_1$ for each state was proportional to R_0 for that state before any prevention measures were used. Some states, similar to New York State, like Massachusetts and Illinois, have big metropolitan areas (Boston in MA, and Chicago in IL) with high population density and heavily public transportation. Thus, the estimate $\hat{\beta}_1$ of these

states were relatively higher than the rest. Other states, like Mississippi and New Mexico, have no such big cities, and usually had lower estimate $\hat{\beta}_1$.

Overall, the data from most states showed a change point pattern. Before the point, the daily new cases were proportional to the total cases, similar to the whole US data. By comparing the change point and the date when the state emergency was declared in Table 6, we found that the average delay-period is 21.8 days. This suggested that if there is another change point, what happen 3 weeks before would likely be the causes of the change.

4. Consistency and Asymptotic Normality

Here, we present consistency and asymptotic normality results and omit the proofs. What is novel here is that we model the change point and the possible non-linear relationship simultaneously, whereas a typical change point model involves in only $(x-\delta)_+$. This is *a continuous second-order free spline model with one knot*.

To prove asymptotic normality, we have to deal with the irregular criterion function $\left(y_i - \boldsymbol{\beta}^T \boldsymbol{z_i}(\delta)\right)^2$, in which the truncated power function x_+ is not differentiable. Thanks to Theorem 5.23 of van der Vaart (1998), we have obtained a quick result at the price of boundedness Assumption 4. In other words, with careful elaboration, we believe this assumption (and some of others as well) can be relaxed to the boundedness assumption of the knot parameter δ as in the case of consistency, see Wu, *et. al.* (2019).

Consider that $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ satisfy the second-order free spline model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x-\delta)_+ + \beta_4 (x-\delta)_+^2 + \epsilon_i, \ i = 1, \dots, n,$$

where $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are i.i.d. random errors with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2 < \infty$, $x_1, x_2, \dots, x_n$ are assumed to be non-random, both β and δ are unknown parameters to be estimated.

Denote $\boldsymbol{\beta} = (\beta_0, \ \beta_1, \ \beta_2, \ \beta_3, \ \beta_4)^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$, and $\boldsymbol{z_i}(\delta) = (1, \ x_i, \ x_2^2, \ (x-\delta)_+, \ (x-\delta)_+^2)^T$**.** Using these symbols, we can write

$$y_i = \boldsymbol{\beta}^T z_i(\delta) + \epsilon_i, \qquad i = 1, \dots, n.$$

We estimate $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$ by the least squares estimate (LSE) $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}^T, \hat{\delta}\right)^T$, that is,

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S_n(\boldsymbol{\theta}), where \ S_n(\boldsymbol{\theta}) = S_n(\boldsymbol{\beta}^T, \delta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^T \boldsymbol{z_i}(\delta)\right)^2. \tag{6}$$

For $\delta \in \Delta \subset R$ fixed, the minimization (6) simplifies to the usual LSE problem. Let $\boldsymbol{Z}(\delta)$ be the $n \times 5$ matrix consisting of $\boldsymbol{z_1}(\delta), \boldsymbol{z_2}(\delta), \dots, \boldsymbol{z_n}(\delta)$ as its rows, and let $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$. If $\boldsymbol{Z}(\delta)$ has full rank 5, then the LES $\hat{\beta}(\delta)$ is given by

$$\hat{\beta}(\delta) = [\boldsymbol{Z}^T(\delta)\boldsymbol{Z}(\delta)]^{-1}\boldsymbol{Z}^T(\delta)\boldsymbol{y}. \tag{7}$$

As a result, the minimization (6) becomes minimizing the new objective over $\delta \in \Delta$:

$$\hat{\delta} = \arg\min_{\delta \in \Delta} \tilde{S}_n(\delta), \ S_n(\boldsymbol{\beta}^T(\delta), \delta) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^T(\delta)\boldsymbol{z_i}(\delta)\right)^2.$$

Assumption 1. There exists a compact subset Δ of $\mathbb{R}$ and a matrix function $\boldsymbol{M}(\delta_1, \delta_2)$, $\delta_1, \delta_2 \in \Delta$, such that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z_i}(\delta_1)\boldsymbol{z}_i(\delta_2)^T \to \boldsymbol{M}(\delta_1, \delta_2),$$

uniformly in $\delta_1, \delta_2 \in \Delta$, that $\boldsymbol{M}(\delta, \delta)$ is positive definite on Δ, and that $\boldsymbol{T}(\delta) = \boldsymbol{M}(\delta_0, \delta_0) - \boldsymbol{M}(\delta_0, \delta)\boldsymbol{M^{-1}}(\delta, \delta)\boldsymbol{M}(\delta_0, \delta)$ has a unique zero solution at $\delta = \delta_0$.

Assumption 2. For large n, $\sup_{1\le i\le n}\{|x_i|\} \le M_x < \infty$ for some constant M_x.

Note **Assumption 1** ensures that the maximizer is well-separated and unique. It is a typical assumption for establishing consistency of M-estimators, see Chapter 5 of Van der Vaart (1998), Yu and Ruppert (2002) and Wu, *et al.* (2019).

Theorem 1. Assume Assumptions 1 and 2. Then the LSE $\widehat{\boldsymbol{\theta}}$ converges in probability the true value $\boldsymbol{\theta_0} = (\boldsymbol{\beta_0^T}, \delta_0)$ of parameter, *i.e.*,$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta_0}$, in probability.

Remark. If X_i are random, consistency still holds provided that ϵ_i and X_i are independent with $EX^2 < \infty$, and the convergence in **Assumption 1** is modified to convergence in probability.

We need the following assumptions to assure asymptotic normality.

Assumption 3. $X_1, X_2, \dots, X_n$ are i.i.d. with a common continuous density function f, X_i and ϵ_i are independent for all i, and $E(X^8) < \infty$.

Because x_+ is not differentiable, asymptotic normality was proved using the empirical process theory. This requires the square-integrability of the envelope function, which is a polynomial of x of fourth degree, leading to finite 8th moment assumption.

Assumption 4. There exists a neighborhood of $\boldsymbol{\theta_0}$, such that $\forall \boldsymbol{\theta} \in N(\boldsymbol{\theta_0})$, $\|\boldsymbol{\theta}\| \le B_0 < \infty$ for some constant $B_0 > 0$**.**

Let $\mu(\boldsymbol{\theta}) = E\big(\boldsymbol{S}_n(\boldsymbol{\beta}^T, \delta)\big)$, $\dot{\mu}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\mu(\boldsymbol{\theta})$ be the 6-dimensioinal derivative vector and $\boldsymbol{V}(\boldsymbol{\theta_0}) = \ddot{\mu}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\mu(\boldsymbol{\theta})$ be the 6-by-6 matrix of second partial derivatives.

Assumption 5. $\dot{\mu}(\boldsymbol{\theta_0}) = \boldsymbol{0}$ and the matrix $V(\boldsymbol{\theta_0})$ is nonsingular.

Theorem 2. Assume Assumptions 3-5. If the LSE is consistent, *i.e.*, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \to \boldsymbol{\theta_0}$ in probability, then $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is asymptotically linear,

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\big) = -\boldsymbol{V^{-1}}(\boldsymbol{\theta_0})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{m}_{\boldsymbol{\theta_0}}(X_i, Y_i) + o_p(1)$$

where $\dot{\boldsymbol{m}}_{\theta}(x, y) = \frac{\partial}{\partial(\boldsymbol{\beta}^T, \delta)^T}\big(y - \boldsymbol{\beta}^T\boldsymbol{z}(\delta)\big)^2$. Hence, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is asymptotically normal,

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\big) \Longrightarrow N\Big(0, \boldsymbol{V^{-1}}(\boldsymbol{\theta_0})E\big[\dot{\boldsymbol{m}}_{\boldsymbol{\theta_0}}(X_1, Y_1)\dot{\boldsymbol{m}}_{\boldsymbol{\theta_0}}(X_1, Y_1)^T\big]\boldsymbol{V}^{-1}(\boldsymbol{\theta_0})\Big).$$

5. Future study

During the preparation of this paper, we have noticed that there was a second outbreak in the US at the end of June. Our approach can be easily generalized to multiple change points. Currently, we work on the theoretical development in the framework of time series model with multiple change points.

Acknowledgements

The authors are indeed thankful to Prof. Jyoti Sarkar and Prof. Vinod Kumar Gupta, Chair Editor, for their suggestions and encouragement.

References

Bai, J. (1999). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics,* **79(4)**, 551-563.

Julious, S. (2001). Inference and estimation in a changepoint regression problem. *The Statistician*, **50**, 51-61.

Sanche, S., Lin, Y., Xu, C., Romero-Severson, E., Hengartner, N. and Ke, R. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, **26(7)**, 1470-1477.

Van der Vaart, A. W. (1998). *M- and Z- Estimators*. *Asymptotic Statistics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802256.006.

Wu, J., Peng, H. and Tu, W. (2019). Large-sample estimation and inference in multivariate single-index models. *Journal of Multivariate Analysis*, **171**, 382-396.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models, *Journal of American Statistical Association,* **97,** 1042–1054.

Statistics and Applications ISSN 2452-7395 (online)
Volume 18, No. 1, 2020 (New Series), pp 319-331

Estimation of Quality Adjusted Life Year (QALY) for Different States of India During COVID-19

Gurprit Grover[1] and Radhika Magan[2]

[1]*Professor, Department of Statistics, Faculty of Mathematical Sciences, University of Delhi*
[2]*Research Scholar, Department of Statistics, Faculty of Mathematical Sciences, University of Delhi*

Received: 08 July 2020; Revised: 22 July 2020; Accepted: 29 July 2020

Abstract

Covid-19 is an incessant pandemic which is widespread worldwide. Various epidemic models have been used for forecasting the covid-19 cases in India. In this paper we have tried to estimate quality adjusted life year (QALY) for the covid-19 infected patients from 26 March, 2020 till 28 May, 2020 in different states of India. A regression equation for time varying reproduction number has been defined using the basic Susceptible-Infective-Recovered (SIR) model, which is further used to obtain the utility function. The average QALY per month for each state has been computed on the basis of the proposed utility function. Various states are categorised as severe, moderate and controlled regions for the covid-19 pandemic based on QALY values.

Key words: QALY; Utilities; TSIR; Exponential regression; India; Covid-19.

1. Introduction

Pandemic has always attracted global attention due to widespread devastation caused to health of human beings as well as economy of a nation. Dazak *et al*. (2018) have estimated that around 1.67 million yet to be discovered viral species from key zoonotic families exist in mammal and bird hosts. More than 50 percent of these viral species have the potential to cause severe infections via transmission to humans. COVID-19 has a faster spread rate than its ancestors like SARS-COV and MERS-COV but lower mortality rate as stated by Giordano *et al.* (2020). In India covid-19 cases are widespread in all the states. The government had imposed a series of lockdown in five different phases across the country in order to further prevent the spread of virus through community transmission.

Researchers are using various mathematical models in order to study the crucial epidemiological properties of this epidemic. Akshaya *et al*. (2020) have stated how different forecasting techniques have played important roles in capturing the probability of infection and reproduction rate. Wu *et al*. (2020) have indicated that 86% of the infected individuals are expected to remain asymptomatic which are the main sole for spreading the infection under community transmission. Biswas *et al.* (2020) have stated on the basis of Euclidean network that an infected individual can infect another individual with distance (l), rate of infection (δ) and probability proportional to $l^{-\delta}$. In the early days of outbreak, government was keen on tracing contacts of persons who were closely related to the infected individuals.

Corresponding author: Radhika Magan
Email : maganradhika@gmail.com

Those individuals were isolated in order to prevent the further spread of disease. Ghosh *et al*. (2020) analysed the case counts in India using standard epidemiological models and projected on the basis of crisis at present.

Ranjan (2020) reported that the early action of lockdown in India has a favourable effect in limiting the epidemic size. Deo *et al.* (2020) have estimated the reproduction number which was significantly reduced due to lockdown measures. Thus, lockdown not only prevented the rise in the number of cases but also created substantial economic loss to the weaker sections of the society. Chatterjee *et al.* (2020) have performed a review of the pandemic with the current evidence. They highlighted the key areas where research needs attention in order to create critical intelligence for the prevention of its spread. Ferguson *et al.* (2020) figures out South Korea, Taiwan and New Zealand among the few countries who have precisely managed to fight back against the virus.

In this paper we have tried to estimate QALY for the covid-19 infected patients from 26 March 2020 till 28 May 2020. We have estimated QALY by time series modelling of epidemiological model namely, Susceptible Infective Recovered (SIR) model with the exponential form. It enables us to estimate the utilities which in turn help us to compute the Quality Adjusted Life Year (QALY) per month for each state. This quality of life approach is first and foremost attempt in the time of pandemic which is carried over different states of India.

2. Methodology

2.1. Epidemic SIR model linked with exponential regression

Researchers have done a lot of manifold classifications based on SIR models during the covid-19 pandemic for different countries. Jewell *et al.* (2020) describes the underlying principles and value of projections in pandemic models. The early models relates to the region when virus had been circulating in a community. Meanwhile their projections were not robust and reliable. All these models have one thing in common that is the peak which is predicted on the basis of the number of infectives. Prakash *et al.* (2020) have replaced this peak as an artefact of plateau and described it with the help of persistence number. This peak acts as a plateau which grows flat and last for many weeks with no downward trend due to increase in the number of containment zones.

