Ratan Dasgupta *Editor*

Advances in Growth Curve and Structural Equation Modeling

Topics from the Indian Statistical Institute on the 125th Birth Anniversary of PC Mahalanobis

Springer

Advances in Growth Curve and Structural Equation Modeling

Ratan Dasgupta
Editor

Advances in Growth Curve and Structural Equation Modeling

Topics from the Indian Statistical Institute on the 125th Birth Anniversary of PC Mahalanobis

Springer

Editor
Ratan Dasgupta
Theoretical Statistics and Mathematics Unit
Indian Statistical Institute
Kolkata, West Bengal, India

ISBN 978-981-13-1842-9 ISBN 978-981-13-1843-6 (eBook)
https://doi.org/10.1007/978-981-13-1843-6

Library of Congress Control Number: 2018943729

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dedicated to the loving memory of my parents Manoranjan Dasgupta and Saraswati Dasgupta.

Preface

Growth curve model (GCM) is an empirical model for the evolution of a certain characteristic of interest varying over time. These models are of application in different branches of science. The present volume is based on the conference on GCM held in the Indian Statistical Institute, Giridih, during October 23–24, 2017, as a part of the celebration of the 125th birth anniversary of Prof. P. C. Mahalanobis, founder of the Indian Statistical Institute. He was a pioneer of research in statistical theory and methodologies in India and abroad.

The conference proceedings are based on 11 research papers, theoretical and applied, mainly contributed by the participants of the conference. We further invited our colleagues and researchers working on GCM and related topics to contribute to the volume. One could submit more than one paper if interested. All the papers were peer-reviewed, and the revisions culminated in the compilation of research activities in progress for different branches of science currently undertaken in association with scientists from the Indian Statistical Institute. The endeavor will be considered successful and rewarding if the volume can give some idea about solving theoretical and practical problems in the broad area of growth curve model in which many researchers from different branches of science are now interested in.

The present volume is fifth in a series of Springer publications on GCM from selected research works undertaken by the scientists of Indian Statistical Institute. Some of the growth experiments are conducted in the farm of Indian Statistical Institute, Giridih. These have an initial role in development of Growth Curve Models, especially in connection with Elephant foot yam. Some of the associated workers from farm are Aslam Ansari, Sorai Kisku and Hari Sankar Saw. The workers cooperated in conducting successful field experiments. The studies are sponsored by the office of three successive directors of Indian Statistical Institute: Professors Sankar K. Pal, Bimal Roy and Sanghamitra Bandyopadhyay. Ms. Pooja

Sengupta, Purnendu Pradhan, Sourab Bhowmick, Ms. Sucharita Maity helped in data analysis of project linked experiments.

I am thankful to my wife Soma and son Debkumar for their understanding and cooperation when following busy schedules were part of daily activities in last several years, so as to process academic works for publication in a time bound program.

Kolkata, India Ratan Dasgupta
May 2018

Contents

Editor and Contributors

About the Editor

Prof. Ratan Dasgupta is Senior Professor of statistics at the Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata. His research interests include nonparametric statistics, rates of convergence in the central limit theorem, and the application of statistics to industrial quality control, biology, physics, sociology, agriculture, education, environment, and other natural sciences. He has been at the forefront of promoting the theory and applications of growth curve modeling. His knowledge, expertise, and extensive publication record on the topic, as well as his outstanding theoretical skills, make him uniquely qualified to edit this volume.

Contributors

Premananda Bharati Indian Statistical Institute, Kolkata, India

Susmita Bharati Indian Statistical Institute, Kolkata, India

Samarjit Das Indian Statistical Institute, Kolkata, India

Ratan Dasgupta Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata, India

Padmabati Gahan Department of Business Administration, Sambalpur University, Burla, India

Hema Gupta (Joshi) Department of Botany, Visva-Bharati, Santiniketan, West Bengal, India

S. Krithika Department of Clinical and Experimental Epilepsy, Institute of Neurology, University College (UCL), London, UK

Santu Malakar Department of Botany, Visva-Bharati, Santiniketan, West Bengal, India

Nripes K. Mandal Department of Statistics, University of Calcutta, Kolkata, India

Manisha Pal Department of Statistics, University of Calcutta, Kolkata, India

Manoranjan Pal Indian Statistical Institute, Kolkata, India

Monalisha Pattnaik Department of Statistics School of Mathematical Sciences, Sambalpur University, Burla, India

Sedigheh Mirzaei Salehabadi E. K. Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA

Sandip Sarkar Centre for Studies in Social Sciences, Kolkata, India

Soumendu Sen International Institute for Population Sciences, Mumbai, India

Debasis Sengupta Applied Statistical Unit, Indian Statistical Institute, Kolkata, India

Bikas K. Sinha Indian Statistical Institute, Kolkata, India

T. S. Vasulu Biological Anthropology Unit, Indian Statistical Institute, Kolkata, India

Mahalanobis Distance and Longitudinal Growth

Ratan Dasgupta

Abstract Mahalanobis distance is unperturbed by inclusion of highly correlated additional variable, unlike Euclidean distance (Dasgupta 2008). With an application of law of iterated logarithm, an almost sure result is proved in this direction for convergence of sample D^2 statistic to Δ^2, the population Mahalanobis distance squared under mild assumptions, improving earlier results. A rate of convergence $O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1})$ a.s., depending on sample size n and magnitude of correlation $|\rho|$, is proved. The results are of relevance in analysis of longitudinal growth data where some of the variables may be highly correlated. Growth experiments on coconut trees are continued in Sunderban to examine its adaptability in saline soil. Moderate salinity of soil is known to be conducive for coconut tree growth. To examine adaptability of coconut trees in saline soil of Sunderban, longitudinal growth of 42 plants at two time points is observed on six growth characteristics per plant and relevant D^2 statistic is computed to assess tree growth with gradual proximity of plantations towards a saline water river *Bidyadhari*. Principal components of $p = 6$ variables at two time points on the year 2015 and 2017 of $n = 42$ trees are compared to examine the growth status. The angle θ between two growth vectors x and y defined by the relation $\cos\theta = \frac{(x.y)}{||x||.||y||} \in [-1, 1]$ is examined to check direction of growth variability. The values of $\cos\theta$ quantify directional variation present in principal components. These are used to study the growth patterns of coconut trees on a number of aspects, e.g. growth in size, shape. As the variables are scaled in computation of angle, relative variability of the characteristics is examined. When stability of growth is attained, variation over time of each coordinate variable would be negligible after scaling, and the angle between the two growth vectors corresponding to two time points would be small. We compute $n = 42$ angles corresponding to 42 trees, between two growth vectors of $p = 6$ principal components at two time points, and examine whether growth has reached stability. Angle based on principal components may detect minute growth variations, as principal components maximise variance maintaining orthogonality of components to each other. We examine

R. Dasgupta (✉)
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203 B T Road, Kolkata 700108, India
e-mail: rdgupta@isical.ac.in; ratandasgupta@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_1

growth in different orthogonal directions, as the first principal component usually represents the size; the remaining components represent shape and other characteristics. The estimates of angles are independent for 42 plants, and large values of the angles indicate variability of growth direction over time. The angles computed from the growth vectors of original variables are also considered. The angles made by growth vectors of original variables are usually of less variation than those based on principal components. With increase in number of growth variables, variation in growth is seen to be more prominent. In Dasgupta (2017), moderate salinity of soil is found to be conducive for growth of coconut trees. In the present study, in conformity with previous findings, we observe that for some plants the angle between growth vectors increases, when proximity of plants to the river of saline water increases to a moderate level. This is so reflected in variation of angles with respect to the increase in plant's serial numbers from 1 to 30, in 42 plants. Higher serial numbers represent gradual proximity of plants towards saline water river in the group of 42 plants. Clustering of points with large values of angle is observed in the scatter diagrams near serial number 30, where salinity is moderate. The presence of 'whiskers' in a specific region is indicated. The outliers seem to be present near the plant no. 30 with moderate salinity of soil. The plant with largest identification serial number 42 in the riverbank is closest to the river with salinity of water at about 33 g/l, before onset of monsoon.

Keywords Mahalanobis distance · Euclidean distance · Bhattacharya affinity Law of iterated logarithm · Principal components · Proliferation rate

AMS 2000 Classifications Primary 62H99 · Secondary 62P10

1 Introduction and Results on Mahalanobis Distance

Consider $x = (x_1, x_2, ..., x_p)'$, a p-dimensional random variable with mean $\mu = (\mu_1, \mu_2, ..., \mu_p)'$ and a positive definite dispersion matrix $\sum = ((\sigma_{ij}))_{\text{pxp}}$, where $\sigma_{ij} = \text{cov}(x_i, x_j)$, the covariance between ith and jth coordinate variables. Below we recapitulate the set-up considered in Dasgupta (2008). The Mahalanobis distance squared between the random vector x and μ is defined as,

$$\begin{aligned} \Delta^2(x, \mu) &= (x - \mu)' \sum^{-1} (x - \mu) \\ &= (x - \mu)'(x - \mu) = ||x - \mu||^2, \text{ if } \sum = I; \end{aligned} \tag{1.1}$$

i.e. when the variables are uncorrelated.

This distance differs from the usual Euclidian distance $||x - \mu||$; in Δ^2, the relative dispersions and the correlations of the elements of x are taken into account.

For two multivariate populations with means $\mu^{(1)}$ and $\mu^{(2)}$, and *a common dispersion matrix* $\sum$, the Mahalanobis distance squared between the two means or, two populations, is defined as,

$$\Delta^2(\mu^{(1)}, \mu^{(2)}) = (\mu^{(1)} - \mu^{(2)})' \sum^{-1} (\mu^{(1)} - \mu^{(2)}) \tag{1.2}$$

Let C be a matrix such that the dispersion matrix is decomposed in the form $\sum = CC'$ and let $\nu^{(i)} = C^{-1}\mu^{(i)},\ i = 1, 2$. Then,

$$\Delta^2 = (\nu^{(1)} - \nu^{(2)})'(\nu^{(1)} - \nu^{(2)}) = \sum_{i=1}^{p} (\nu_i^{(1)} - \nu_i^{(2)})^2 \tag{1.3}$$

which is the Euclidian distance squared, in terms of transformed coordinates. Such transformations are common to disentangle the coordinates in multivariate analysis.