SIR models encapsulate the number of susceptibles to the number of infectives, further to the number of recovered and death cases for a disease. It is also known as the compartmental model in epidemiology. Wu *et al.* (2020) have defined variants of SIR models for policy decisions in China. As human to human transmission occurs there is high rise in the number of infectives. It is deducible that the susceptibles are more likely to get infected and people in the infected stage are either likely to enter the stage of recovered cases or death. The infectious period determined by Ma (2020) states that it is exponentially distributed with mean $(1/\gamma)$. Also, Wallinga *et al.* (2006) have introduced a nonparametric method to develop reproduction number from exponential growth rate.

The flow of individuals from susceptible to infective to recover as well as to death cases have been monitored with the help of system of non-linear differential equations, which are defined as:

$$\frac{\delta S}{\delta t} = -\frac{\beta IS}{N} \tag{1}$$

$$\frac{\delta I}{\delta t} = \frac{\beta IS}{N} - \gamma I \tag{2}$$

$$\frac{\delta R}{\delta t} = \gamma I \tag{3}$$

$$R_0 = \frac{\beta}{\gamma} \tag{4}$$

where $S = S(t)$ is the total number of confirmed cases in a particular state, $I = I(t)$ is the number of active positive cases, $R = R(t)$ denotes the number of recovered individuals. β is the transmission rate, $\gamma$is the recovery rate. The reproduction number (R_0) can be rewritten as the ratio of transmission rate to recovery rate as given in equation (4). When the disease is transmitted from one to person to another under the assumption that the whole population is vulnerable to the exposure of infection, this rate has a steady increase. The individuals are not vaccinated because the infection erupted for the first time with no way to control the spread. This number reproduces itself in the case of communicable infection. When $R_0 > 1$ then the situation of epidemic results and if $0 < R_0 < 1$ then the infection will eventually die out soon.

2.2. Regression equation for time varying reproduction number (R_t)

We redefine reproduction number (R_t) as a function of t in order to depict the time dependency of the SIR model for the study of disease progression. Under the assumption, $S+I+R = N$ which is the total population for a state. The initial conditions are defined as $S(t) \geq 0$, $I(0)=0$, $R(0)=0$ converge to an equilibrium.

Using equations (1) to (4) we can redefine β and R_{t} as:

$$\beta = \left(\frac{\delta I}{\delta t} + \frac{\delta R}{\delta t}\right)\frac{N}{IS} \tag{5}$$

$$R_{\mathrm{t}} = \left(\frac{\acute{I}+\acute{R}}{\acute{R}}\right)\frac{N}{S} = \left(\frac{\Delta I(t)}{\Delta R(t)} - 1\right)\frac{N}{S} = \frac{CN}{S} \tag{6}$$

where $\acute{I}, \acute{R}$ are the partial derivates of infective and recovered cases respectively (Derivation is given in appendix), and $C = \left(\frac{\Delta I(t)}{\Delta R(t)} - 1\right)$ is calculated as the ratio of change in number of infectives to change in number of recovered cases. The values of R_{t} follows exponential distribution (observed on the basis of AIC values). Thus, we can link the parameters obtained from the above equation (6) to the exponential form by means of link function:

$$R_{\mathrm{t}} = Ae^{\alpha t} \tag{7}$$

where $A = \frac{CN}{S}$ and α is the parameter of the exponential model.

2.3. Utility function

In economic theory, utility is defined as a production function of demand and supply. It is differentiated with respect to time in order to get the preference value for a consumer at a point of time. Borrowing the same concept into health preference for different states of India, we define utility as a function of confirmed, recovered cases and R_t. The production function for utility is defined in a multiplicative form as:

$$U_t = kA_tB_t \tag{8}$$

where $A_t = \frac{1}{S(t)}$, $B_t = R_t$ and the constant k gives the ratio of estimated coefficient of variable C to the number of active infective cases.

Then the utilities for different states can be estimated by:

$$U(t) = \frac{Estimated\ coefficient\ of\ variable\ C * R_t}{Estimated\ coefficient\ of\ active\ cases * S(t)} \tag{9}$$

The state-wise utility values are calculated using equation (9). These are further multiplied with the average length of stay in hospitals in order to get the QALY values over a period of 3 months.

3. Data

The study includes data for the daily number of cases from different states of India. The daily case counts for the covid-19 infected patients from 26 March, 2020 till 28 May, 2020 has been obtained from the websites of Ministry of Health and Family Welfare (MOHFW, Government of India), Covid19India organisation, worldometer India tracker. Among the 28 states and 8 union territories, we have included 20 states and 5 union territories in our dataset. The remaining states and union territories have been excluded due to non-availability of data records for the duration of 3 months (March-May). The count for a daily case was accumulated on a weekly basis. These were further aggregated on monthly basis. The Table 4 for different values of R_t is given in appendix which has been taken from COVID-19 India organisation data operations group. The data for different states with total number of confirmed, active, recovered and death cases is presented below in Table 1.

Table 1: State wise data of COVID-19 cases as on 28 May 2020

State	Confirmed	Recovered	Deaths	Active
Andhra Pradesh	8929	4307	106	4516
Bihar	7808	5631	51	2126
Chandigarh	406	316	6	84
Chhattisgarh	2255	1421	11	823
Delhi	59746	33013	2175	24558
Gujarat	27317	19357	1664	6296
Haryana	10709	5557	161	4991
Himachal Pradesh	702	419	7	263
Jammu and Kashmir	5956	3382	82	2492
Jharkhand	2089	1406	11	672
Karnataka	9150	5618	138	3390
Kerala	3173	1659	22	1490
Ladakh	837	134	1	702
Madhya Pradesh	11903	9015	515	2373
Maharashtra	132075	65744	6170	60147
Manipur	841	250	1	591
Odisha	5303	3720	21	1562
Puducherry	383	149	8	226
Punjab	4074	2700	99	1275

Rajasthan	14997	11652	349	2996
Tamil Nadu	59377	32754	757	25866
Telangana	7802	3731	210	3861
Uttar Pradesh	17731	10995	550	6186
Uttarakhand	2344	1500	27	802
West Bengal	13945	8297	555	5093

4. Implementation and Results

Ma *et al*. (2014) have used exponential curve at the time of onset but as the spread increases it tries to flatten out in due course of time. Ghosh and Mondal (2020) have identified the number of corona positive cases in the month of March by extrapolation of exponential growth model. They have used low time axis values on the basis of sigmoid function whose growth is saturated with an assumption of 10^4 or 10^5 values in different states. In the early month of March till late May there is a deviation in the number of cases from exponential growth to non-exponential growth. Prakash *et al.* (2020) have indicated that the number of daily new cases increases as the number of cumulative infections. Ma (2020) in the initial growth phase of cumulative number of cases has derived a linear relationship with time by using a log linear scale. Since an epidemic grows exponentially in an initial phase. Guerrero (2020) forecasted the spread of virus by using logistic and SIR model combination. Giordano *et al*. (2020) have defined eight stages of an infection and called the model as SIDARTHE.

Assuming that the entire population across all the states have equal likely chance of being susceptible to infection, the best distributional fit to the reproduction number (R_t) in all the states is determined on the basis of Akaike Information Criterion (AIC) values using the fitdistrplus package in the *R* programming language as shown below in Table 2.

Table 2: Potential form of distributions with AIC values

Distribution	Log Likelihood	AIC
Normal	–46.36955	88.7391
Exponential	–31.01179	60.02358
Log Logistic	–46.01647	88.0329
Log Normal	–39.34114	74.68229
Weibull	–45.80388	87.60776
Gamma	–41.745	79.4915

The model having least AIC value is the best model. From Table 2, we can choose the distribution on the basis of least AIC value and maximum log likelihood value. Regardless of their random movement within the population, exponential distribution best fit the data. The model equation (7) indicates that

$$R_t \sim Exponential(9.397762).$$

After taking logarithm of the equation (7) and running the regression we get:

$$logR_t = -1.80 + .0813\ logN - 0.268\ logC - 0.150\ logS + 9.3977\mathrm{t} + \varepsilon \qquad (10)$$

where ε is the random error component which follows normal distribution.

The value of coefficient of determination for the above model is 0.9075. This implies that 90.75% of the total variation in the number of reproduction number which varies by time is explained by the set of confirmed, recovered and active cases. Then we differentiate the model equation in order to get the utilities for each state. This analytic model (in equation 10) serves to be better choice for estimation of quality of life for covid-19 patients. It eliminates the drawback of solving the differential equations again and again along with differentiating the likelihood function.

5. Quality Adjusted Life Year (QALY)

Quality Adjusted Life Year (QALY) is a metric used by health economists to evaluate new and innovative healthcare treatment for any particular disease. It is an important measurement of health outcome which gives the quality adjusted life years for an individual or group of individuals. Drummond *et al.* (1997) have introduced the quality of life which can be quantified by using the concept of utility. Whitehead and Ali (2010) have combined the effects of health care interventions on mortality as well as morbidity. Their definition of QALY goes around a single index termed as common currency enabling comparison across different disease areas which can further be extended to different states. Thus, QALY is a summary measure which incorporates the impact on quantity as well as quality of life.

QALD (Quality Adjusted Life Days) for childbirth and maternity service in India have also been estimated by Grover *et al*. (2019). These QALDs are estimated for different quintiles which are classified on the basis of usual monthly per capita expenditure. Deo and Grover (2020) have defined utility as a function of longitudinal covariate which is significantly associated with a disease progression. In this paper we estimate QALY by linking the utility function with the conventional epidemiological models. On the basis of utility function and average length of stay in hospital (ALOS) QALY's for different states can be estimated by:

$$QALY = Utility * Averagelengthofstayinhospital \tag{11}$$

ALOS in the hospital has been accumulated from a weekly data base. Using the datasets of total confirmed cases, recovered cases, deaths and active cases given in Table 1, we have estimated QALY for the corona virus affected patients for various states as given in Table 3 below:

Table 3: QALY's based on different states of India

State	R_t	U_t	ALOS	Q_M
Andhra Pradesh	0.15	2.215	13.17	0.317
Bihar	0.12	4.115	8.04	0.360
Chandigarh	0.06	5.415	13.54	0.797
Chhattisgarh	0.11	3.070	5.856	0.295
Delhi	0.13	2.726	12.646	0.375
Gujarat	0.13	4.861	10.21	0.539
Haryana	0.08	2.404	9.582	0.250
Himachal Pradesh	0.08	2.990	6.81	0.221
Jammu and Kashmir	0.13	2.678	13.01	0.379
Jharkhand	0.17	3.483	1.16	0.44
Karnataka	0.09	3.024	15.22	0.500

Kerala	0.07	2.386	10.26	0.66
Ladakh	0.01	1.336	25.77	0.374
Madhya Pradesh	0.16	5.620	8.23	0.503
Maharashtra	0.13	2.460	13.7	0.366
Manipur	0.06	1.594	2.26	0.39
Odisha	0.11	3.804	13.22	0.547
Puducherry	0.06	1.899	9.1	0.88
Punjab	0.11	3.580	16.55	0.644
Rajasthan	0.12	5.608	11.09	0.676
Tamil Nadu	0.17	2.572	10.74	0.300
Telangana	0.12	2.264	10.91	0.268
Uttar Pradesh	0.11	3.211	13.48	0.471
Uttarakhand	0.08	3.274	9.7	0.345
West Bengal	0.13	3.068	10.05	0.335

There is a huge variation in QALY per month (Q_M) values across different states of India. It indicates that few states are on the verge of better quality of life with QALY value closer to 1 than other states which are in worse condition with QALY value close to 0. QALY provides a better tool to policy makers for identifying how preventive measures implemented in various states have impacted differently. They help us to conclude on those states of India whose QALY value is close to 1 thus, indicating adequate lockdown and preventive measures which were taken timely in order to curb the virus. Thus the disease progression and QALY variation will help the policy makers to initiate new frameworks for states with lower quality of life for corona virus affected regions. One way of representation through QALY is done by means of classifying the states with values greater than 0.5 or less than it.

Classification I: *QALY* values > 0.5

States: Chandigarh, Gujarat, Karnataka, Kerela, Madhya Pradesh, Odisha, Puducherry, Punjab, Rajasthan

Classification II: *QALY* values < 0.5

States: Andhra Pradesh, Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jammu and Kashmir, Jharkhand, Ladakh, Maharashtra, Manipur, Tamil Nadu, Telangana, Uttar Pradesh, Uttarakhand, West Bengal

Thus from the above classification we can observe that there are 16 states which needs more preventive measures and strict lockdown guidelines in order to get better QALY values. While classification I indicates that these states also need to follow continued guidelines under covid-19 since their QALY values are not so much closer to one.

6. Discussion

India is a densely populated country with restricted infrastructure for healthcare systems in order to tackle a pandemic. With due demands of hospital beds, the state and central government are working towards creation of new corona isolation wards, medical equipment like ventilators, testing kits, personal protective equipment (PPE) kits, sanitizers, masks *etc*. Ranjan (2020) clearly states that the immediate action of lockdown imposed by

the Indian government proved to be fruitful in early spread of infection as compared worldwide.