Let $x_1^{(1)}, \cdots, x_{n_1}^{(1)}$ be a sample of size n_1, from a population with mean vector and dispersion matrix as $(\mu^{(1)}, \sum)$, and $x_1^{(2)}, \cdots, x_{n_2}^{(2)}$ be a sample of size n_2, from the other population with mean vector and dispersion matrix $(\mu^{(2)}, \sum)$, as considered in Dasgupta (2008). Then, the sample estimate of $\mu^{(1)}$ is $\overline{x}^{(1)} = \sum_{i=1}^{n_1} x_i^{(1)}/n_1$ and that of $\mu^{(2)}$ is $\overline{x}^{(2)} = \sum_{i=1}^{n_2} x_i^{(2)}/n_2$, and an unbiased estimate S of the common dispersion matrix $\sum$ is given by,

$$(n_1 + n_2 - 2)S = \sum_{i=1}^{n_1} (x_i^{(1)} - \overline{x}^{(1)})(x_i^{(1)} - \overline{x}^{(1)})' + \sum_{i=1}^{n_2} (x_i^{(2)} - \overline{x}^{(2)})(x_i^{(2)} - \overline{x}^{(2)})'$$

$$\text{i.e. } n\,S = (n_1 - 1)S_1 + (n_2 - 1)S_2, \quad n = n_1 + n_2 - 2 \tag{1.4}$$

An estimate of population distance squared Δ^2 in (1.2) is provided by sample Mahalanobis distance squared,

$$D^2 = (\overline{x}^{(1)} - \overline{x}^{(2)})' S^{-1} (\overline{x}^{(1)} - \overline{x}^{(2)}) \tag{1.5}$$

Mahalanobis distance takes into account the covariance structure of the coordinate random variables $x_i, i = 1, \cdots, p$. If the correlation coefficient between coordinates is near to 1, essentially there is a single coordinate. The effect of such strong dependence on the sample version of Mahalanobis distance is studied in Dasgupta (2008). We shall modify some of the steps in the proof presented therein and obtain improved almost sure rate of convergence for sample D^2 statistic to population Mahalanobis distance squared under milder assumptions.

Without loss of generality, let there be two coordinates and the covariance structure be standardised, i.e. $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Then, $\Sigma^{-1} = (1-\rho^2)^{-1}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$.

Almost sure behaviour of D^2 statistic as $|\rho| \to 1$ is of interest.

Consider the distance between $(\overline{x}_1, \overline{y}_1) = \hat{\mu}^{(1)}$ and $(\overline{x}_2, \overline{y}_2) = \hat{\mu}^{(2)}$, i.e. the D^2 statistic based on a sample of size n each, from two populations $(\mu^{(1)}, \Sigma)$ and $(\mu^{(2)}, \Sigma)$, where $\mu^{(1)} = (\mu_x^{(1)}, \mu_y^{(1)})$, $\mu^{(2)} = (\mu_x^{(2)}, \mu_y^{(2)})$. An estimate of common dispersion matrix Σ is S given by (1.4). Now $S^{-1} \to \Sigma^{-1}$ with probability 1, as $n \to \infty$ by strong law of large numbers. Thus,

$$\begin{aligned} &\lim_{n\to\infty} D^2(\hat{\mu}^{(1)}, \hat{\mu}^{(2)}) \\ = &\lim_{n\to\infty} (\overline{x}_1 - \overline{x}_2, \overline{y}_1 - \overline{y}_2) S^{-1} (\overline{x}_1 - \overline{x}_2, \overline{y}_1 - \overline{y}_2)' \\ &= (1-\rho^2)^{-1} \lim_{n\to\infty} (d_1, d_2) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} (d_1, d_2)' \\ &= (1-\rho^2)^{-1} \lim_{n\to\infty} (d_1^2 - 2\rho d_1 d_2 + d_2^2) \\ &= \lim_{n\to\infty} \{d_1^2 + \frac{(d_2 - \rho d_1)^2}{(1-\rho^2)}\} \end{aligned} \tag{1.6}$$

where $d_1 = \overline{x}_1 - \overline{x}_2,\ d_2 = \overline{y}_1 - \overline{y}_2$.

Let the regression of y on x be linear, i.e.

$Y = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \mu_y + \rho(x - \mu_x)$, since $\sigma_x = \sigma_y = 1$.

Thus, $y = Y + e$ where $e = e_i$ is the error of prediction in the ith sample.

So, $(y - \mu_y) = \rho(x - \mu_x) + e$, $Var(e) = (1-\rho^2)$, $E(e) = 0$.

i.e. $(\overline{y} - \mu_y) = \rho(\overline{x} - \mu_x) + \overline{e}$, $\overline{e} = \sum_i e_i/n \to 0$, with probability 1; as $n \to \infty$.

i.e. $(\overline{y}^{(1)} - \mu_y^{(1)}) = \rho(\overline{x}^{(1)} - \mu_x^{(1)}) + \overline{e}^{(1)}$, for the first sample. Similarly,

$(\overline{y}^{(2)} - \mu_y^{(2)}) = \rho(\overline{x}^{(2)} - \mu^{(2)}) + \overline{e}^{(2)}$, for the second sample.

Thus, $d_2 - \mu_y^{(1)} + \mu_y^{(2)} = \rho(d_1 - \mu_x^{(1)} + \mu_x^{(2)}) + \overline{e}^{(1)} - \overline{e}^{(2)}$, and $d_2 - \rho d_1 \cong (\mu_y^{(1)} - \rho\mu_x^{(1)}) - (\mu_y^{(2)} - \rho\mu_x^{(2)}) + O_e(n^{-1/2}\log\log n) = O_e(n^{-1/2}\log\log n)$ almost surely, from law of iterated logarithm, where the symbol $\cong$ here is interpreted to state that the lim sup of absolute difference of expression in LHS and first term in RHS of the symbol is of the exact order $O_e(n^{-1/2}\log\log n)$ a.s., i.e.

$\limsup_{n\to\infty} |d_2 - \rho d_1 - (\mu_y^{(1)} - \rho\mu_x^{(1)}) - (\mu_y^{(2)} - \rho\mu_x^{(2)})| = O_e(n^{-1/2}\log\log n)$ a.s.

Hence,

$$\begin{aligned} D^2(\hat{\mu}^{(1)}, \hat{\mu}^{(2)}) &\cong (\overline{x}_1 - \overline{x}_2)^2 + O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1}) \\ &= (\mu_x^{(1)} - \mu_x^{(2)})^2 + O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1}) \\ &= \Delta^2(\mu_x^{(1)}, \mu_x^{(2)}) + O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1}) \text{a.s.} \end{aligned} \tag{1.7}$$

In other words,

$\limsup_{n\to\infty} |D^2(\hat{\mu}^{(1)}, \hat{\mu}^{(2)}) - \Delta^2(\mu_x^{(1)}, \mu_x^{(2)})| = O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1})$ a.s.

The RHS of (1.7) is based on a single variable x. So, D^2 statistic is unperturbed when variables with high correlations are added in the set of variables. For limiting D^2 approximation, the error in deleting a correlated coordinate in Δ^2 is $O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1})$, almost surely, under the assumption of linear regression. Thus, D is robust with respect to inclusion of highly correlated variables. Here, n represents the common sample size. In the case of different sample sizes from two populations, n may be considered to be the minimum of two sample sizes.

The results stated below are improved versions of Proposition 1 and Remark 1 stated in Dasgupta (2008). This also extends the result stated in Remark 2 of Dasgupta (2008), proved under stronger assumption of finiteness of fourth moments.

Proposition 1 *The D^2 statistic is unperturbed when an additional variable with high correlation is included to the set of variables. Under the assumption of linear regression and existence of the dispersion matrix $\sum$, the error term in the limiting D^2 approximation by Δ^2, deleting the correlated coordinate from Δ^2, is $O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1})$ a.s.*

Remark 1 Assumption of linear regression is satisfied for normal distribution. When $|\rho| = 1$, then Σ is singular, and S is so almost surely; the Mahalanobis distance is not defined. If the absolute value of correlation between any two variables in the set of variables is high, $|\rho|$ is near 1, and any one of the two correlated variables may be dropped. The growth of $n(\log\log n)^{-2}$ has to be proportional to $(1-\rho^2)^{-1} = O_e((1-|\rho|)^{-1}/2),\ |\rho| \to 1,$ so that error of approximation by dropping the correlated variable in Δ^2 is small almost surely.

In the next section, we shall assess the longitudinal growth status of coconut trees in saline soil of Sunderban at two time points. Apart from asymptotic testing based on D^2 statistic on growth, further comparisons are done based on angles θ made by original growth vectors and by vectors of principal components. From scatter diagram of $\cos\theta$ versus serial number of plants with gradual proximity to the river, it appears that the stability in growth is yet to be achieved; moderate salinity is seen to induce growth variability. Under certain conditions, distribution of $\cos\theta$ is approximated by a Weibull distribution. The distribution fits the observed data well. Discussions of the results are presented in Sect. 3.

2 Growth Status of Coconut Trees

Growth comparisons of coconut trees at two time points here are based on D^2 statistics. The comparison is further enhanced by examining the angle between two growth vectors, while taking different number of growth variables in two occasions. The major classification of coconut based on stature or height is tall palm and dwarf palm, see e.g. www.bioversityinternational.org/fileadmin/bioversity/publications/Web_version/108/ch02.htm.

A longitudinal study of growth in plants is of interest. Growth characteristics of individual plants under study show an increasing trend over time in general, indicating improved growth status.

2.1 Comparison of Coconut Tree Growth Over Two Time Points in Saline Soil and D^2 Statistics

Growth experiment was initiated in the year 1987, to see the adaptability of coconut trees in the saline soil of Sunderban, West Bengal.