Salman and Salem (2020) have also listed the age group and immunity developed due to BCG vaccination which has favoured lower mortality rate in India. Also the testing rate is lower in India as compared to other countries which under estimates our number of positive cases. Testing of samples was done in India with restrictions. It majorly targets those individuals which show severe symptoms of prolonged high fever, acute respiratory syndrome patients, people travelling from high risk countries with their immediate contacts, symptomatic health care workers/professionals. People with mild to moderate symptoms are advised for home quarantine with few general medications of fever, multivitamins, cough syrups, immunity booster food etc.

Ferguson *et al*. (2020) reveals that if all the countries adopt social distancing, testing and isolation of infected cases then the global death would cut down by 1.9 million by the end of 2020. Mair (2020) has stated how the different economic situations will change due to corona virus. In order to prioritise the protection of livelihoods we have to respond towards the pandemic with extreme combinations. The vaccination introduced for this infection need to be made by keeping in mind about its cost effectiveness for our population. Grover and Aggarwal (2020) have proposed cost effectiveness analysis on the basis of health outcome DALY. Shankar *et al.* (2020) have stated the mitigation strategy on how to closely monitor the effective reproduction number below one which is useful to prevent the spread.

States which lie in severe category further require lockdown measures as well as strict adherence to the guidelines of prevention to covid-19. For allay of our estimation procedures we have limited our models by considering homogenous distribution of population across all the states. It fails to capture the variations in population density for rural as well as urban India. Due to non-availability of data based on age, gender, occupation, travel history etc we could not resort to stratification on the basis of different predictors. Mandal *et al*. (2020) have stated that the probability of an infected air traveller coming back to India as the final destination which further import the risk in Delhi, followed by Mumbai, Kolkata, Bengaluru, Chennai, Hyderabad, Kochi. Menon (2020) highlights the differences among the states in terms of population density. Mumbai has higher population density with closer contacts in terms of transmission from one person to another than in comparison to spare populated Arunachal Pradesh.

The city of Maharashtra, Mumbai appears similar to epicentre Wuhan in China due to high call in the number of COVID-19 cases but the slum area Dharavi has placed an extraordinary example of combating with the virus. Due to excessive testing and following the guidelines for the prevention under covid-19 there has been speedy decline in the number of deaths and active cases. Masih (2020) list Kerala as the first state in the country to report a corona virus case. They had maximum influx of students returning from China as a carrier. But their health infrastructure followed district monitoring, risk communication, and engagement of health professionals with aggressive testing.

Singhal (2020) has listed the laboratory parameters such as white blood cell count, lymphocyte count, platelet count, procalcitonin etc which can be assessed for the estimation of quality of life when the virus hits the body. The scope of estimation for QALY can further be extended for patients who are elderly with underlying co-morbidities such as hypertension, diabetes, cardiovascular disease etc in order to study the variations with adverse outcomes.

These epidemic also teach us lessons how to build a stronger healthcare infrastructure with good investment and community engagement.

7. Conclusions

Azad and Poonia (2020) have listed short term forecasts for the infection spread across the Indian states. On the similar lines, Ghosh *et al*. (2020) have divided the states into three different zones based on daily infection rate (*DIR*) as severe, moderate and controlled. We have considered exponential regression alongwith SIR model on the dataset of different states and the analysis done by Ghosh *et al*. (2020) goes in conjunction with each other, thus fitting the scenario of infections precisely and robustly. We further establish a link between the states in terms of *DIR* and Q_M.

States with an increasing trend in *DIR* such as Maharashtra ($Q_M = 0.36$); Delhi ($Q_M = 0.37$); Bihar ($Q_M = 0.36$); Andhra Pradesh ($Q_M = 0.31$); Uttar Pradesh ($Q_M = 0.47$); Haryana ($Q_M = 0.25$); Tamil Nadu ($Q_M = 0.30$); West Bengal ($Q_M = 0.33$); Chattisgarh ($Q_M = 0.29$); Himachal Pradesh ($Q_M = 0.22$); Jammu and Kashmir ($Q_M = 0.37$); Jharkhand ($Q_M = 0.44$); Ladakh ($Q_M = 0.37$); Manipur ($Q_M = 0.39$); Telangana ($Q_M = 0.26$); Uttarakhand ($Q_M = 0.34$); Thus all the above states are densely populated with high *DIR* values and lower *QALY* values (less than 0.5). They belong to the category of severe states affected by covid-19.

States with decreasing trend in *DIR* and non-increasing growth in active cases such as Gujarat ($Q_M = 0.53$); Madhya Pradesh ($Q_M = 0.50$); Karnataka ($Q_M = 0.5$); Odisha ($Q_M = 0.54$); These states are termed as moderate regions.

States with decreasing trend in *DIR* and decreasing growth in active cases such as Kerala (Q_M =0.66); Chandigarh ($Q_M = 0.79$); Rajasthan ($Q_M = 0.67$); Punjab ($Q_M = 0.64$); Puducherry ($Q_M = 0.88$); and higher *QALY* values (greater than 0.5) will lie under controlled regions against covid-19.

Acknowledgement

We would like to thank the editors and reviewers for their comments and suggestions that helped us to significantly improve our manuscript.

References

Azad, S. and Poonia, N. (2020). Short-term forecasts of COVID-19 spread across Indian States until May 2020 under the worst-case scenario. *Arvix.org*. (doi: 10.20944/preprints202004.0491.v1)

Biswas, K., Khaleque, A. and Sen, P. (2020) . Covid-19 spread : Reproduction of data and prediction using SIR model on Euclidean network. *ArXiv.org, Cornell University*. arXiv:2003.07063

Chatterjee, P., Nagi, N., Agarwal, A., Das, B., Banerjee, S., Sarkar, S., Gupta, N. and Gangakhedkar, R. R. (2020) . The 2019 novel corona virus disease (COVID-19) pandemic : A review of the current evidence. *Indian Journal of Medical Research*, **151**, 147-159.

Covid19India.org (2020). India Covid-19 tracker.Accessed from https://www.covid19.india.org/

Dazak, P., Carroll, D., Wolfe, N. and Mazet, J. (2018) . The global virome project. *Science*, **359** , 872-874.

Deo, V., Chetiya, A. R., Deka, B. and Grover, G. (2020) . Forecasting transmission dynamics of COVID-19 in India under containment measures - A time dependent state space SIR approach. *Statistics and Applications*, **18(1)**, 157-180.

Deo, V. and Grover, G. (2020) . Evaluating quality adjusted life years in absence of standard utility values – A dynamic joint modelling approach. *Journal of Communicable Disease*, **51(3)**, 1-9.

Drummond, M. F., Stoddart, G. L., Torrance, G. W. and O'Brien, B. (1997). Methods for the Economic Evaluation of Health Care Programmes. *Oxford Medical Publication*, Second Edition. pp 316.

Ferguson, M. C., Bartsch, S. M., McKinnell, J. A., O'Shea, K. J., Siegmund, S. S. and Lee, B. Y. (2020). The potential health care costs and resource use associated with COVID-19 in United States. *Health Affairs*, **39(6)**, 1-7.

Ghosh, S. and Mondal, S. (2020). Possibilities of exponential or Sigmoid growth of Covid-19 data in different states of India. *Indian Journal of Applied Research*, **10(6)**, 1-12.

Ghosh, P., Ray, D., Bhattacharyya, R., Wang, L., Salavtore, M., Mohammed, S., Halder, A., Zhou, Y., Song, P., Purkayastha, S., Bose, D., Banerjee, M. and Baladandayuthapani, V. (2020). Predictions and role of interventions for COVID-19 outbreak in India. *Harvard Data Science Review*, Special issue-1,Covid-19, June, **2020(1)**, 1-45.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Filippo, A. D., Matteo, A. D. and Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population wide interventions in Italy. *Nature Medicine*, **26**, 855-860.

Grover, G., Das, N. R. and Magan, R. (2019). On the estimation of QALD (Quality Adjusted Life Days) based on expenditure data for Childbirth and Maternity services in India using utility theory. *Indian Journal of Economics and Development*, **7 (6)**,1-9.

Grover, G. and Aggarwal, S. (2020). A study comparing cost effectiveness of combination therapy for preventing opportunistic infections among HIV infected adults on antiretroviral therapy. *Value in Health in Regional Issues* (To appear).

Guerrero, D. (2020). Spread of COVID-19: a study case of Honduras, forecasting with Logistic Model and SIR model. *Zenodo*, 1-13.

He, X., Lau, E. H. Y. and Wu, P. (2020) . Temporal dynamics in viral shedding and transmissibility of Covid-19. *Nature Medicine*, **26**, 672-675.

Jewell, N. P., Lewnard, J. A. and Jewell, B. L. (2020) . Predictive mathematical models of the COVID-19 pandemic. *JAMA Network*, **323(19)**.

Kotwal, A., Yadav, A. K., Yadav, J., Kotwal, J. and Khune, S. (2020). Predictive models of COVID-19 in India: A rapid review. *Medical Journal, Armed Forces India*, Advance online publication. https://doi.org/10.1016/j.mjafi.2020.06.001

Ma. J. (2020). Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, **5**, 129-141.

Ma, J., Dushoff, J., Bolker, B. M. and Earn, D. J. D. (2014) . Estimating initial epidemic growth rates. *Bulletin of Mathematical Biology*, **76(1)**, 245-260.

Mair, S. (2020). How will corona virus change the world? *The Conversation, BBC Future*. 31 March 2020. https://www.bbc.com/future/article/20200331-covid-19-how-will-the-coronavirus-change-the-world

Masih, N. (2020).Aggressive testing, contract tracing, cooked meals: How the Indian state of Kerala flattened its coronavirus curve. *The Washington Post*.

Mandal, S., Bhatnagar, T., Arinaminpathy, N., Agarwal, A., Chowdhury, A., Murhekar, M., Gangakhedkar, R. R. and Sarkar, S. (2020). Prudent public intervention strategies to control the coronavirus disease 2019 transmission in India: A mathematical model based approach. *Indian Journal of Medical Research*, **151**, 190-199.

Menon, G. I. (2020). COVID-19 Pandemic: Should you believe what the models say about India? *Science*: *The Wire*.

Minisitry of Health, Family Welfare and Planning, Government of India. https://www.mygov.in/corona-data/covid19-statewise-status

Prakash, M. K. and Ansumali, S. (2020) . A very flat peak: Why standard SEIR models miss the plateau of COVID-19 infections and how it can be corrected. *medRxiv, The preprint server for health sciences , Chan ZukerbergIntiative.* https://doi.org/10.1101/2020.04.07.20055772

Ranjan, R. (2020) . Predictions for COVID-19 outbreak in India using epidemiological models. *medRxiv*. doi: https://doi.org/10.1101/2020.04.02.20051466.

Salman, S. and Salem, M. L. (2020) . The mystery behind childhood sparing by COVID-19. *International Journal of Cancer and Biomedical Research*, **5(1)**, 11-13.

Shankar, S., Kumar, A., Chatterjee, K. and Chatterjee, K. (2020). Healthcare impact of COVID-19 epidemic in India: A stochastic mathematical model. *Medical Journal Armed Forces*, **76** , 147-155.

Singhal, T. (2020). A review of Corona virus disease-2019 (COVID-19). *The Indian Journal of Paediatrics*, **87(4)**, 281-286.

Thompson, R., Stockwin, J., Van, G. R., Polonsky, J., Kamvar, Z., Demarsh, P., Dalqwist, E., Li, S., Miguel, E. and Jombart, T. (2019). Improved inference of time varying reproduction numbers during infectious disease outbreaks. *Epidemics*, **29 (100356),** 1-11.

Wallinga, J. and Lipsitch, M. (2006) . How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B*: *Biological Sciences*, **274**, 599-604.

Whitehead, S. J. and Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *Oxford University Press, British Medical Bulletin*, **96**, 5-21.

Worldometer coronavirus.https://www.worldometers.info/coronavirus/country/india

Wu, P., Hao, X., Lau, E. H. Y., Wong, J. K., Leung, K. S. M., Wu, J. T., Cowling, B. J. and Leung, G. M. (2020). Real time tentative assessment of the epidemiological characteristics of novel corona virus infections in Wuhan, China. *Euro Surveillance*, **25(3),** 1-6.

APPENDIX

1. Formulation of R_t

Using the identity from numerical analysis which links the difference operator in finite differences as:

$$(1+\delta)^n = 1 + \Delta \tag{12}$$

$$\delta = \frac{\Delta}{n} + \frac{\frac{1}{n}\left(\frac{1}{n}-1\right)\Delta^2}{2!} + \frac{\frac{1}{n}\left(\frac{1}{n}-1\right)\left(\frac{1}{n}-2\right)\Delta^3}{3!} + \cdots \tag{13}$$

Substituting the above identity in equation (6) for single term and ignoring higher order difference operators, we get

$$\delta I = \frac{\Delta}{n} I.$$

The equation (6) can be rewritten as:

$$R_t = \left(\frac{\left\{\frac{I(t+1)-I(t)}{\Delta t}\right\} + \left\{\frac{R(t+1)-R(t)}{\Delta t}\right\}}{\left\{\frac{R(t+1)-R(t)}{\Delta t}\right\}} \right) \frac{N}{S} \tag{14}$$

$$= \left(\frac{\Delta I(t) + \Delta R(t)}{\Delta R(t)} \right) \frac{N}{S}$$

$$= \left(\frac{\Delta I(t)}{\Delta R(t)} - 1 \right) \frac{N}{S} = \frac{CN}{S}$$

2. Formulation of the utility function

The utility function has been defined as the production function of susceptible, recovered and R_t.

$$U_t = f(S, R, R_t) \tag{15}$$

The marginal utility from equation (6) and (7) is obtained as :

$$\frac{\delta R_t}{\delta t} = \frac{CN}{S} e^{\alpha t} \alpha$$

$$\frac{\delta R}{\delta t} = \frac{-N(I+R)}{R_t S t^2} + \frac{N(I+R)}{St} \frac{(-1)}{R_t^2} \frac{\delta R_t}{\delta t}$$

Thus we can conclude that,

$$U_t \propto \frac{1}{S}, U_t \propto R_t \tag{16}$$

3. R_t values for different states of India

The values of R_t were accessed from covid-19 India 2020 tracker with different time points *i.e.*, t = 7, 14, 21,...., 70 days are presented below in Table 4.