Selection of land was made in District Seed Farm, Manmathanagar, near Gosaba, Sunderban. The piece of land, given by Farm on lease to Indian Statistical Institute for coconut cultivation, was by the side of river *Bidyadhari*, flowing near the farm boundary. The experimental plot was in lowland area subjected to water stagnation in rainy season. Coconut trees were planted in several rows parallel to river flow; these are now a little bit tilted towards the edge of the plot at the right-hand side, while facing the river from farm; direction of water flow is towards the right-hand side, merging at sea. Farm land erosion caused by river *Bidyadhari* is continuing over years, thus substantially damaging riverbank tree plantations of palm and coconut.

Growth measurements on 42 coconut plants are serially presented in Tables 1 and 2; these are recorded on two time points on the dates 15 October 2015 and 11 August 2017, via digital photography, assisted by marker of known length. Plates 1 and 2 show the same plant on two occasions. See also Dasgupta (2017) and the references given therein. First four variables are recorded in inch. The sixth variable, a visual grade of the trees on a scale of $\{0, 1, 2, \cdots, 10\}$, represents overall growth status as perceived by the photographers. The detailed procedure is described in Dasgupta (2017). We adopt similar notations in the present paper.

Plants with higher serial numbers are gradually located towards the saline water river *Bidyadhari*, with increasing proximity to the river. The plant with largest identification serial number 42 is the closest to river with salinity of water at 33 g/l.

M.D squared from origin for first set of data is $(\overline{x}^{(1)})' S^{*^{-1}} (\overline{x}^{(1)}) = 50.7664$, and for the second set of data, this is 87.32682. Since the M.D. squared from zero for the second population is higher than that for first population, the growth status of the plants in 2017 seems superior. See Dasgupta (2008) for a quality index based on Mahalanobis distance.

Under the assumption of same dispersion matrix in two populations, the common dispersion matrix may be estimated by the mean of the estimates of dispersions in two populations, $S = \hat{\Sigma} = (\hat{\Sigma_1} + \hat{\Sigma_2})/2$, for the sample sizes are same.

An estimate of population Mahalanobis distance squared Δ^2 above is provided by sample Mahalanobis distance squared,

$$D^2 = (\overline{x}^{(1)} - \overline{x}^{(2)})' S^{-1} (\overline{x}^{(1)} - \overline{x}^{(2)})$$

In the present case, calculated $D^2 = 2.3437$. A test for equality of population mean vectors at two different time points will be made by D^2 statistic later.

Calculation of Bhattacharya affinity, see Bhattacharyya (1946), for two multinormal distribution, leads to an analogue of Mahalanobis distance (Dasgupta 2008). See also Dasgupta (2013).

The extended definition of Δ^2 to the case $\Sigma_1 \neq \Sigma_2$, for *two multivariate normal densities* $\phi_1 = N_p(\mu^{(1)}, \Sigma_1)$ and $\phi_2 = N_p(\mu^{(2)}, \Sigma_2)$, reduces to,

Table 1 Growth data of 42 coconut plants on 15 October 2015

Plant height	Girth base	Girth middle	Girth top	No. of leaves	Visual grade
148.822082	14.03154574	7.191167192	6.138801	23	7
102.2938742	9.435430464	7.019039735	5.062914	10	4
120.4598379	10.75534267	6.555637436	5.838615	9	3
185.8499142	19.55060034	8.344768439	7.987136	19	6
176.3194544	13.67049533	7.48384781	6.486001	19	6
144.7686185	6.242751563	7.111995452	12.88061	15	6
159.9441734	17.10189702	7.985907859	5.876423	23	8
109.3419023	11.4344473	6.878534704	5.359897	12	5
146.2217877	13.51173184	7.454748603	6.056983	21	7
125.8354497	11.84074074	6.692592593	5.515873	15	6
152.2417401	12.78248899	7.807268722	6.046806	16	7
115.9326062	13.86209583	7.093494351	6.335411	10	6
120.2912723	9.208201893	7.088853838	5.627234	14	6
132.3534304	8.573111573	6.646569647	5.586972	17	7
149.0929936	9.207643312	6.994267516	5.666242	21	8
159.9528937	19.73687752	9.634589502	7.950875	22	8
169.6016797	23.76424715	10.50629874	8.838632	18	8
223.4524469	18.86703601	8.599261311	7.187442	15	5
255.3632205	21.87374199	10.936871	8.011894	11	4
141.5605263	10.60789474	7.803508772	5.242982	15	6
149.6491935	10.64919355	7.062096774	6.61371	25	8
219.3491012	19.46263009	10.12582781	7.758751	24	9
179.5303327	17.68101761	7.752446184	4.896282	9	4
207.2668559	20.61816891	8.286728176	3.84741	0	0
191.4384314	25.72862745	9.484705882	6.541176	18	6
174.3558974	21.66974359	9.979487179	7.270769	17	7
143.7243202	23.09667674	11.0234139	6.299094	11	4
216.810582	20.44550265	10.0021164	7.648677	20	8
106.288269	11.53361946	6.860515021	5.965665	15	7
202.4938272	25.47673314	11.61633428	7.920228	19	5
168.2967827	29.47931714	9.035456336	6.753775	18	6
177.6712062	24.42866407	9.194552529	8.022698	15	8
202.967893	23.52307692	10.13444816	7.810033	18	8
145.136036	19.66036036	8.640540541	7.012613	19	7
252.4100418	23.40899582	6.25209205	5.379707	13	7
219.7382979	19.3712766	8.428723404	5.619149	15	7
344.72	22.84108108	9.467027027	7.663784	21	8
267.7084095	22.23491773	8.639853748	6.352834	25	8
212.6939655	22.66738506	8.587643678	6.091236	13	7
284.0339384	24.68429361	8.447513812	6.143646	25	7
183.5746073	13.28141361	7.277486911	6.003927	14	5
144.5329001	13.88870837	5.645816409	4.403737	20	7

First four variables are in inch

Table 2 Growth data of 42 coconut plants on 11 August 2017

Plant height	Girth base	Girth middle	Girth top	No. of leaves	Visual grade
182.8486981	14.97636553	8.222318332	6.656162	21	8
130.0155954	9.096759889	6.930864677	5.0125	6	5
151.8393025	9.932252001	6.952576401	6.207658	10	6
207.8823128	17.63974238	9.845437608	8.614758	17	7
215.6967942	13.01618586	5.820902992	5.820903	12	6
160.2616157	12.18947451	7.593443137	6.194651	13	6
178.6543793	12.84913545	8.017860519	6.784344	16	7
146.4150568	11.87675328	7.788034934	5.451624	8	5
186.921522	14.41681951	7.705541463	7.705541	18	7
183.4416	11.82563774	7.499184906	5.960891	12	5
208.8812512	12.33154374	7.717700847	6.543268	13	5
179.3280815	14.33838987	8.249484581	9.03515	16	6
186.4370965	9.572564856	7.293382748	6.38171	11	5
180.4478975	9.668728344	7.356641131	5.885313	17	6
190.1685252	9.897928058	7.148503597	5.865439	19	7
195.8455031	20.85047182	9.6805762	7.260432	20	8
199.9722079	22.56240735	9.736474541	7.864076	13	5
260.3002333	16.27761111	7.926488889	8.209578	10	4
252.7738583	22.7320935	10.53224409	6.845959	18	7
178.0169483	12.61364699	6.142297663	6.471349	13	5
190.3183009	10.83699653	8.050340278	6.398988	20	7
261.68124	19.17353393	10.28301261	7.283801	19	7
210.8159545	15.59831147	7.958322178	4.45666	2	3
222.2688309	16.89339648	9.412035183	6.757359	14	6
205.5701558	18.66799252	9.333996262	8.667282	17	7
180.169775	22.55721744	10.17474489	7.871029	8	4
292.2336142	22.48807241	9.351475655	7.904223	18	6
253.7158084	19.93794349	9.202127764	7.66844	19	7
135.0320778	10.64337831	5.738169175	5.738169	11	4
240.14036	23.09041924	10.99543773	7.696806	10	4
202.6257772	19.20309653	9.932636139	6.180307	14	4
274.679628	21.73946043	7.418204506	5.769715	17	6
189.9747579	19.80342931	9.671442221	8.980625	16	7
251.6492288	22.79705226	10.57984746	8.060836	17	7
225.8897679	21.04209821	11.146625	9.55425	20	7
334.9526196	24.23179348	7.754173913	8.584978	21	8
251.9083789	21.27749041	8.981453237	7.698388	20	8
438.8771743	30.8715883	10.84677427	8.343673	23	8
274.752614	23.69632168	8.979658741	7.732484	17	7
377.2787193	25.78757473	9.101496962	6.501069	24	8
280.134769	22.8593422	7.763550181	6.253971	18	6
325.7921281	25.86184934	9.572242938	8.396704	17	6

First four variables are in inch

Plate 1. Coconut plant no. 1 on 15 October 2015

$$\Delta^2_{\phi_1,\phi_2} = (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) + 4 \log \frac{|\Sigma|}{(|\Sigma_1||\Sigma_2|)^{1/2}};\ \Sigma = (\Sigma_1 + \Sigma_2)/2.$$

This reduces to usual Mahalanobis distance squared when the dispersion matrices are equal. The above distance squared $\Delta^2_{\phi_1,\phi_2}$ can be estimated from sample by,

$$D^2 = (\overline{x}^{(1)} - \overline{x}^{(2)})' \bar{S}^{-1} (\overline{x}^{(1)} - \overline{x}^{(2)}) + 4 \log \frac{|\bar{S}|}{(|S_1||S_2|)^{1/2}}$$

$$\begin{aligned} \text{where } \bar{S} = \hat{\Sigma} &= (\hat{\Sigma}_1 + \hat{\Sigma}_2)/2, \\ &= (S_1 + S_2)/2. \end{aligned}$$

Plate 2. Coconut plant no. 1 on 11 August 2017

An index of comparison for growth scenarios at two time points can be computed by $(\text{M.D})^2$ from origin, when dispersions are not equal.

$$D_i^2 = (\overline{x}^{(i)})' \bar{S}^{-1} (\overline{x}^{(i)}) + 4 \log \frac{|\bar{S}|}{(|S_1||S_2|)^{1/2}}, \ i = 1, 2$$

From Tables 1 and 2, the computed values are $D_1^2 = 55.7760$ and $D_2^2 = 66.8262$.