Table 4: The values of R_t

State	R_1	R_2	R_3	R_4	R_5
Andhra Pradesh	0.188	0.245	0.245	0.08	0.145
Bihar	0.107	0.115	0.12	0.072	0.137
Chandigarh	0.068	0.092	0.079	0.012	0.053
Chhattisgarh	0.165	0.188	0.047	0.103	0.104
Delhi	0.134	0.154	0.22	0.107	0.131
Gujarat	0.132	0.152	0.106	0.165	0.144
Haryana	0.06	0.063	0.119	0.081	0.076

Himachal Pradesh	0.024	0.057	0.178	0.088	0.087
Jammu and Kashmir	0.127	0.229	0.128	0.1006	0.135
Jharkhand	0.148	0.167	0.189	0.222	0.245
Karnataka	0.171	0.146	0.07	0.055	0.084
Kerala	0.15	0.139	0.057	0.02	0.058
Ladakh	0.002	0.0039	0.006	0.015	0.011
Madhya Pradesh	0.197	0.258	0.182	0.136	0.162
Maharashtra	0.179	0.1387	0.15	0.127	0.130
Manipur	0.055	0.0619	0.064	0.075	0.080
Odisha	0.043	0.062	0.227	0.094	0.112
Puducherry	0.056	0.0727	0.147	0.029	0.061
Punjab	0.284	0.111	0.082	0.09	0.077
Rajasthan	0.12	0.125	0.166	0.138	0.122
Tamil Nadu	0.26	0.297	0.244	0.08	0.149
Telangana	0.309	0.146	0.158	0.076	0.103
Uttar Pradesh	0.096	0.115	0.146	0.09	0.115
Uttarakhand	0.059	0.084	0.145	0.05	0.082
West Bengal	0.145	0.192	0.143	0.113	0.123
State	**R_6**	**R_7**	**R_8**	**R_9**	**R_{10}**
Andhra Pradesh	0.156	0.178	0.117	0.189	0.120
Bihar	0.145	0.131	0.118	0.107	0.099
Chandigarh	0.137	0.128	0.119	0.114	0.112
Chhattisgarh	0.053	0.062	0.065	0.062	0.060
Delhi	0.104	0.094	0.085	0.080	0.085
Gujarat	0.131	0.121	0.112	0.105	0.099
Haryana	0.144	0.134	0.123	0.114	0.106
Himachal Pradesh	0.076	0.073	0.071	0.067	0.064
Jammu and Kashmir	0.087	0.075	0.070	0.067	0.071
Jharkhand	0.135	0.121	0.110	0.102	0.096
Karnataka	0.199	0.187	0.146	0.122	0.118
Kerala	0.084	0.076	0.071	0.068	0.068
Ladakh	0.058	0.051	0.045	0.042	0.041
Madhya Pradesh	0.011	0.017	0.020	0.021	0.021
Maharashtra	0.162	0.145	0.130	0.120	0.111
Manipur	0.130	0.124	0.115	0.109	0.104
Odisha	0.076	0.054	0.048	0.032	0.018
Puducherry	0.112	0.104	0.104	0.105	0.102
Punjab	0.061	0.054	0.051	0.051	0.055
Rajasthan	0.077	0.089	0.088	0.082	0.075
Tamil Nadu	0.122	0.110	0.101	0.094	0.088
Telangana	0.149	0.139	0.134	0.126	0.118
Uttar Pradesh	0.103	0.090	0.080	0.074	0.069
Uttarakhand	0.115	0.106	0.097	0.090	0.085
West Bengal	0.082	0.073	0.066	0.062	0.070

Statistics and Applications {ISSN 2452-7395(online)}
Volume 18, No. 1, 2020 (New Series), pp 333–350

Epidemiological Dynamics of the COVID-19 Pandemic in India: An Interim Assessment

Sitabhra Sinha
The Institute of Mathematical Sciences
CIT Campus, Taramani, Chennai 600113, India

Received: 20 July 2020; Revised: 29 July 2020; Accepted: 30 July 2020

Abstract

We have analyzed the time-series for the number of active cases of COVID-19 pandemic in India, as well as, in other countries around the world using a variety of statistical fitting procedures. We obtain robust estimates of the exponential growth rate for the number of active cases, which is then used for calculating the reproduction number of the epidemic. We estimate the basic reproduction number of COVID-19 epidemic in India to be $R_0 \sim 1.82 \pm 0.02$, a value that lies at the lower end of the spectrum of values of different regions around the world where there have been major outbreaks of the disease. We have also investigated the change in the effective reproduction number over time, particularly following the introduction of unprecedented non-pharmaceutical interventions such as the stay-at-home order (lockdown) imposed over the entire country from 24 March 2020, and continued at varying levels of strictness, and with regional variations, up to the present (July). We observe that the reproduction number showed a large reduction within a couple of weeks of the imposition of lockdown, suggesting that this measure played a role (along with others such as compliance with physical distancing rules in public and use of masks) in reducing the rate of spreading of the contagion, although it was unable to break the chain of infection. We also note that there is considerable regional variation across India in the dynamics of the epidemic, with different regions registering rise and fall in the growth rate of the disease at different times.

Key words: COVID-19; Corona virus; Reproduction number; Epidemiological dynamics; Pandemic.

AMS Subject Classifications: 92D30, 62M10

1. Introduction

The rapid spread of Coronavirus disease 2019 (COVID-19), that results from infection with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pathogen, has in the few months following its initial identification in December 2019 (Wang, *et al.*, 2020) not only brought to mind recent pandemics such as the 2009 swine flu pandemic that is believed

Corresponding Author: Sitabhra Sinha
Email: sitabhra@imsc.res.in

to have affected around $11\% - 18\%$ of the world population (see Kelly, *et al.*, 2011), but also raised the specter of it eventually growing to rival the 1918-20 "Spanish Flu" pandemic. Believed to be one of the deadliest pandemics in recent history, the 1918-20 pandemic caused between 15 million (Spreeuwenberg, *et al.*, 2018) and 50 million deaths (Johnson and Mueller, 2002), with a third of these occurring in British-ruled India (Reyes, *et al.*, 2018). Indeed, in scale, COVID-19 has already surpassed the other two coronavirus epidemics of recent times, *viz.*, the 2002-04 SARS outbreak and the 2012-14 MERS outbreak. Compared to the 8439 cases that were reported for SARS worldwide (WHO, 2003), and the even smaller 2500+ cases of MERS from 2012 till date (WHO, 2020a), there have already been, as of 15 July 2020, over 13 million confirmed cases of COVID-19 and 574,466 deaths (WHO, 2020b) spread across more than 200 countries and territories.

In the absence of any possibility that a vaccine for COVID-19 will be available in the immediate future, the epidemic is also proving to be a testing ground for extreme non-pharmaceutical interventions that different countries have implemented in order to contain the spread of the disease. The principal among these are the stay-at-home orders (colloquially referred to as lockdowns) imposed on the populace of several countries around the world by their governments [see Wikipedia (2020) for a list] after its apparent success in containing the initial outbreak of the disease in Hubei province of China [Lai, *et al.* (2020)]. Such unprecedented measures have provoked controversy, not least in India where an initial 21-day lockdown was initiated on 24 March 2020 with the stated purpose of bringing the epidemic to an end by breaking the chain of infection through enforced distancing [see India Today (2020)]. While the physical isolation of individuals through such measures is very likely to restrict the likelihood of transmission of the pathogen from infected to susceptible individuals, there are high social and economic costs accompanying such a measure. Lockdowns also affect different sections of the population asymmetrically, and can result in aggravating existing inequalities in society - making it unsuitable for prolonged use. Thus, following the initial period of nationwide lockdown in India, it is now primarily being applied in a more restrictive manner at specific locations where the number of active cases is increasing in a particularly alarming rate, to ensure that the medical infrastructure is not overwhelmed by the rising number of infected individuals who need to be hospitalized.

To gauge the efficacy of such non-pharmaceutical interventions, it is imperative to understand the epidemiological dynamics of this novel infectious disease - particularly, as it manifests in diverse manners at different locations. In this paper, the transmissibility of COVID-19 has been investigated with special focus on India, but also considering many other locations around the world where there have been major outbreaks. For this purpose, we have estimated at each location the reproduction number of the disease (the basic reproduction number R_0 and effective reproduction number R are defined in the next section), which measures how rapidly the number of active cases of the disease changes over time (*active cases* refers to the individuals who are infected with the disease at a given time and who can potentially infect others by passing the pathogen to non-infected individuals, *e.g.*, via contact). As can be easily explained using the mathematical theory of epidemics, the reproduction number has to be greater than 1 for an epidemic to occur, and the larger the number, the faster the disease will spread. Although several studies have already appeared

that compute the basic/effective reproduction numbers for different locations at different times, as the numerical value of the reproduction number typically depends to an extent on the exact model used to estimate it from data, most of these numbers that appear in different studies cannot be compared to each other. Thus, in order to compare the epidemiological dynamics of COVID-19 across different geographical regions and temporal phases of the epidemic, the corresponding values of R_0 and R need to be calculated in a consistent manner across both space and time. With this aim in view, in this paper the reproduction numbers for the epidemic have been calculated not only for India but also for several countries where there have been major outbreaks, as well as, for different regions within India. This allows us to obtain an understanding of the spatio-temporal diversity in the spreading dynamics of the disease within India, apart from that between India and other parts of the world. The paper is organized as follows. In Section 2, we briefly describe the sources of the data used for the analysis and the method of calculating the key epidemiological parameters R_0 and R. In Section 3, the international situation is discussed with analysis of the epidemiological dynamics for the world as a whole and that of selected countries. In Section 4, we focus on India with data for the entire country, as well as for individual states and districts. We conclude with a discussion on the limitations of the study and its implications in Section 5.

2. Data Sources and Methods

Data aggregated for the world as a whole and at the level of each country affected by the epidemic was obtained from CSSE (2020), an online data repository on GitHub of COVID-19 cases worldwide that is operated by the Johns Hopkins University Center for Systems Science and Engineering. Information about the cumulative number of confirmed cases, deaths and recovered cases are updated daily, beginning from 22 January 2020. The data is collated from a large number of sources, such as various national government health departments, as well as, from the World Health Organization (WHO), the US and European CDCs and aggregating sites such as *WorldoMeters*. Disaggregated data for India was obtained from COVID19-India (2020), a crowdsourced database of COVID-19 cases. Its volunteers collate information from health bulletins issued periodically by various governmental organizations, as well as other sources, and compile the obtained numbers to create district-level and state-level daily time-series for confirmed cases, recovered cases, deaths, active cases and number of individuals tested for the disease. At the state-level, data is available from 14 March 2020, while for the bulk of the districts the time-series information is obtainable from 21 April 2020 onwards.

We consider the time-evolution of the number of active cases, *i.e.*, the number of individuals who remain infected with the virus on a particular date, which is obtained by subtracting the cumulative number of deaths and recoveries announced up to that date from the cumulative number of confirmed cases till then. During the initial phase of an epidemic, the number of infected individuals is expected to increase exponentially with time, as is the case for any multiplicative process (such as, a chain reaction) where the value adopted by a variable at each instant is obtained by multiplying the value at the immediately preceding instant by a constant factor. This is easy to see from the mathematical models of epidemiological dynamics that stem from the pioneering work of Kermack and McKendrick (1927) [for an example of how such theoretical modeling can accurately describe the empirical data

from influenza epidemics see Spicer and Lawrence (1984)]. In the most basic setting, one can divide the entire population comprising N individuals into three compartments, corresponding to those who are *susceptible* to contracting the disease ($\mathcal{S}$), those who are at present *infected* ($\mathcal{I}$), and those who have *recovered* or are *removed* by death ($\mathcal{R}$). Neglecting any demographic changes during the period that one is considering, the time-evolution of the number of individuals in each compartment can be described by the system of differential equations:

$$\frac{d\mathcal{S}}{dt} = -\beta\mathcal{S}\mathcal{I}, \frac{d\mathcal{I}}{dt} = \beta\mathcal{S}\mathcal{I} - \gamma\mathcal{I}, \frac{d\mathcal{R}}{dt} = \gamma\mathcal{I}, \tag{1}$$

where the parameters β is the rate of infection transmission through contact between an infected and a susceptible individual, while γ is the recovery rate ($= \tau^{-1}$, *i.e.*, the reciprocal of the time duration during which a person is free to pass on the infection to others). It is easy to see that as the total population is conserved ($\mathcal{S} + \mathcal{I} + \mathcal{R} = N$), only two of the equations are independent. Furthermore, at the earliest stage of the epidemic, we can assume the susceptible population size to be effectively equal to the total population N, and thus a constant. Thus, we are left with a single differential equation that describes the evolution of $\mathcal{I}$. Normalizing the variables by the total population size N and solving the equation, we see that the fraction of infected individuals in the population $i(= \mathcal{I}/N)$ will evolve from its initial value i_0 as $i(t) = i_0 \exp(\{N\beta - \gamma\}t) = i_0 \exp(\{[R_0 - 1]/\tau\}t)$. Here the parameter $R_0 = N\beta/\gamma$ is the key epidemiological parameter *basic reproduction number*, which is defined as the average number of secondary infections that results from a primary infection at the earliest stage of the epidemic, i.e., before a significant fraction of the population has been exposed to the disease. Note that the expression of R_0 remains unchanged even if we augment this basic model with an additional compartment $\mathcal{E}$ for the subpopulation of *exposed* individuals who have been infected but are not infectious, taking into account the latent or pre-infectious period after an infection.