The second component in $D_i^2, i = 1, 2$, which is common, is 1.65154; and that can be attributed to variation in two dispersions, which is small in comparison with the total. The pseudo-likelihood ratio test for equality of dispersions in high dimensions as proposed in Bai et al. (2009) performs well even in small or moderate dimensions p. The value of LRT statistic is $21 \times 1.65154 = 34.68234$ to be compared with a

Table 3 Mean vector of the growth variables

Date	Plant height	Girth base	Girth middle	Girth top	No. of leaves	Visual grade
15 October 2015	177.002162	17.425946	8.280335	6.564682	16.642857	6.333333
11 August 2017	223.728814	17.56483	8.581136	7.077882	15.357143	6.119048

chi-square variable with $p(p+1)/2 = 21$ degrees of freedom. The computed value is significant at 5% level, and p value of significance is 0.0306, indicating dispersion matrices are possibly different.

We may then carry out asymptotic test for equality of mean vectors, in the case of unequal dispersion, based on the test statistic

$$D^2 = (\overline{x}^{(1)} - \overline{x}^{(2)})' \bar{S}^{-1} (\overline{x}^{(1)} - \overline{x}^{(2)}) = 2.3437$$

as computed earlier. The sample mean vectors of two populations are given in Table 3.

Now calculated $\chi^2 = 2.3437 \times 42 = 98.44$, which is highly significant with 6 d.f. indicating superior growth status of coconut trees in the year 2017.

Such comparisons can be carried out with first two variables, viz. height of trees and diameter at base. The mean vectors are (177.00216, 17.42595) and (223.72881, 17.56483), respectively, for population 1 and 2.

Mahalanobis distance-squared statistic from origin for first population is 12.51286. And for second population, it is 12.57239.

An asymptotic test for equality of two mean vectors is given by $\chi^2 = (12.57239 - 12.51286) \times 42 = 2.50$ with 2 d.f. p value is $p = 0.2865$, which is not significant.

Increasing the number of variables from 2 to 6 thus seems to provide sharper result in analysis with D^2 statistic.

2.2 Angle Between Principal Component Vectors and Growth Stability

Principal components of $p = 6$ variables at two time points on the year 2015 and 2017 of $n = 42$ trees may be compared to examine stability in growth for plants.

The angle θ between two vectors x and y is defined by the relation $\cos\theta = \frac{(x.y)}{||x||.||y||} \in [-1, 1]$. The angle is nonzero in the presence of growth variability. We examine the angles to assess variation in principal components, in order to study the growth patterns on a number of aspects, e.g. growth in size, shape of the coconut trees. As the variables are scaled in computation of angle, relative variability of characteristics is under study. When stability of growth is attained, variation over time of each coordinate variable would be negligible after scaling, and the angle between the two growth vectors would then be small. We compute $n = 42$ angles between two vectors (x, y) of $p = 6$ principal components at two time points to

examine whether growth has reached stability. Scatter diagram of angles versus plant serial numbers is relevant to assess interdependence of growth variation with salinity. The 42 cosine of angles are as follows.
$\cos\theta$ from principal components for plant number 1–42 sequentially from left to right, row-wise, are as follows.

0.9959022,	0.9995068,	0.9958342,	−0.8865450,	0.2487164,	0.9532908,	0.9373912,
0.9994929,	0.9977222,	0.9981561,	0.9717075,	0.9889522,	0.9949258,	0.9992459,
0.9955459,	0.9919808,	0.8305336,	0.9858285,	0.9783585,	0.9954143,	0.9846454,
0.9919392,	0.4603025,	−0.2504891,	−0.8198713,	0.5934564,	−0.9538746,	0.9988497,
0.9994820,	0.9362831,	0.6495822,	0.1642608,	−0.9428566,	−0.9526317,	0.3158574,
0.9974439,	0.9881725,	0.9969175,	0.9913660,	0.9986049,	0.7257185,	−0.9898743

When growth has nearly stalled in a plant while approaching stability at a mature stage, θ is nearly zero, with $\cos\theta$ near to 1 for that plant.

This helps to examine growth in different directions, as the first principal component usually represents the size, the other components represent shape and other characteristics.

For data on 15 October 2015 on 6 variables, the eigen values are

2720.5350094, 27.7911905, 20.6237735, 2.3091141, 0.9462715, 0.5248376.

And for data of 11 August 2017 on 6 variables, the eigen values are

4078.4498645, 16.5547668, 11.5264570, 0.8972969, 0.5329889, 0.3576735.

The first two eigen values are high in each case, indicating first two principal components explain major variations in data sets.

The estimates of angles are independent over 42 individual plants, and large values of these indicate substantial variability in growth over time.

The angles with original untransformed variables to examine growth at two time points may not provide enough insight of data compared to the analysis based on principal components, as the following analysis explains.

$\cos\theta$ from raw data represented in the above-mentioned fashion are as follows.

0.9991628,	0.9983003,	0.9995149,	0.9995947,	0.9983531,	0.9979777,	0.9979520,
0.9980999,	0.9987273,	0.9978514,	0.9984415,	0.9989607,	0.9975797,	0.9991055,
0.9989513,	0.9991094,	0.9984541,	0.9993050,	0.9995218,	0.9991605,	0.9978894,
0.9990824,	0.9988288,	0.9973372,	0.9989839,	0.9984767,	0.9954318,	0.9996345,
0.9969608,	0.9981453,	0.9961068,	0.9972916,	0.9998835,	0.9966824,	0.9988164,
0.9997185,	0.9995702,	0.9989958,	0.9997428,	0.9995118,	0.9997470,	0.9957904

See Figs. 1 and 2 for histogram of $\cos\theta$ based on 6 principal components, and the 6 original variables, respectively, at two time points. Variation is higher in Fig. 1, and the variable is spanned over $[-1, 1]$.

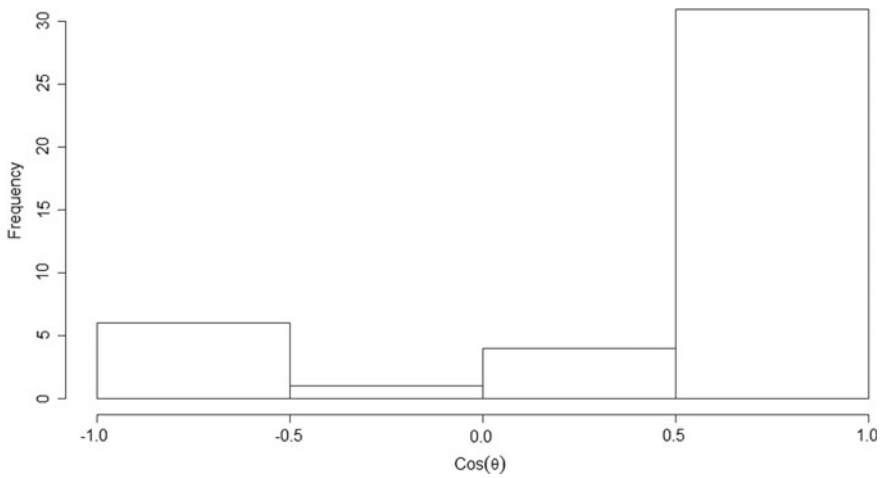

Fig. 1 Histogram of cos(angle) between principal components of six growth characteristics at two time points

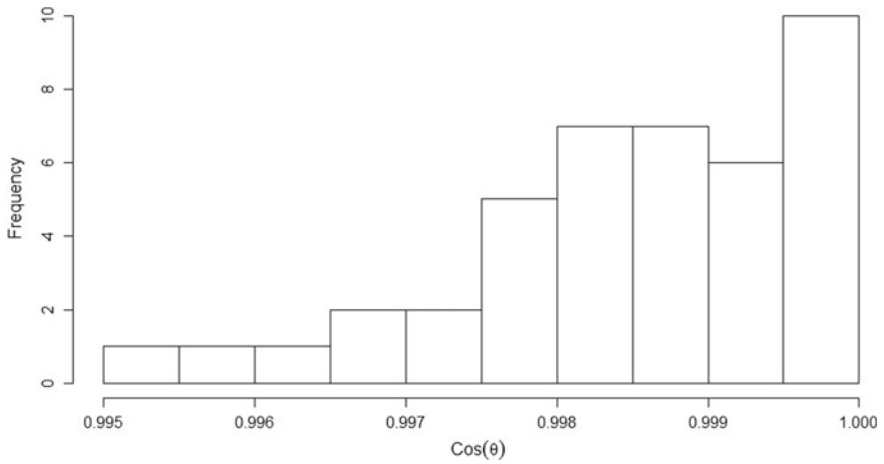

Fig. 2 Histogram of cos(angle) between six growth characteristics at two time points

Variability is maximised with orthogonal components at two time points in principal component analysis. The direction of relative growth is amply reflected through the magnitude of angles.

With variable number increased, variation in growth is expected to be more prominent in computed angle between two time points for original variable vector and for vector of principal component.

Figures 3 and 4 show the similar analysis based only on two main variables, viz. height of the plant and diameter at base of the plants for $\cos\theta$ with principal

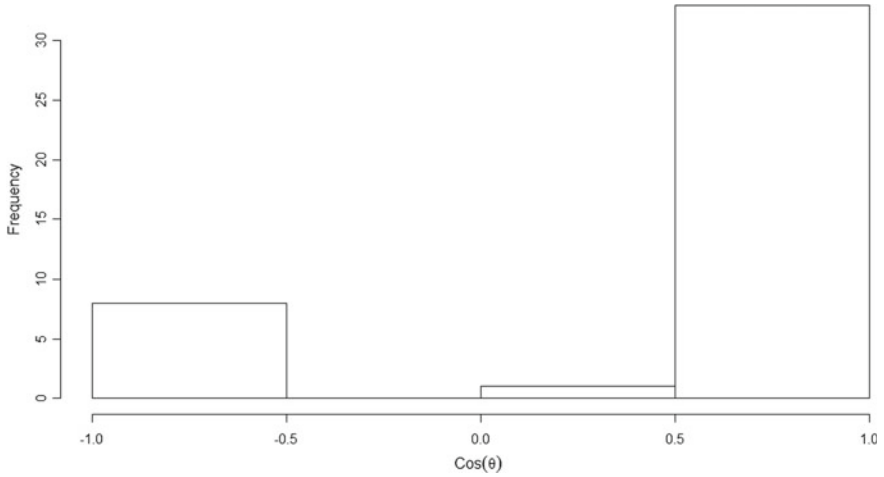

Fig. 3 Histogram of cos(angle) between principal components of two growth characteristics at two time points

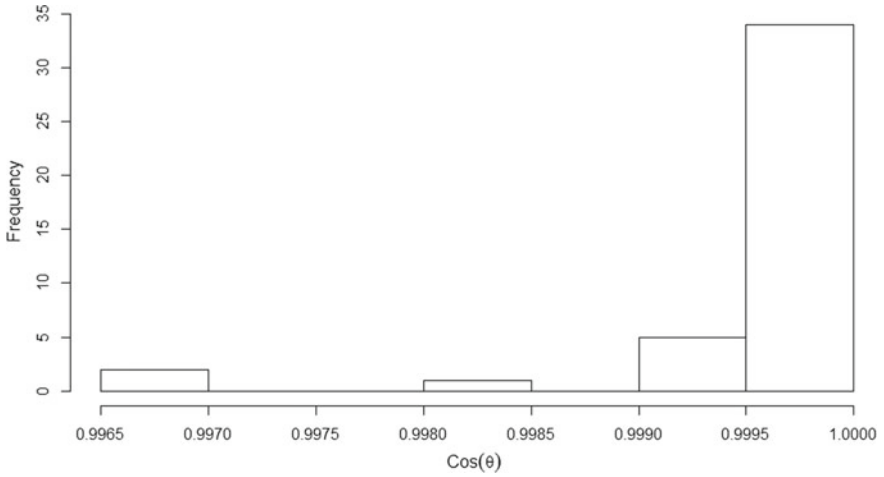

Fig. 4 Histogram of cos(angle) between two growth characteristics at two time points

components and original variable. The patterns shown in Figs. 3 and 4 are similar to those seen in Figs. 1 and 2, respectively, based on all the 6 variables, implying angle computed on less number of variables does not affect the conclusion much for the present data, in comparison with conclusion based on D^2 statistics with less number of variables.

For data on 15 October 2015 of 2 variables, the eigen values are 2718.35190, 20.03124. And for data on 11 August 2017 of 2 variables, the eigen values are 4070.39422, 10.33131.

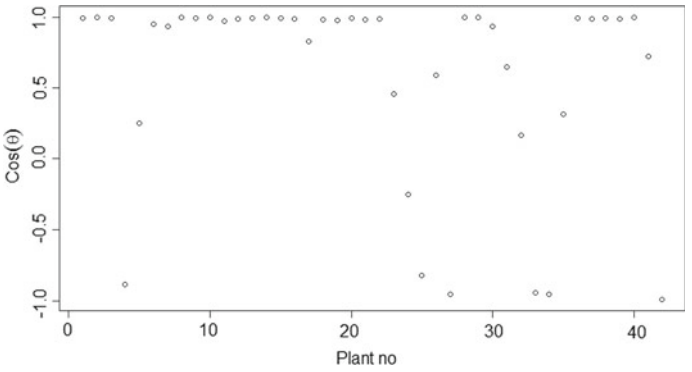

Fig. 5 Variation of cosine(angle) of PC with respect to proximity to river (6 variables)

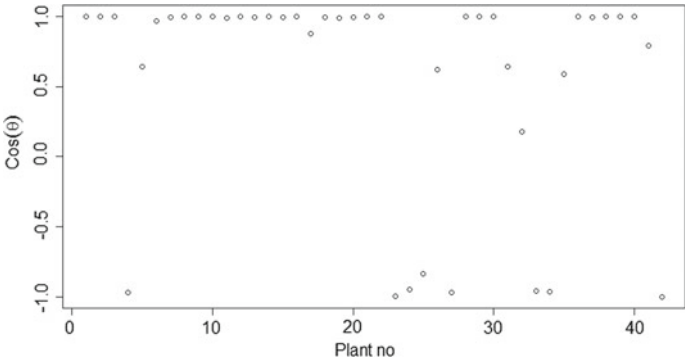

Fig. 6 Variation of cosine(angle) of PC with respect to proximity to river (2 variables)

In an earlier study (Dasgupta 2017), it is seen that moderate salinity of soil is conducive to the growth of coconut tree. In the present analysis, we observe that for some plants the angle between growth vectors increases when proximity to river with saline water increases to a moderate level, as reflected in the increase in relevant plants serial number to 30; see Figs. 5, 6, 7, and 8, where a number of points show clustering towards the base line near plant serial number 30, situated at moderate distance from the saline water river. Plants with higher serial numbers are gradually closer towards the saline water river *Bidyadhari* in a set of 42 plants; the plant with largest identification serial number 42 is closest to the river with salinity of water at 33 g/l.

The points near the line $y = 1$ *in Figs.* 5, 6, 7 and 8 *indicate that growth variation may have stalled as* $\cos\theta \approx 1$. *However, marked downward tendency of points are observed near plant serial number 30, indicating that growth variation is present in the plants where salinity is moderate.*

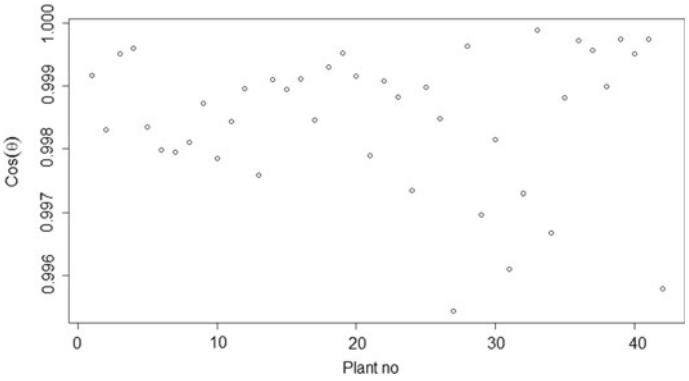

Fig. 7 Variation of cosine(angle) of raw data with respect to proximity to river (6 variables)

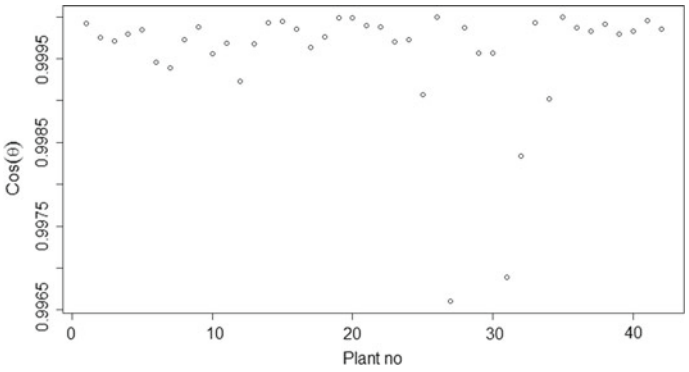

Fig. 8 Variation of cosine(angle) of raw data with respect to proximity to river (2 variables)

Earlier findings that moderate salinity is conducive for coconut tree growth is reconfirmed in the present analysis for Sunderban coconut tree plantations based on cosine of angle between growth vectors.

One has the following series expansion.

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \cdots \tag{2.1}$$

Proliferation rate $(2 + o(1))/\theta$ of the quantity $f(\theta) = 1 - \cos\theta \approx \frac{\theta^2}{2!}(1 + o(1))$ over plant serial number may be studied for similarity check with that computed from observed $(1 - \cos\theta)$, near plant number 30 by derivative estimation technique, in search of assignable causes for variation of points downward, as shown in Figs. 5, 6, 7 and 8. In a subsequent study, we plan further investigation on theoretical proliferation rates along with computed rates from observed data and look for prob-

able causes for deviation observed near plant number 30. Computation of such rates $\frac{d}{d\theta} \log f(\theta)$ from observed data is discussed in Dasgupta (2018).

In Figs. 5, 6, 7 and 8, scatter diagram shows clustering of points with low value of $\cos\theta$ near the base line around serial no. 30 of coconut plant. This is an indication of directional change in growth variability under moderate soil salinity around plant no. 30.

The location of lowest observation (2.5th percentile of 42 observations), i.e. the minimum of $\cos\theta$, indicates the presence of 'whiskers' in a specific region; the outliers seem to be present near the plant no. 30 with moderate salinity of soil. To infer about outliers, one may consider data points on or below 2.5 percentile.

Distributional characteristic of the computed angle is also of interest. Angle between growth vectors does not depend on the magnitude of the vectors. Consider the two-dimensional case. Angle θ between the growth vectors joining origin with the points, $x = (x_1, x_2)$ and $y = (y_1, y_2) = (0, 1)$ say, may be written as $\cos\theta = \frac{(x.y)}{||x||.||y||} = \frac{x_2}{(x_1^2+x_2^2)^{1/2}} = x_2$; for $y = (0, 1)$ and x lying on the unit circle, $\cos\theta$ is an increasing function of x_2 from 0 to 1. In three-dimensional unit sphere, a similar relation holds, the angle of vectors joining the points, $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3) = (0, 0, 1)$ say, $\cos\theta = x_3$. Such a reformulation of coordinates is possible by a change of base points in measurements, keeping the magnitude of angle unchanged.