Thus, R_0 can be estimated from the empirical time-series of the number of infected individuals (*i.e.*, the active cases) by accurately fitting it to an exponential growth curve, *viz.*, $i(t) \sim \exp(\lambda t)$, and obtaining the most reliable estimate for λ. Using the equivalence $\lambda = (R_0 - 1)/\tau$, and equating τ with the *generation time*, i.e., the mean interval between a person getting infected by another individual (the "infector") and the time at which the "infector" was infected, R_0 can be calculated from the data. We have used a generation time of 5.2 days that was estimated by Ganyani, *et al.* (2020) from the Singapore cluster of cases. The fitting procedure is carried out using a nonlinear least squares approach implemented by the function *fit* in *MATLAB* R2009b software (Mathworks (2009)). We have also obtained the 95% confidence bounds using the function *confint* which does the calculation through QR decomposition of the Jacobian. To assess the quality of fitting, we calculate the correlation coefficient r between the logarithm of the number of active cases and time, as well as the p-value indicating the measure of significance. We only use those estimates of λ for calculating R_0 for which $r > 0.99$ and $p \leq 0.002$. We have earlier shown in Jesan, *et al.* (2011) that using the above fitting procedure yields values of R_0 that are consistent with those calculated using alternative methods, such as bootstrapping. To aid fitting when the data exhibits large

fluctuations we have performed smoothing using a 3-day moving average.

Once the epidemic has had time to penetrate substantially into the population and/or containment measures put in place have had a discernible effect, it is no longer possible to view the process as a contagion freely infecting every contact of an infected person. At this stage, we speak of the growth rate in terms of the time-dependent *effective reproduction number*, R. As in the case of R_0, R is also defined as the mean number of infections arising from a single infected individual (with the difference that now we can no longer assume the population to be almost entirely susceptible to the disease). Thus, it can be estimated using the above technique provided that the susceptible population does not decrease perceptibly over the period in which the estimation is being done. Also, similar methods can be applied to calculate the reproduction number at different spatial scales. Obviously the smaller the area being considered the smaller is the total population, so that it becomes more likely that there will be discernible changes in the susceptible population as the epidemic progresses and consequently one has to be more careful in using the above procedure.

3. COVID-19: International Scenario

The outbreak of a novel disease (to be named COVID-19 eventually) came to the attention of public health authorities towards the end of December 2019 with the occurrence of a large number of pneumonia cases of unknown causes in Wuhan, the capital of Hubei province in China. Huang, *et. al.* (2020) have traced the earliest human infected case to 01 Dec 2019, although it is likely that the virus had been circulating in the population even earlier. Subsequently, the extremely rapid rise of the number of infections made the authorities impose unprecedented city-wide stay-at-home orders (lockdown) in Wuhan and other cities in the province on 23 January 2020. Less restrictive measures for ensuring physical distancing were introduced in several other locations in China. As a consequence, there was discernible decrease in the rate of growth in infections and from 17 February 2020, the number of active cases began to decline. However, as is evident from the time-series shown in Figure 1 (left), the active case count for the entire world started to increase again from 5 March 2020. This resulted from the focal point of the epidemic shifting outside China (where it continued to decline) to countries such as Italy and the United States of America. Indeed, by 8 March 2020, the total number of active cases of the disease outside China exceeded that from China for the first time, and while there have been subsequently resurgences of COVID-19 cases in China, these have been fairly limited in size. With the rapid spread of the epidemic across different countries in Europe and the Americas, apart from Asia, on 11 March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak to be a global pandemic.

Figure 1 (right) shows the reproduction number R estimated from the time-series of the total number of active cases across all countries by using a moving window having different starting dates (t) and interval lengths (Δt). The choices of t and Δt for which the correlation coefficient r between time and logarithm of the number of active cases (that measures how closely the curve describing the number of active cases fits an exponential function) is greater than 0.998 are indicated within the black dotted contour lines (the regions within the blue dotted lines have $r > 0.995$). The corresponding measure of significance is $p < 0.001$. As

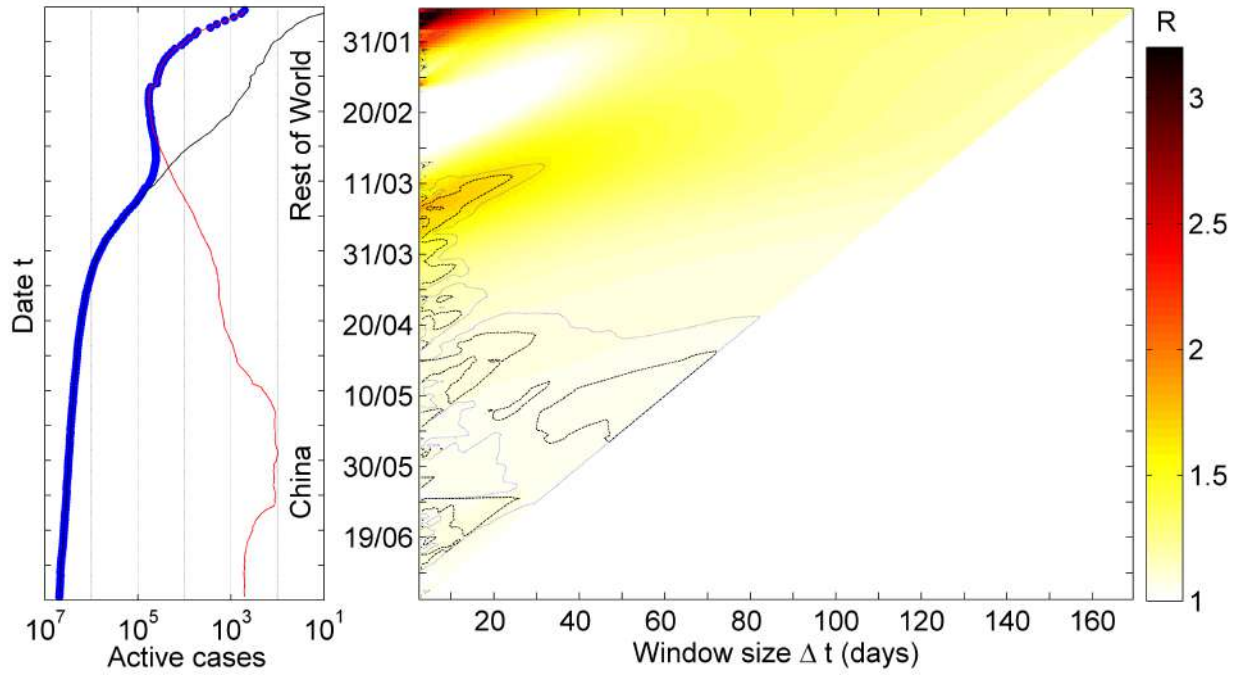


Figure 1: (Left) The time-series for the number of active cases of COVID-19 for the entire world (the data for China and the rest of the World are also shown separately) and (right) the estimated reproduction number R over time windows with different starting dates (t) and temporal intervals (Δt).

the data source we have used does not have information for the period prior to 22 January 2020, we cannot estimate the basic reproduction number. We note that R was 2.02 ± 0.16 between 23-27 January 2020, but then decreased to 1.85 ± 0.13 (02-05 February 2020) and subsequently to 1.31 ± 0.06 (07-10 February 2020). Throughout the second half of February R appeared to be equal or less than 1 so that, even though there were countries such as Italy and United Kingdom that had large outbreaks during the period, it still seemed possible that the disease can be contained and prevented from becoming a global pandemic. However, it started rising again in March, increasing from 1.18 ± 0.04 (06-09 March 2020) to 1.66 ± 0.02 (09 March-01 April 2020) - possibly resulting from the large number of outbreaks that occurred across countries in Europe and Latin America, as well as, USA and South Africa, at this time. From the month of April onward, however, we have seen a steady decrease in the global R, from 1.25 ± 0.01 (03-13 April 2020) to 1.16 ± 0.01 (12-21 April 2020) and 1.12 ± 0.01 (20-28 April 2020), notwithstanding the fact that new territories have been affected by the disease. Over the last couple of months, R has stood at 1.065 ± 0.001 (28 April-08 July 2020) which probably reflects the success of European countries in containing the epidemic and the fact that in USA, despite the large absolute number of cases, the growth rate has decreased substantially.

As the United States of America has the highest number of confirmed cases and is therefore contributing to the value of R for the entire world more than any other country, in Figure 2 we specifically look into how the situation has evolved there, beginning from

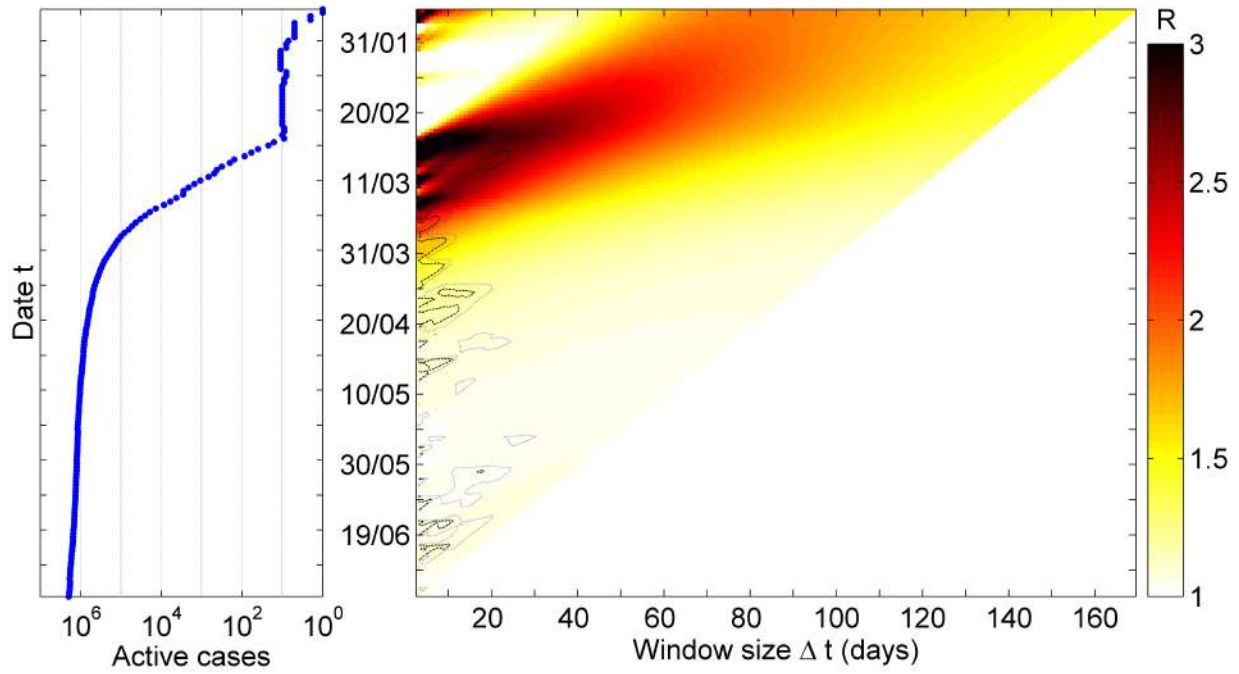


Figure 2: (Left) The time-series for the number of active cases of COVID-19 for United States of America and (right) the estimated reproduction number R over time windows with different starting dates (t) and temporal intervals (Δt).

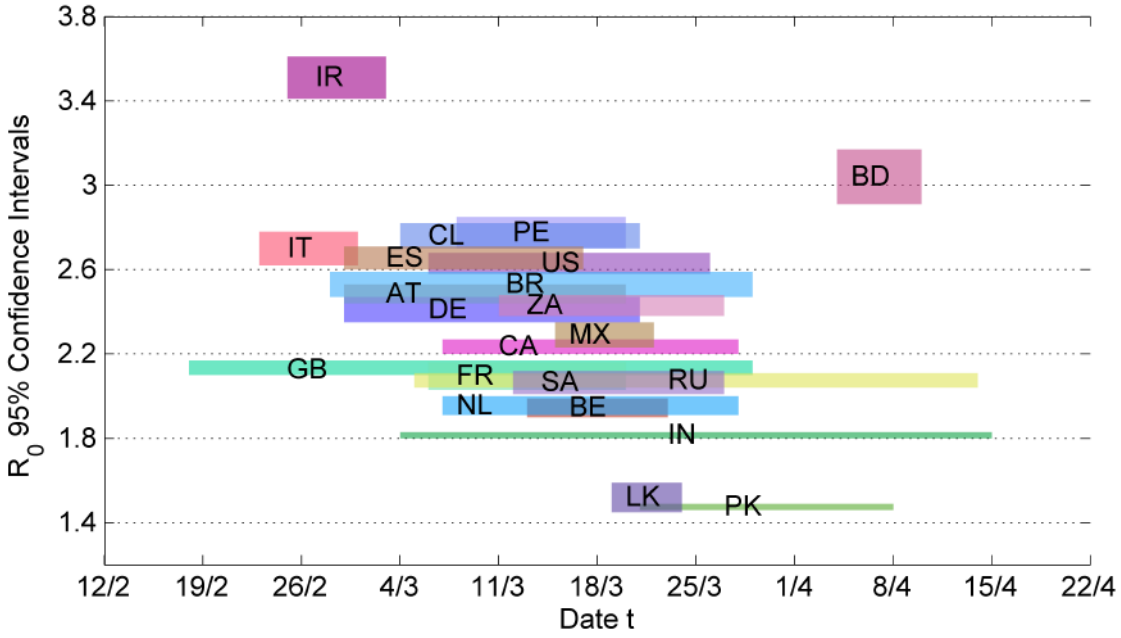


Figure 3: The 95% confidence intervals (represented by the vertical extent of the colored bars representing different nations) for the basic reproduction number R_0 estimated for COVID-19 outbreak in 22 countries across the world with the horizontal extent of the colored bars indicating the time period used for the estimation.

late January. The first case of COVID-19 in USA was reported on 19 January 2020 from Snohomish County in Washington state (Holshue, *et al.*, 2020). Following this, as seen from Figure 2 (left), the number of active cases rose only slightly to reach double digits and remained steady at fairly low numbers (~ 10) for the entire month of February. However, between 03-26 March 2020, the number of cases rose rapidly marking the advent of the epidemic with a basic reproduction number $R_0 = 2.63 \pm 0.05$. The case fatality ratio (CFR) was also high at this early period, peaking at 0.114. It is to be noted, however, that there is a high degree of heterogeneity in the disease incidence across the country, with the state of New York (followed by California, Florida and Texas) accounting for a large fraction of the cases.

As the epidemic unfolded, the value of the effective reproduction number has undergone several changes as can be seen from Figure 2 (right). Between 27 March-4 April 2020, the number reduced from the value of R_0 (mentioned above) to 1.68 ± 0.03. It further reduced to 1.38 ± 0.02 during 4-11 April 2020, 1.18 ± 0.01 during 11-26 April 2020, 1.10 ± 0.01 during 30 April-10 May 2020, and reached its lowest value so far $R = 1.05 \pm 0.01$ in the first half of June (1-18 June 2020). In more recent periods, it has marginally increased back to 1.10 ± 0.01 (22 June-1 July 2020). The trajectory of the epidemic in USA as described by these reproduction numbers is qualitatively similar to that seen for India (described below), with an initial period of extremely rapid spread lasting for about a month followed by gradual reduction in the transmission, with R eventually settling to a value just higher than 1 about four months after the outbreak established itself in the local population.

To see how much variation there is across geographical regions in the rate at which the epidemic has spread, in Figure 3 we graphically represent the 95% confidence intervals of the basic reproduction number for several countries where there have been major outbreaks of COVID-19, along with the period corresponding to the initial phase of growth of the epidemic (over which R_0 has been estimated). China is not included because as mentioned earlier, the data sources being used do not include information on the initial phase of the outbreak in China. The two letter symbols associated with each colored bar indicate the different countries (see Table 1, which provides the numerical values of R_0 for these and several additional countries). Note that all the countries which are currently in the top 15 in terms of confirmed cases have been included.

Even a cursory glance at Figure 3 is sufficient to establish a few exceptional features underlining the diversity in COVID-19 epidemiological dynamics in different locations. While the bulk of the countries investigated have had their R_0 values lying between 2 and 2.8, there have been exceptions such as Iran, which had an unusually high R_0. In contrast, R_0 for countries in South Asia such as Sri Lanka, Pakistan and Nepal (see Table 1) have been very low, with the notable exception of Bangladesh which had a R_0 of around 3. The R_0 for India, while higher than its southern and western neighbors, is still at the lower end of the range of values for the basic reproduction number that we have estimated for different countries.

4. COVID-19: The Situation in India

Having discussed the international situation, we now focus on how the epidemic has

Table 1: Basic reproduction numbers (R_0) estimated for COVID-19 outbreaks in different countries

Region	Country	R_0	95% CI	r	p	Period	Peak CFR
North America	USA (US)	2.63	[2.58, 2.68]	0.998	< 0.001	03-26 Mar 2020	0.114
	Mexico (MX)	2.29	[2.23, 2.35]	0.999	< 0.001	15-22 Mar 2020	0.124
	Canada (CA)	2.23	[2.20, 2.27]	0.998	< 0.001	07-28 Mar 2020	0.082
South America	Brazil (BR)	2.53	[2.47, 2.59]	0.995	< 0.001	26 Feb-29 Mar 2020	0.070
	Peru (PE)	2.77	[2.70, 2.85]	0.998	< 0.001	08-20 Mar 2020	0.042
	Chile (CL)	2.76	[2.70, 2.82]	0.998	< 0.001	04-21 Mar 2020	0.022
Europe	Italy (IT)	2.70	[2.62, 2.78]	0.999	< 0.001	23 Feb-01 Mar 2020	0.145
	Spain (ES)	2.66	[2.60, 2.71]	0.998	0.001	29 Feb-Mar 17 2020	0.122
	UK (GB)	2.13	[2.10, 2.17]	0.995	< 0.001	18 Feb-29 Mar 2020	0.155
	France (FR)	2.10	[2.03, 2.16]	0.995	< 0.001	06-20 Mar 2020	0.159
	Belgium (BE)	1.94	[1.90, 1.99]	0.998	< 0.001	13-23 Mar 2020	0.165
	Netherlands (NL)	1.96	[1.91, 2.00]	0.995	< 0.001	07-28 Mar 2020	0.129
	Germany (DE)	2.41	[2.35, 2.47]	0.995	< 0.001	29 Feb-21 Mar 2020	0.047
	Austria (AT)	2.49	[2.44, 2.53]	0.998	< 0.001	29 Feb-20 Mar 2020	0.040
	Russia (RU)	2.08	[2.04, 2.11]	0.995	< 0.001	05 Mar-14 Apr 2020	0.015
Africa	South Africa (ZA)	2.43	[2.38, 2.48]	0.998	< 0.001	11-27 Mar 2020	0.022
Middle East	Iran (IR)	3.51	[3.41, 3.61]	0.999	< 0.001	25 Feb-03 Mar 2020	0.079
	Turkey (TR)	4.75	[4.52, 4.97]	0.998	< 0.001	14-21 Mar 2020	0.028
	Saudi Arabia (SA)	2.06	[2.01, 2.12]	0.996	< 0.001	12-27 Mar 2020	0.015
South Asia	India (IN)	1.82	[1.80, 1.83]	0.998	< 0.001	4 Mar-15 Apr 2020	0.036
	Pakistan (PK)	1.48	[1.46, 1.49]	0.998	< 0.001	21 Mar-08 Apr 2020	0.024
	Bangladesh (BD)	3.04	[2.91, 3.17]	0.998	< 0.001	04-10 Apr 2020	0.128
	Sri Lanka (LK)	1.52	[1.45, 1.59]	0.995	< 0.001	19-24 Mar 2020	0.037
	Nepal (NP)	1.48	[1.47, 1.49]	0.995	< 0.001	25 Mar-1 Jul 2020	0.007

developed in India. The first confirmed case of COVID-19 in India was recorded when an Indian student at Wuhan returned to Kerala on 30 January 2020. Subsequently two more Indian students returning from Wuhan were also tested to be positive for the disease in early February 2020. All of them subsequently recovered without having passed the infection to anybody else in India. However, the situation altered beginning from 02 March 2020, when an Indian citizen who had traveled to Austria tested positive in East Delhi, who was soon found to have infected six of his family members in Agra (all of whom tested positive on March 4). In a parallel development, after a member of an Italian tourist group tested positive on 03 March 2020, 16 other members of the group were found to have contracted the infection on the next day. Thus, with 22 new infected individuals being detected on 04 March 2020, the number of COVID-19 cases in India crossed single digits for the first time (see Figure 4). Following this the number of new cases steadily rose [see Figure 5 (left)], alarming the public health authorities into considering implementing extreme non-pharmaceutical intervention measures such as those successfully employed to control the epidemic in China. A "Janata curfew" (People's or self-imposed curfew) for 14 hours on 22 March 2020, a Sunday, tested the feasibility of imposing a nation-wide lockdown, and it was followed up by a stay-at-home order implemented from 24 March 2020. Figure 4 shows also the composition of the new cases reported each day between those returning from abroad and those who contracted it

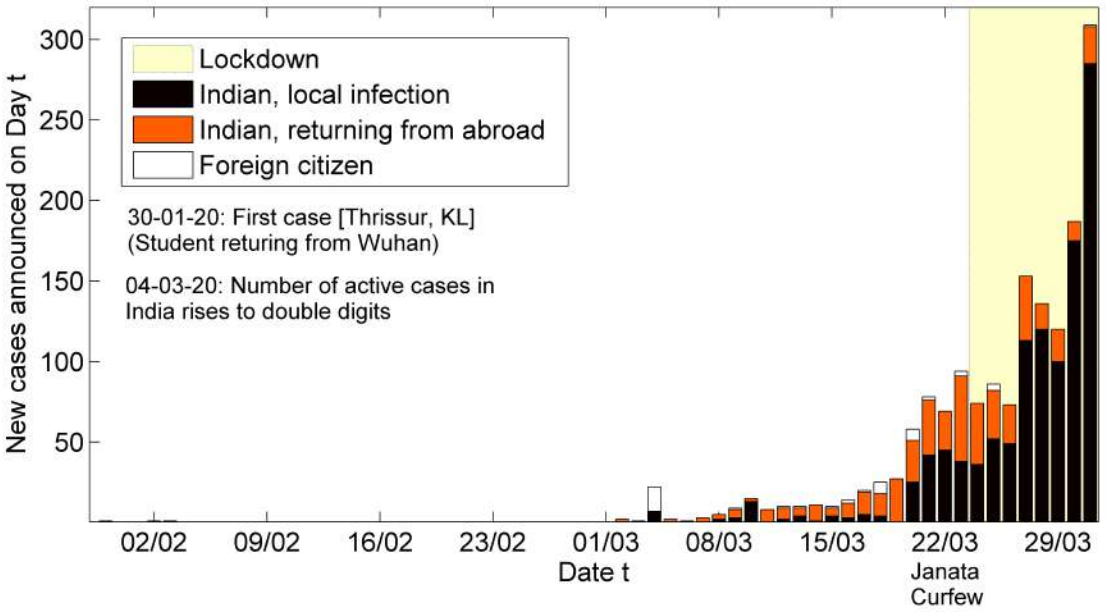


Figure 4: Initial phase of the COVID-19 epidemic in India, showing the first 1635 **confirmed cases and distinguishing between infections that were imported from abroad, either by returning Indians or visiting foreign citizens, and those that occurred among the local population.**

locally, and we note that, after 24 March 2020, the bulk of the cases were of the latter kind. It suggests that by this time, the disease had established itself in the local population.

Figure 5 shows how the reproduction number for the disease has evolved in India in response to the various measures that were successively put in place. Specifically, we indicate the various stages of the lockdown (differing in terms of the severity of the measures imposed to ensure social distancing) that was imposed from 23 March 2020, *viz.*, Phase 1 (23 March-14 April 2020), Phase 2 (15 April-3 May 2020), Phase 3 (4-17 May 2020), and Phase 4 (18-31 May 2020), followed by Unlock 1 (1-30 June 2020) and Unlock 2 (1-31 July 2020) which is still underway. In the initial stage between 04 March-15 April 2020 over which the R_0 value is estimated, the number of active cases rose from 25 to 10485. After this period, the rate of spreading lessened to a large extent and the effective reproduction number between 14 April-16 May 2020 was estimated to be 1.28 ± 0.01, which is a reduction of 30% from the value of $R_0 (\simeq 1.82)$. To understand the significance of this change in R, we note that had the epidemic continued with its initial growth rate for much longer, then we would have been very likely to have crossed 1 million active cases before the middle of May (see Figure 6). This would have undoubtedly put enormous stress on the medical infrastructure of the country.