Thus, an increase in a coordinate here may result in an increase in the magnitude of $\cos\theta$. The coordinate vectors of growth characteristic may fluctuate over time, and the maximum may affect the angle computed for growth at second time point. Therefore, it may be a good idea to attempt an extreme value fit to $(1 - \cos\theta)$. Weibull fit by Minitab is shown in Figs. 9 and 10. A satisfactory Weibull fit for $(1 - \cos\theta)$ based on six growth variables is shown in Fig. 10. The fit seems marginal in Fig. 9

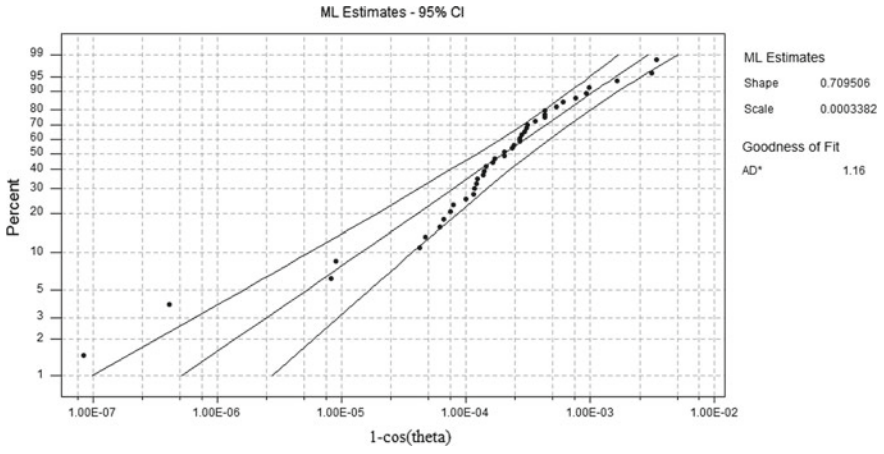

Fig. 9 Weibull plot for $(1 - \cos\theta)$ with 2 variables from raw data

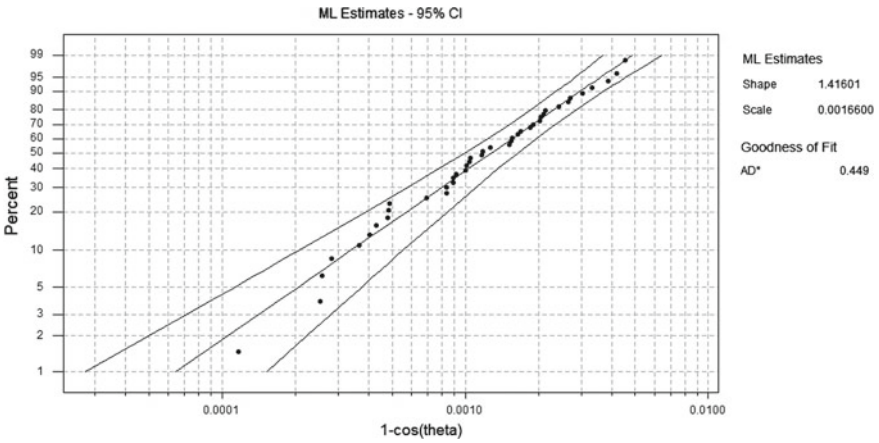

Fig. 10 Weibull plot for $(1 - \cos\theta)$ with 6 variables from raw data

with two points falling outside 95% confidence band, and the fit is better in Fig. 10 with value of the Anderson Darling statistic to be 0.449.

The six growth characteristics of plants, viz. plant height, plant girth at base, girth at middle of the tree, girth at top of tree (from where the leaves started expansion), number of leaves, and a visual grade on growth assigned to the tree by the photographer, are serially assigned to the 42 plants in Table 1. First four variables shown in tables are measured in inch.

The six growth characteristics of serially assigned 42 plants on 11 August 2017 are shown in Table 2. Except for the last variable recorded by another photographer, the variables on later date are in increasing order over plants in general, compared to earlier date of taking measurement.

Mean vector of growth variables on two occasions is shown in Table 3. Except for last two variables, the mean of other four variables is of higher magnitude in the later date.

In the histogram of $\cos\theta$ between principal components, the variable at base has spread on $[-1, 1]$. A large proportion of variable is within $[0.5, 1]$. The distribution is negatively skew.

For angles based on original variables, distribution of $\cos\theta$ is within a narrow range of $[0.995, 1]$ compared to that in Fig. 1. The distribution is negatively skew.

When the number of growth characteristic is two, viz. height and diameter of tree at base, the pattern of histogram for $\cos\theta$ with principal components does not change much compared to full set of six variables.

The spread of the variable is within a small range for histogram of $\cos\theta$ between two growth characteristics, and pattern of histogram remains the same as that for six growth characteristics.

Although most of the points are near 1, still a cluster of points with high variation in $\cos\theta$ are seen around serial number 30, referring to moderate salinity of soil.

The picture pattern remains the same as that of Fig. 5, even with less number of variables, indicating analysis with principal components is affective.

The scatter is now more evenly spread in a low range of variation in $\cos\theta$. Still a cluster of points are seen around serial number 30.

With number of variables reduced to 2, a cluster of points are still seen around serial number 30, and the points have variation in $\cos\theta$ values, indicating growth has not stalled in moderate salinity of soil.

Weibull plot of $(1 - \cos\theta)$ where θ is the angle between two growth vectors at two time points with 2 growth variables, viz. height and girth at base, shows a marginal fit; the value of A.D. statistic is 1.16. Two points fall outside the 95% confidence band.

In Fig. 9, we consider all the six growth characteristics computed at two time points to find the angle θ between the two growth vectors. Weibull fit is satisfactory, with value of A.D. statistic as 0.449. All the points lie within the 95% confidence band.

3 Discussions

Mahalanobis distance remains unperturbed with respect to inclusion of highly correlated additional variables. Such highly correlated variables can be dropped from the calculation with nominal error committed. An almost sure rate of convergence in sample D^2 statistic to Δ^2, the population Mahalanobis distance squared is computed with an error bound $O_e(n^{-1}(\log\log n)^2(1-\rho^2)^{-1})$ a.s., under minimal assumptions like existence of dispersion matrix and validity of linear regression, when the correlated coordinate is dropped from the calculation of D^2 used to estimate Δ^2, the population Mahalanobis distance squared.

Based on six growth characteristics, we study the longitudinal growth scenario at two time points by D^2 statistic for coconut trees planted in the saline soil of Sunderban by side of the river *Bidyadhari*. Stability of tree growth in plantation seems yet to be achieved. A second technique based on the angle θ made by growth vectors of original variables at two time points, as well as angle made by principal components of the growth characteristics, is used to check for stability in growth over time. Scatter diagrams of $\cos\theta$ indicate that moderate salinity of soil is conducive to coconut tree growth variation.

The number of growth variables considered affects the conclusion to some extent in the analysis presented. Even with less number of variables per plant, the conclusion that growth is present in moderate salinity does not change much when angle of growth vectors is considered. This is true especially for angles made by vectors of principal components.

References

Bai, Z., Jiang, D., Yao, J., & Zheng, S. (2009). Corrections to LRT on large dimensional covariance matrix by RMT. http://arxiv.org/pdf/0902.0552.pdf

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. Sankhya, A.7. 401–406.

Dasgupta, R. (2008). Quality index and Mahalanobis D^2 statistics. In *Proceedings of ISI Platinum Jubilee Conference World Scientific Advances in Multivariate Statistical Methods* (pp. 367–382)

Dasgupta, R. (2013). Optimal-time harvest of elephant foot yam and related theoretical issues. Advances in growth curve models: Topics from the Indian Statistical Institute. Springer Proceedings in Mathematics & Statistics, 46. Chapter, 6, 101–130.

Dasgupta, R. (2017). Coconut plant growth, Mahalanobis distance, and Jeffreys' prior. In *Growth curve models and applications* (Chap. 5, pp. 115–126). Springer.

Dasgupta, R. (2018). Longitudinal studies on mathematical aptitude and intelligence quotient of North Eastern tribes in Tripura. In *Advances in Growth Curve and Structural Equation Modeling: Proceedings 2017* (Chap. 1). Springer. (To appear).

www.bioversityinternational.org/fileadmin/bioversity/publications/Web_version/108/ch02.htm

Stock Market Growth Link in Asian Emerging Countries: Evidence from Granger Causality and Co-integration Tests

Monalisha Pattnaik and Padmabati Gahan

Abstract For both risk management and foreign equity portfolio investment purposes, financial integration of Indian stock market is crucial especially among seven dominant players in the Asian emerging stock markets. The purpose of this paper is to examine the evidence on the integration of Indian stock market with other Asian emerging stock market prices. Daily time series data spanning the period from December 2000 to March 2016 has been used. The unit root test, the co-integration test and the Granger causality test for testing cause–effect relationship of India with the set of seven country stock price indices, including that of Hong Kong, Indonesia, Malaysia, South Korea, Philippines, China and Taiwan have been applied to derive the long-run and short-term equilibrium relationships. The findings of the study establish that there is not any evidence of co-integration between India and the other countries like Malaysia, Philippines and China in the sample; however, there is strong evidence of co-integration between India and these countries like Hong Kong, Indonesia, South Korea and Taiwan. It is shown that the stock markets of India with Hong Kong, Indonesia, South Korea and Taiwan do have a long-run relationship. Through multivariate co-integration test, it is concluded that Philippines stock market influences both the Indian stock market and Hong Kong stock market positively and significantly where Taiwan stock market influences both the Indonesian stock market and Malaysian stock market positively and significantly. Through Granger causality test, it is clear that the Indian capital market is not getting significantly affected by all of the emerging market economies. The study emphasises on the evidence of long-run equilibrium relationships among emerging Asian economies.