The reduction in R may be attributed at least partially to the imposition of the lockdown and other related measures (such as, asking people to wear masks in public, *etc.*), especially as the deviation from the initial trend can be observed from the data by 06 April 2020 (see Figure 6), *i.e.*, after approximately two weeks following the imposition of Phase

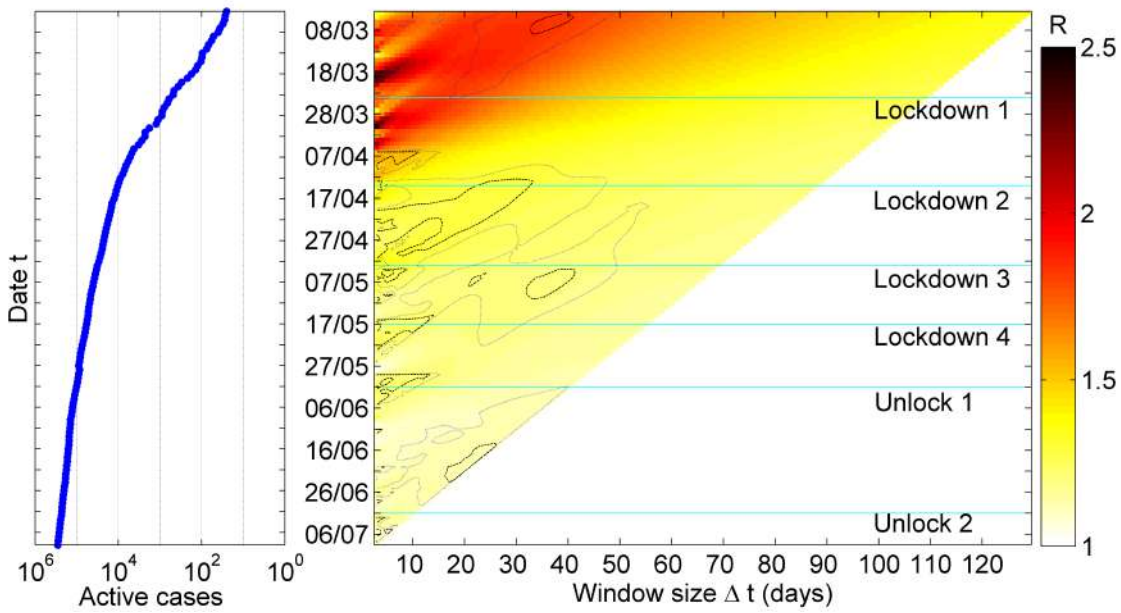


Figure 5: (Left) The time-series for the number of active cases of COVID-19 for India and (right) the estimated reproduction number R over time windows with different starting dates (t) and temporal intervals (Δt). The periods of different stay-at-home orders (referred to as Lockdown and Unlock) are indicated.

1 of the lockdown. The duration of this lag between an intervention and its manifestation in terms of changes in the number of cases is because a person infected with SARS-CoV-2 can take up to 14 days to manifest symptoms, upon which time they can be tested and then quarantined. Until this time, such individuals may be freely circulating in the population and aiding in the transmission of the pathogen. Thus, the bulk of the confirmed cases that were reported in the days immediately following the lockdown imposition would have resulted from infections that took place in the period prior to it.

Between 16-28 May 2020, the value of R slipped further to 1.22 ± 0.01, followed by a marginal decrease to $R = 1.21 \pm 0.01$ during 29 May-11 June 2020. Subsequently, during 12 June-11 July 2020, R reached its lowest value (up till the time of writing) of 1.13 ± 0.01. However, this continually decreasing trend in R was then broken and the most recent value was estimated to be 1.19 ± 0.01 between 11-20 July 2020. As already hinted in the previous section, this trajectory of the spreading dynamics, decrease in R for four months followed by a slight upward turn, resembles that of USA. Taking into account the approximately two-week delay between an event involving a population and its effect manifesting in the epidemiological data, the present rise in R can be possibly related to the relaxation of lockdown norms in the second half of June.

To obtain a better understanding of the temporal variation in the epidemiological dynamics, we should consider more spatially detailed information. As seen from the pie chart in Figure 7, the disease has not affected all regions of the country uniformly. Maharashtra

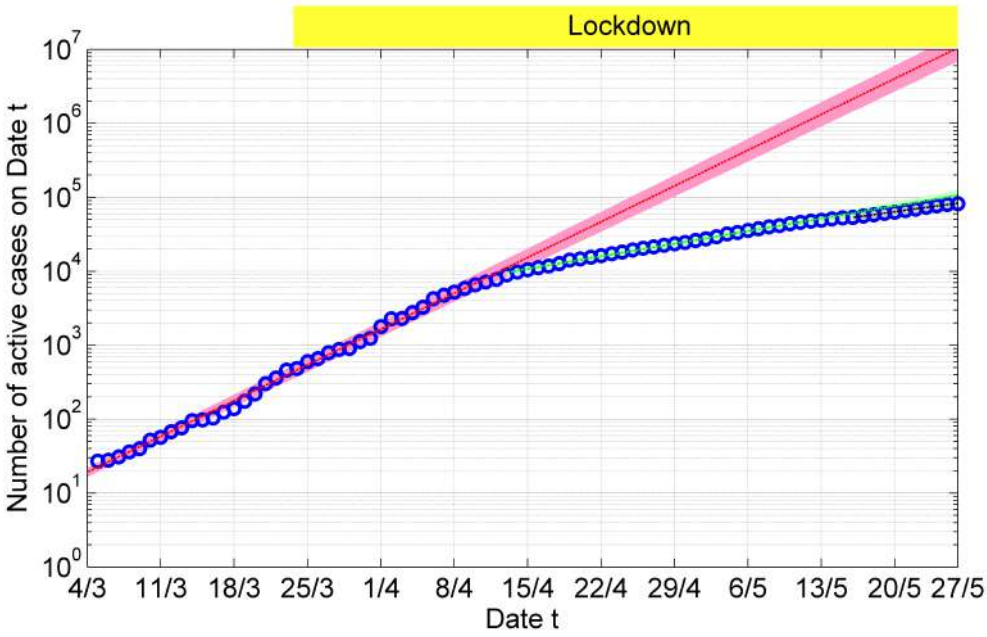


Figure 6: The progress of COVID-19 epidemic in India between 04 March-27 May 2020 showing the daily number of active cases (circles) in logarithmic scale. Log scale is used to visualize the quality of fit of the data to an exponential curve, that manifests as a straight line in such a scale. The red dotted curve indicates the projected increase in active cases in April and May had the epidemic continued to progress according to the rate given by the basic reproduction number of 1.82. The 95% confidence intervals are indicated by the shaded regions, and the period under lockdown by the colored horizontal bar on top.

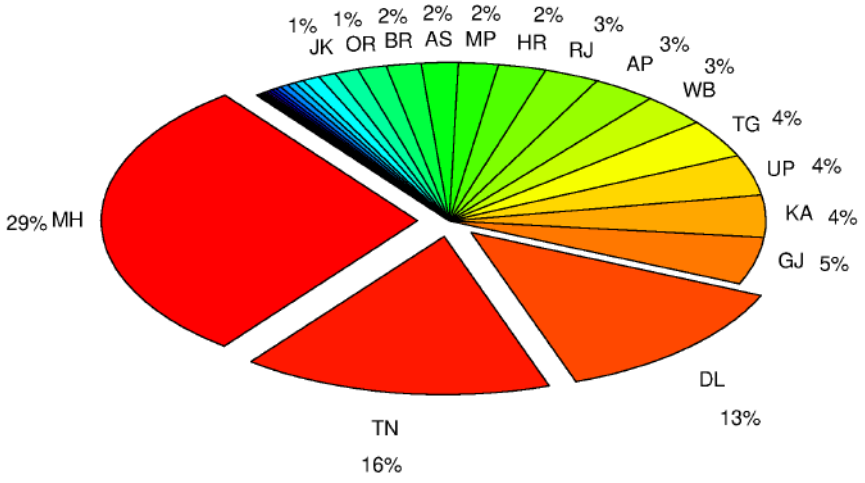


Figure 7: Pie chart showing the percentage contribution of the different states of India to the total number of confirmed cases of COVID-19 till 11 July 2020.

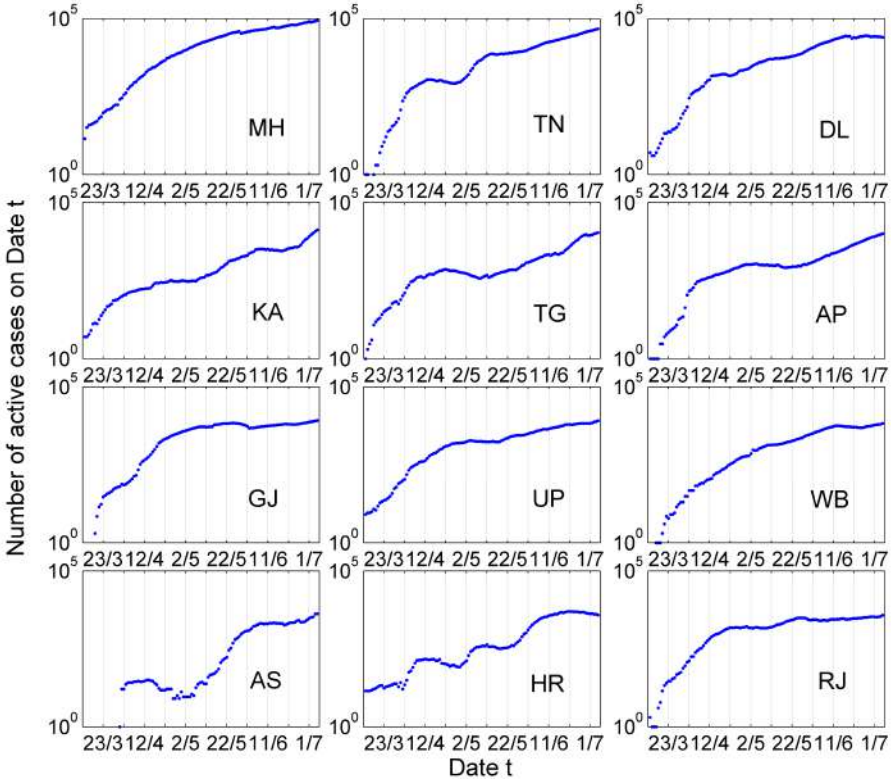


Figure 8: Progress of the epidemic in Indian states having the highest number of active cases on 05 July 2020 (in decreasing order, left to right & top to bottom).

has contributed to almost a third of the entire burden of the country, and along with Tamil Nadu and Delhi, are the three states which account for about 60% of all confirmed cases till date. On the other hand, both Kerala and Punjab, which were some of the first states to be affected, surprisingly had less than 1% of the total number of confirmed cases till July. Figure 8 also suggests a substantial amount of spatio-temporal heterogeneity in the manner in which the epidemic has evolved across the country, with the disease peaking early in west (*e.g.,* Maharashtra) and north (*e.g.,* Delhi), followed by the south (*e.g.,* Karnataka) and much later by the east (*e.g.,* West Bengal and Assam). While many of the states show a trajectory similar to that for the country, *viz.* a rapid growth phase initially, followed by a slowing of the spreading while continuing to be an epidemic (i.e., $R > 1$), certain states like Tamil Nadu and Haryana exhibit multiple rounds of growth and decay of the epidemic. Thus, between 12-30 April 2020, Tamil Nadu had a value of R around or less than 1, so that the number of active cases were decreasing over time indicating that more recoveries were happening than new infections. [1] However, the appearance of a cluster of cases originating in the Koyambedu wholesale market in Chennai towards the end of this period resulted in the

[1]We would like to note here that $R \sim 1$ does not necessarily imply that the number of active cases has remained constant, for instance, because the number of new infections reported each day is exactly balanced by the number of infected individuals who are recovering daily. Rather, the growth is slower than exponential (sub-exponential), *e.g.,* following a trend that fits a polynomial trend.

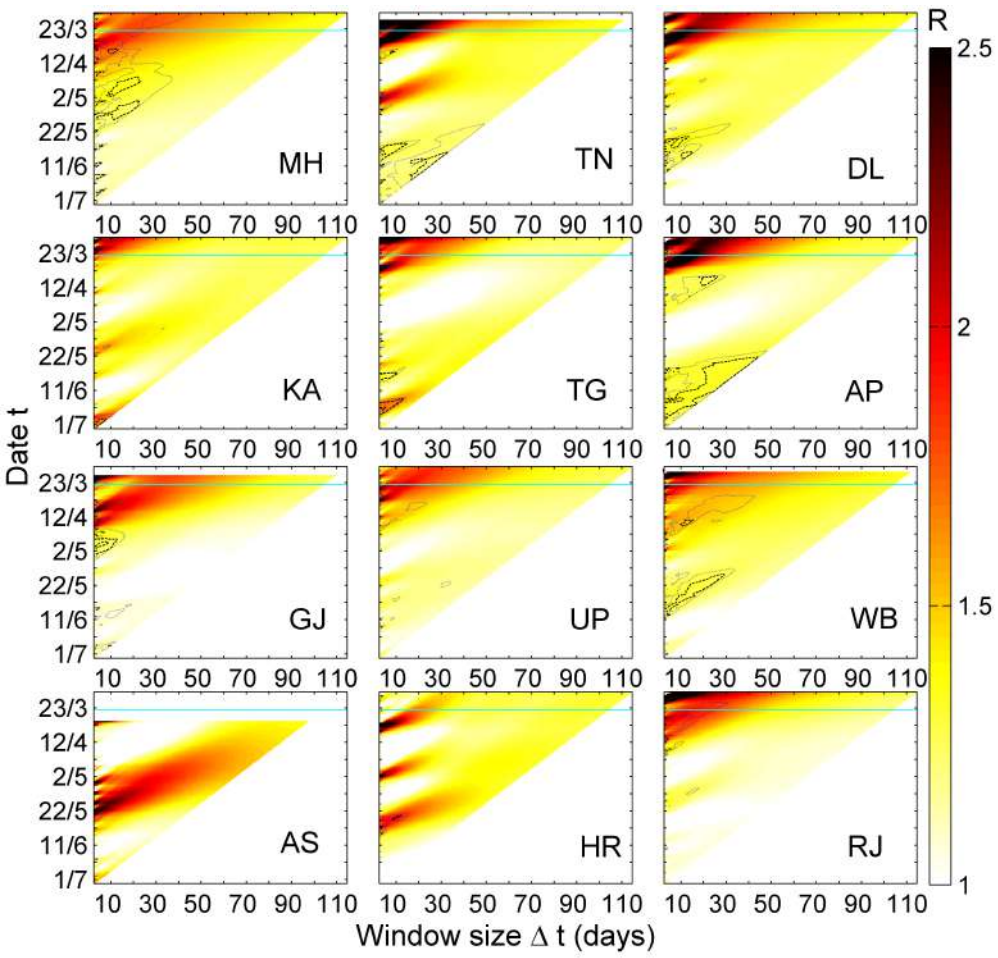


Figure 9: The basic/effective reproduction numbers for the states of India having the highest number of active cases as on 05 July 2020, estimated from the time-series of the number of active cases by using a moving window having different starting dates and interval lengths (Δt).

number of infections across the state to increase rapidly and R rose to 2.01 ± 0.10 between 30 April-07 May 2020. Subsequently, it decreased to 1.56 ± 0.06 between 06-13 May 2020 and then further to 1.31 ± 0.03 during 30 May-04 June 2020.