Keywords Stock markets · Co-integration · Emerging Asia

M. Pattnaik (✉)
Department of Statistics School of Mathematical Sciences, Sambalpur University, Jyoti Vihar, Burla 768019, India
e-mail: monalisha_1977@yahoo.com

P. Gahan
Department of Business Administration, Sambalpur University, Jyoti Vihar, Burla 768019, India
e-mail: pgahan7@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_2

1 Introduction

Financial integration makes significant positive effect on GDP growth in an economy and also ensures increased capital flows from industrialized countries towards developing as well as emerging countries. As financial market integration proceeds, geography becomes less relevant to finance and thereby the less fortunate countries get a source to fund their economic growth. A liberal and global trading as well as financial system is essential for the developing countries to achieve greater integration with the world economy and thereby ensure more rapid economic growth. In the post-globalization era of India, given that its re-emergence and grand strategy as a rising power have been predicated on economic reintegration into the global economy; the co-integration study of Indian economy with the world economy has bound to have strategic policy implications.

2 Literature Review

Since the stock market crash of October 1987, research on stock market integration has been well studied. Initially most of the studies are focussed on developed economies but after the post-Asian crisis, the literature has started focussing on emerging Asian markets. In recent past, few studies are made on the co-integration of Indian markets.

Mittoo (1992) examined the integration of Canadian and US stock markets in a period that is relatively free from capital controls. The analysis is done over a 10-year period 1977–86. Instrument used Canadian TSE 35 index stocks and US stocks matched on the basis of size and industry. Capital asset-pricing model and multi-factor pricing model (APT framework) are employed in the study. Yuhn (1997) tried to explore if there is any link between financial integration and market efficiency by developing a dynamic representation of co-integration that is consistent with the efficient markets hypothesis in five most industrialized and advanced markets. In this study, it was first tested by Philips (1987), Philips and Perron (1988), and Perron (1988) test if individual national stock price and dividend series contain units roots. The empirical analysis being performed under the study indicated that the market USA and Canada are efficient, but the stock markets of Japan, UK, and Germany are not efficient which as per the authors is because of the stringent capital controls levied until the late 1970s.

Bollerslev and Jubinski (1999) studied from July 2, 1962 to December 29, 1995. The data consists of the bivariate absolute return and trading volume series of each of the 100 firms included in the January 10, 1997, revision of the S&P 100 broad-based composite index. This study examines the behaviour of equity trading volume and volatility for the individual firms composing the Standard & Poor 100 composite

index. Multivariate spectral methods and mixture of distributions (MDH) are used. Tabak et al. (2002) provided further evidence on the linkage between daily close quotes for stock prices of Latin American Equity markets (Argentina, Brazil, Chile, Colombia, Peru, Mexico and Venezuela) and the US equity market from January 1995 and till March 2001. For assessing whether the indices have unit roots, the most widely accepted test, i.e. Augmented Dickey and Fuller (1979) test, was performed and then the Johansen's method for test of co-integration and Granger's approach for the test of causality were adopted. The results shown that shocks in the US stock market have a heterogeneous effect of Latin American stock markets and there is relatively a greater degree of integration between USA and Mexico than others.

Wong et al. (2004) studied the co-movement of stock markets in major developed countries and Asian emerging markets and also investigated the level of declining in benefits of international diversifications. It is derived that that stock markets of Singapore and Taiwan are co-integrated with that of Japan while Hong Kong is co-integrating with the USA and the UK. Further, no evidence of any co-integration between Malaysia, Thailand and Korea and the developed markets of the USA, and the UK and Japan were detected.

Tahai et al. (2004) investigated financial co-integration of equity markets. For the test of stationarity, the augmented Dickey–Fuller test was implemented and then a vector error correction model for I(2) processes was introduced that allows for linear deterministic trends in the I(1). For this purpose, the authors employed monthly stock indices from March 1978 till December 1997. As explained by the authors, it could be because of the increased globalization that has brought about a shift from traditional joint ventures to strategic alliances set up between global competitors within these countries. However, the authors also accept that the degrees of global integration as measured in this study may not be truly indicative of a firm's international involvement.

Bose and Mukherjee (2006) studied the inter-linkage between the Indian Stock Market and some other emerging and developed markets like USA and Japan and seven Asian markets, namely Hong Kong, Korea, Malaysia, Singapore, Taiwan, Thailand along with India. Period of study was daily data for the period January 1999 to June 2004. Market integration may not be only due to free mobility of capital, but may to a great extent depend on institutional factors. Methodology used is pair-wise and group-wise co-integration and Granger causality tests. The nature of co-movement or integration with emerging Asian markets does not yet warranty any immediate concern regarding possible contagion in case of any financial crisis in the region. The Indian Market is seen to belong to the group of Asian markets co-integrated with them and with the US market.

Wong et al. (2005) empirically investigated if there is any long-run equilibrium relationship and also short-run dynamic linkage between the Indian stock market and three major developed markets considering weekly indices of these stock exchanges

from January 1991 till December 2003 as the sample. The authors justified considering the weekly indices instead of daily data by saying that it avoids representation bias. In addition to this, the authors used Wednesday indices to avoid day-of-the-week effect of stock returns. The indices were adjusted to be in terms of US dollars for better comparison. As the Indian market is found to be co-integrated with other markets, the ECM model was applied to test the Granger causality. Further, the authors applied the multivariate co-integrated system and at last, in order to have a characterization of the long-run dynamics of the system of the stock indices in the study a generalized form of co-integration, known as fractional co-integration was applied. The results of the study revealed that the Indian stock market is co-integrated with stock market in USA, UK and Japan and the unidirectional Granger causality that is running from the USA, UK and Japanese stock markets to the Indian stock market. The authors claim that their findings on co-integration and causality can enable investors to take suitable decisions regarding investments in Indian stock market.

Adjasi et al. (2006) investigated the existence of long-run linkages amongst African stock markets, and also tested if there is short-run dynamic inference that may be present. In this study, seven countries were selected from the African continent on the basis of the age of their stock markets, availability reliable and consistent data and must be with value weighted constructed indices. A dynamic vector autoregressive regression (VAR) that explores both co-integration and Granger causality possibilities was adopted in this study. The findings from the empirical analysis reveal that there is a set of long-run relationships which is unique in nature underlying African stock markets and this long-run relationship hinges two categories of markets: a larger more relatively active market, i.e. South African stock market, and a smaller and inactive market, i.e. Ghana stock market. Hence, it may be concluded that there is long-run interdependence between stock market returns the long-run dynamics.

Gikas et al. (2006) studied the integration of eleven Euro-zone countries and the UK and they estimated a conditional asset-pricing model with a time-varying degree of integration. Jeyanthi et al. (2008) focused on the emerging markets justifying that due high degree of globalization of finance of recent past, there is a significant increment in funds flowing from the developed markets towards the developing markets, and therefore for an efficient management of portfolio, these markets are becoming important. By employing Dickey–Fuller Test and further Augmented Dickey–Fuller Test, the unit roots in stock prices were found. Then, pair-wise co-integration tests implementing co-integrating Regression Durbin–Watson (CRDW) test indicated that there is no evidence of co-integration among the stock prices. Most of the developed markets except Singapore and Hong Kong were found to be earning negative returns during the period of the study while all the emerging market economies except Taiwan were found to be earning positive returns during the period. Since the returns from the emerging markets were found to be reasonably better than the matured American and European markets, the authors concluded that international investors can achieve substantial risk diversification benefits in these markets.

Batori (2009) choose selected diverse markets which are with different levels of development and efficiency. In an aggregate, 15 European equity markets were over the period January 1999 to December 2008 were studied in this paper on the basis of daily data. At the first step of analysis, the data was put for a test of unit root and further the causal relationship among the data series were studied using tests of co-integration, Granger causality, VAR, impulse response functions and variance decomposition procedures. Apart from a numerous evidences of causal and short-term linkages, the results of the study also support a long-term equilibrium relationship among the stock markets.

Bangake and Eggoh (2010) used the approach of assessing relationships between financial development and economic growth with time series and this has been done by analyzing the long-run relationship between 25 OECD countries. First of all the researchers employed the recently developed panel methods of testing the unit roots and then Fully Modified OLS and Dynamic OLS for the test of co-integration since these methods avoid problems of low power associated with the traditional unit root and co-integration tests. Another contribution of this study is that it has taken into account two financial development indicators: the banking sector indicators and the stock markets indices. The results of the study clearly provide evidences for the existences of a single long-run equilibrium between financial development, economic growth and the set of control variables and also points to a bidirectional causality between finance and growth.

Ivanov (2011) studied the co-integrating relation between the S&P 100 and S&P 500 indices in extreme market conditions. By using high frequency, one minute interval data the study examined the influence of multiple crises on the relation of indexes like Black Monday of October 19, 1987, the Friday the 13th mini-crash of October 13, 1989, the 1997 mini-crash of October 27, 1997, the Flash Crash of May 6, 2010 and the Japanese Earthquake of March 11, 2011. The major findings of the study are that there was a linkage breakdown between these two indices the day after the Japanese earthquake, but no linkage-breakdown on or around the other crises. The results of the study also indicate that each crisis has affected differently the co-integrating relation between the indices in consideration.

In the present studies, an attempt is made to study co-integration of India with selected emerging Asian countries. Researches including India and other developing economies have been conducted more in comparison with other countries. The present study has been conducted considering the limitations of the previous studies in various countries.

The rest of the paper is organized as follows. Section 3 defines the objectives and structure of the study, and Sect. 4 introduces the methodologies and data uses in the study. Section 5 discusses empirical estimation and results of the study. Section 6 concludes.

3 Objectives and Structure of the Study

This study aims at identifying the long-run equilibrium relationship of India with seven selected emerging Asian stock exchanges. This study has been undertaken with the following broad objectives:

1. To conduct the Granger causality test of India with the selected emerging Asian economies.
2. To detect whether India is financially integrated with the selected emerging Asian economies, i.e. Hong Kong, Indonesia, Malaysia, South Korea, Philippines, China and Taiwan through the analysis of co-integration.