A perusal of the evolution of the reproduction number for the individual states (see Figure 8 where the horizontal line in each panel indicates the date of imposition of the national lockdown) shows the diversity of outcomes as COVID-19 has spread through India. Not surprisingly, the R value for the state of Maharashtra has largely driven that for the entire country, as it accounts for the largest share of COVID-19 cases among the states, even though it may not have had the highest value of R among them. To see why this is the case, consider a hypothetical situation where a country has an epidemic raging in two of its states with two very different values of R, *e.g.*, 2 in state A and 4 in state B. Consider also that at a given time, state A has 1000 active cases, while state B has 50 cases. After a time period

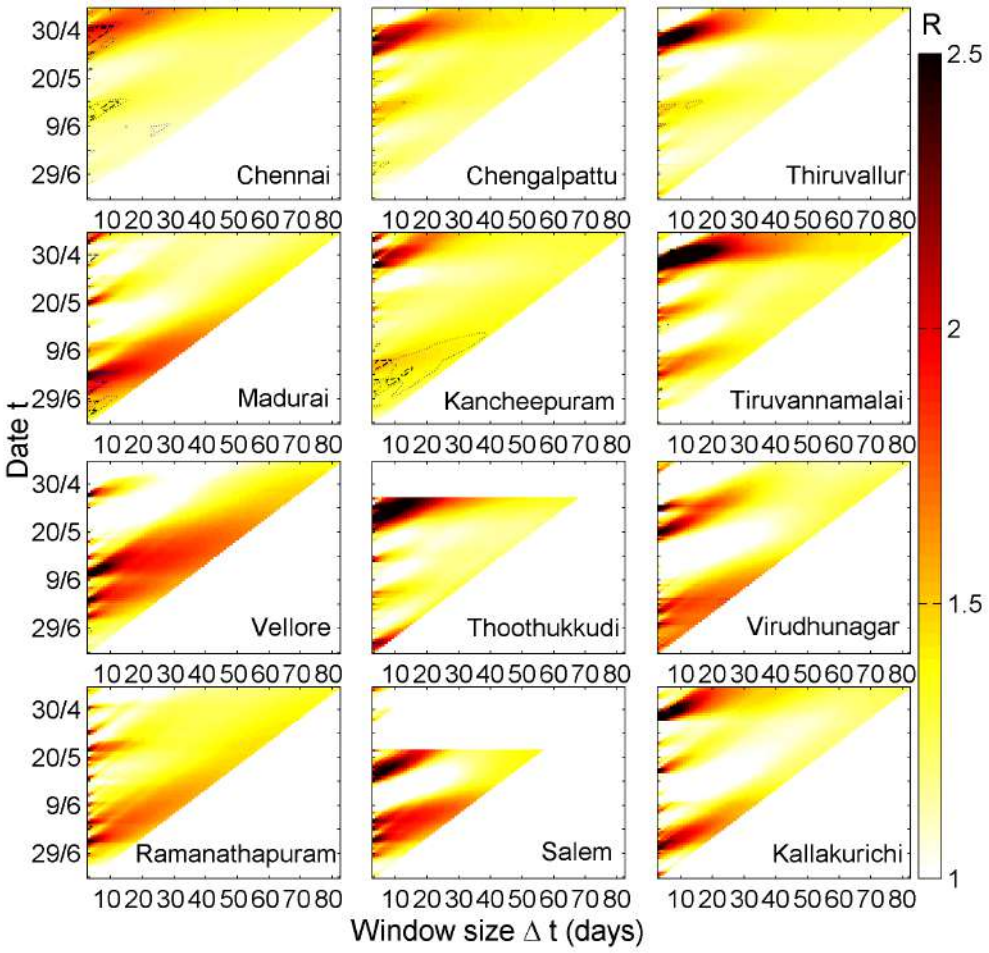


Figure 10: The basic/effective reproduction numbers for the districts of Tamil Nadu having the most confirmed cases as on 15 July 2020, estimated from the time-series of the number of active cases by using a moving window having different starting dates and interval lengths (Δt).

corresponding to one generation interval, state A will have 2000 active cases, while state B will have 200 cases (as per the definition of R). Thus, the effective reproduction number for the country as a whole will be 2.2, a value that is quite close to the R for state A which has the bulk of the active cases even though the epidemic is spreading much slower there than in state B. Note that if the growth rate for the two states remain unchanged, state B will soon surpass state A in terms of active cases and from that point onward will have its R value dominating the national R.

Similar heterogeneity is observed at an even finer spatial scale when we consider the evolution of the epidemic in each district of a state. Figure 10 shows the estimated reproduction numbers for twelve districts in Tamil Nadu that have had the highest number of confirmed cases. It can be easily observed that there has been much more temporal variability in the epidemiological dynamics at this more spatially resolved scale. Most districts show

multiple flare-ups of the epidemic growth rate because of local spreading events, followed by periods during which R has substantially decreased. We also note that the growth is far from being synchronized across a state. For example, while Madurai district had R of 1.41 ± 0.04 during 04-09 July 2020, at around the same time Chennai had the epidemic under effective control so that R was less than 1 (we estimate R as 0.78 ± 0.02 in the slightly later period of 10-14 July 2020). While data for spatial resolution higher than this is not publicly available, it seems reasonable to conclude that the spatio-temporal heterogeneity will be even more pronounced at that scale, not the least because of the relatively stronger daily fluctuations in the number of cases. In contrast, at the level of a state, and more so for a country as large as India, a self-averaging process occurring through rises in some regions being balanced by dips in others, decreases the fluctuations and aids the statistical analysis.

5. Discussion

While the reproduction number is not the only possible metric one can associate with an epidemic, nor does it necessarily contain all relevant information about the epidemic, it is nevertheless extremely informative about the dynamical process by which the disease is spreading (Heesterbeek, 2002). Its value is determined by multiple factors associated with demography and social structure of the population in which the disease is spreading, as well as, the biology of the pathogen, *viz.,* (i) the generation time (which can be considered as the period over which an infected individual passes the pathogen to others), (ii) the mean number of contacts between susceptible and infectious individuals, (iii) the probability of an infection resulting from such a contact, and (iv) the size of the susceptible population (Sinha, 2020). Typically, not all of these factors may be known for an epidemic. Thus, estimating R from empirical data can provide us with a means of making inferences about such factors. The basic reproduction number is also of practical importance from a public health perspective, as using the estimated value of R_0 we can estimate the overall burden of the disease, as well as, in the event of availability of a vaccine for the disease, the fraction of population who will need to be vaccinated to achieve herd immunity. It is for these purposes that it is imperative to accurately estimate R_0. We note in passing that the basic reproduction number for the pandemic is in the same range as the infamous Spanish Flu pandemic of 1918-19 [as estimated by Mills, *et al.* (2004)].

To conclude, it should be stressed that the estimated value of R_0 is at the lower end of the values reported for different regions in which the outbreak of COVID-19 has resulted in a large number of infected cases. Indeed, this seems to be true for most countries from the South Asian region (barring the notable exception of Bangladesh). Preliminary analysis of physical and climatic factors done by us appears to rule out the direct role of these in making R_0 for India low. While reliability of the available epidemiological data may be an issue, it is unlikely that this alone can be the explanation, because under-reporting, as long as it is done consistently at the same level over time, will not significantly alter the estimated value of R. The possibility that genetic or physiological features of the South Asian population may be responsible is a hypothesis that needs further investigation. One of the intriguing questions that arise from the analysis is the fact that the growth rate of the disease has continued to be low despite a large degree of relaxation that has happened in the lockdown norms. As India is still very far from achieving herd immunity, and there appears to be no

evidence that the pathogen has shown any change in its ability to infect, it is possible that the voluntary adherence to public hygiene has been responsible. If true, this may suggest that, under certain conditions, citizens can be mobilized to engage in a mass effort to achieve desirable public health outcomes.

Acknowledgements

I thank Izhar M. Ashraf for technical assistance with acquiring data and Soumya Easwaran for assistance with data analysis in the early phase of the project. The work was supported in part by the Center of Excellence in Complex Systems and Data Science, The Institute of Mathematical Sciences, funded by the Department of Atomic Energy, Government of India.

References

COVID19-India (2020). *A volunteer-driven, crowd-sourced database for COVID-19 stats and patient tracing in India.* https://www.covid19india.org/.

CSSE (2020). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.* GitHub https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J. and Hens, N. (2020). Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv* https://doi.org/10.1101/2020.03.05.20031815

Heesterbeek, J. A. P. (2002). A brief history of R_0 and a recipe for its calculation. *Acta Biotheoretica*, **50(3)**, 189-204.

Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A. and Diaz, G. (2020). First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*, **382(10)**, 929–936.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. and Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, **395(10223)**, 497–506.

India Today (2020). *21-day lockdown in entire India to fight coronavirus, announces PM Narendra Modi.* Last updated March 25, 2020 00:05 IST, Retrieved on 17 July 2020 from https://www.indiatoday.in/india/story/india-lockdown-pm-narendra-modi-speech-coronavirus-1659266-2020-03-24.

MathWorks (2009). *MATLAB R2009b.* The MathWorks Inc., Natick, Massachusetts. https://in.mathworks.com/help/matlab/release-notes-R2009b.html.

Kermack, W.O. and McKendrick, A.G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London,* **A115 (772)**, 700–721.

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., Wesolowski, A., Santillana, M., Zhang, C., Du, X. and Yu, H. (2020). Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak in China. *medRxiv* https://doi.org/10.1101/2020.03.03.20029843.

Jesan, T., Menon, G. I. and Sinha, S. (2011). Epidemiological dynamics of the 2009 influenza A (H1N1) v outbreak in India. *Current Science,* **100(7)**, 1051–1054.

Johnson, N. P. and Mueller, J. (2002). Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the History of Medicine,* **76(1)**, 105–115.

Mills, C. E., Robins, J. M. and Lipsitch, M. (2004). Transmissibility of 1918 pandemic influenza. *Nature*, **432(7019)**, 904–906.

Kelly, H., Peck, H. A., Laurie, K. L., Wu, P., Nishiura, H. and Cowling, B. J. (2011). The age-specific cumulative incidence of infection with pandemic influenza H1N1 2009 was similar in various countries prior to vaccination. *PLoS ONE* **6(8)**, e21828.

Reyes, O., Lee, E. C., Sah, P., Viboud, C., Chandra, S. and Bansal, S. (2018). Spatiotemporal patterns and diffusion of the 1918 influenza pandemic in British India. *American Journal of Epidemiology*, **187(12)**, 2550–2560.

Sinha, S. (2020). Why should we be mindful of nonlinear dynamics in the midst of a global pandemic. *Pramana: Journal of Physics*. To appear.

Spicer, C. C. and Lawrence, C. J. (1984). Epidemic influenza in Greater London. *The Journal of Hygiene,* **93(1)**, 105–112.

Spreeuwenberg, P., Kroneman, M. and Paget, J. (2018) Reassessing the global mortality burden of the 1918 influenza pandemic. *American Journal of Epidemiology,* **187(12)**, 2561–2567.

Wang, C., Horby, P. W., Hayden, F. G. and Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, **395(10223)**, 470–473.

WHO (2020a). *MERS situation update, January 2020.* Retrieved on 17 July 2020 from http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-january-2020.html.

WHO (2020b). *WHO Coronavirus Disease (COVID-19) Dashboard.* Retrieved on 17 July 2020 from https://covid19.who.int/

WHO (2003). *Cumulative Number of Reported Probable Cases of SARS.* Retrieved on 17 July 2020 from https://www.who.int/csr/sars/country/2003_07_04/en/.

WikipediA: The Free Encyclopedia. National responses to the COVID-19 pandemic: Lockdowns. Retrieved on 17 July 2020 from https://en.wikipedia.org/wiki/National_responses_to_the_COVID-19_pandemic#Lockdowns.

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