3.1 Hypotheses of the Study

The following null (H0) and alternative (H1) are framed (see Eq. (1));

1. H0: $\delta = 0$ (Level data series is non-stationary)

   H1: $\delta \neq 0$ (Level data series is stationary)

2. H0: No causality of Indian stock market index with selected stock market indices

   H1: There is causality of Indian stock market index with selected stock market indices.

3. H0: No co-integration of Indian stock market index with selected stock market indices.

   H1: There is co-integration Indian stock market index with selected stock market indices.

4 Methodology

The daily closing level data of the sample indices has been taken from the online database maintained by Web resources such as www.yahoo.com/finance for the period 05.12.2000–09.03.2016. The total number of observations for each case is 3696.

Daily level data of the indices has been converted to daily returns by taking natural logarithm. The present study uses logarithmic difference of the daily level data for two successive periods for the calculation of daily rate of return of the indices.

The logarithmic difference is symmetric between up and down movements and is expressed in percentage. The one day return on the index is calculated as:

$$y_t = \ln(I_t / I_{t-1}) \times 100$$

where

ln (Z) Natural logarithm of 'Z'.
I_t Daily closing level of the index for the date 't'.
I_{t-1} Daily closing level of the index for the previous date t − 1.

4.1 Techniques of Data Analysis

The present study uses the daily return series of the sample indices for the determination of their stationarity and co-integration studies. From the return data, descriptive statistics like mean, standard deviation, skewness, and kurtosis are calculated and then the return series is undergone a test of normality test by using a test called Jarque–Bera (JB) test statistics. The formulas for the above statistics are given below.

4.2 Formula of Descriptive Statistics and JB Test Statistics

1. Daily mean return $(\bar{y}) = \sum_{t=1}^{n} y_t / n$
2. Standard Deviation in return $(\delta) = \left[\sum_{t=1}^{n} (y_t - \bar{y})^2 / n - 1\right]^{1/2}$
3. Skewness in return $(S) = \sum_{t=1}^{n} (y_t - \bar{y})^3 / (n-1)\delta^3$
4. Kurtosis in return $(K) = \sum_{t=1}^{n} (y_t - \bar{y})^4 / (n-1)\delta^4$
5. JB statistics, $JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right)$

where n = no. of observations in return series, S = skewness coefficient and K = kurtosis coefficient

JB test statistic is asymptotically distributed as a chi-squared random variable with two degrees of freedom [$\chi^2(2)$]. It is used under the null hypothesis of normal distribution in the data. If the P-value of the calculated JB statistic is less than the chosen significant level, then the data series is normally distributed.

From the JB test, if it is detected that return distributions of the sample indices over the period of study are not normally distributed, then there is presence of skewness and excess kurtosis which can be modelled for the calculation of the conditional volatility of the return series.

4.3 Test of Stationarity

In order to arrive at meaningful inferences, it is necessary to conduct studies on the stationary data series. To test whether the return series is stationary or non-stationary (with the presence of unit root in the data), the Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests are applied to the data series.

4.4 Augmented Dickey–Fuller (ADF) Test

Stationarity of the return series of the sample indices are tested by conducting ADF test which takes care of the possible serial correlation in the error terms by adding the lagged difference terms of the regressors and thereby overcomes the drawback of the Dickey–Fuller (DF) test based on the assumption of independently and identically distributed error terms. The ADF test is conducted by using the following regression equation:

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \in_t \tag{1}$$

where

Y_t	Return data series over the study period
ΔY_t	First difference, i.e. $\Delta Y_t = Y_t - Y_{t-1}$
δ	Coefficient of the autoregressive (AR1) term of Y_t
α_i	Coefficient of ΔY_{t-i}
β_1	Drift parameter
β_2	Coefficient of the trend variable 't'
t	Time or the trend variable
ΔY_{t-i}	$Y_t - Y_{t-i}$
$\in_t$	A pure white noise error term

In fact, $\delta = \rho - 1$, when $\delta = 0$, then $\rho - 1 = 0, \therefore \ \rho = 1$. This '$\rho$' is the coefficient of y_{t-1} when y_t is the regress and y_{t-1} is the regressor. Hence, the non-stationary series is called the unit root series. In order to test the null hypothesis of 'unit root' in the return series, a "tau (τ) statistics" is used. If the computed value of the tau (τ) statistics is greater than the critical value of the tau (τ) statistics, the null hypothesis of "unit root" in the data will be rejected. The formula for the computation of tau (τ) statistic is stated below:

Computed value of tau (τ) $= \frac{\delta}{\sqrt{\frac{1}{n}}}$, where δ = Coefficient of y_{t-1}, n = no. of observations and $\sqrt{\frac{1}{n}}$ = Standard error of estimate. The critical value of tau (τ)

statistics is computed by Dickey–Fuller on the basis of Monte Carlo simulation (Dickey–Fuller, 1979) which is known as Mackinnon critical value of tau (τ).

4.5 Phillips–Perron Test (PP)

The Phillips–Perron (PP) test is a nonparametric test that does not assume the presence of serial correlations in the error terms like ADF test. This method is similar to the non-augmented DF test. The t-ratio of the coefficient of Y_{t-1} i.e. δ is modified in PP test which is shown below:

$$\text{Modified } t_s = t_s\left(\frac{\gamma_0}{f_0}\right)^{1/2} - \frac{n(f_0-\gamma_0)(se(\delta))}{2f_0^{1/2}s} \tag{2}$$

δ	Coefficient of Y_{t-1}
t_s	T-value of the coefficient of Y_{t-1}, i.e. β_3 of the ADF equation
se (δ)	Standard error of the coefficient s
s	Standard error of the test regression
γ_0	Error variance
f_0	An estimator of the residual spectrum at frequency zero
n	No. of observations

If the absolute computed t_s-value is greater than the Mackinnon Critical tau value, the null hypothesis that the series is non-stationary will be rejected in PP test. Therefore, the alternative hypothesis that the series is stationary will be accepted.

4.6 Test of Co-integration: Engle–Granger Test

In the Engle–Granger method, if the variables are integrated of the same order, the equation is estimated.

$$y_{t1} = \beta_0 + \sum_{j=1}^{n} \beta_j y_{tj} + \varepsilon_t \tag{3}$$

The presence of unit root in the error terms from this regression is tested by using ADF. If the error terms ε_t of the regression are stationary I(0), which means that the variables are co-integrated and that even if the variables in the system are not stationary, the system consisting of these variables has a long-term equilibrium point.

Table 1 Description of data

Sl. no.	Country	Type	Name of the index	Period of study	Total number of observations
1	India	Emerging	Bombay Stock Exchange (BSE)	05/12/2000 till 09/03/2016	3803
2	Hong Kong	Emerging	Hang Seng (HSI)	05/12/2000 till 09/03/2016	3788
3	Indonesia	Emerging	Jakarta Composite (JCI)	05/12/2000 till 09/03/2016	3729
4	Malaysia	Emerging	KLSE	05/12/2000 till 09/03/2016	3763
5	South Korea	Emerging	Seoul Composite (KOSPI)	05/12/2000 till 09/03/2016	3769
6	Philippines	Emerging	PSE	05/12/2000 till 09/03/2016	3824
7	China	Emerging	Shanghai Composite (SSE)	05/12/2000 till 09/03/2016	3696
8	Taiwan	Emerging	TWI	05/12/2000 till 09/03/2016	3763

Source Researcher's Distillation

5 Sample Data

The sample consists of daily closing level data of stock market indices of India and Asian emerging economies like Hong Kong, Indonesia, Malaysia, South Korea, Philippines, China and Taiwan from December 5th, 2000 through March 9th, 2016. The daily closing level data of eight indices has been obtained from www.finance.yahoo.com. The description of the selected economies, their type, period of the study and total number of observations included in the study are shown in Table 1.

6 Empirical Estimation and Results

6.1 Test of Stationarity

To test stationarity of the data, first plot the graph of the variables over time. Figure 1 plots the index values of India, Hong Kong, Indonesia, Malaysia, South Korea, Philippines, China and Taiwan, and Fig. 2 plots the daily stock prices, and Fig. 3 plots the

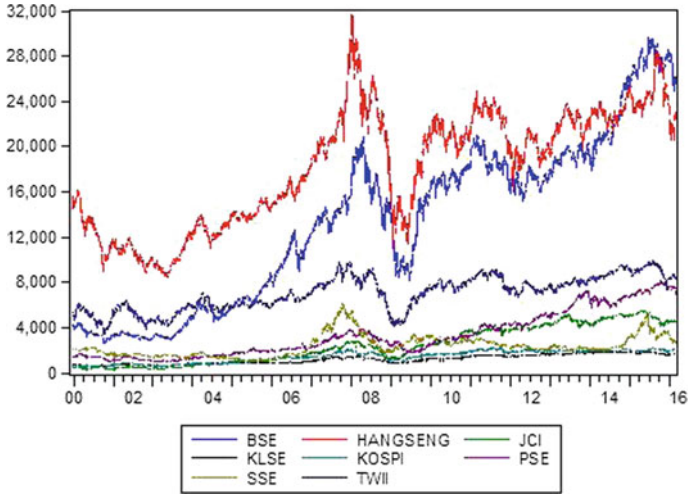

Fig. 1 India (BSE), Hong Kong (HSI), Indonesia (JCI), Malaysia (KLSE), South Korea (KOSPI), Philippines (PSE), China (SSE) and Taiwan (TWI)

daily stock returns of the countries of India, Hong Kong, Indonesia, Malaysia, South Korea, Philippines, China and Taiwan over the period, respectively.

The graphical presentations above in Fig. 2 indicate that the variables are found to be having a trend, implying that the data are non-stationary in nature. However, in order to prove it statistically, the Augmented Dickey–Fuller (ADF) test Eq. (1) for unit root has been conducted and to verify the results of ADF test, the Phillips and Peron (PP) test Eq. (2) of stationarity have been conducted. From the unit root test shown in Table 2, it is found that all level data are non-stationary (the null hypothesis of a unit root is accepted) as the ADF and PP test statistics are more than the 95 per cent critical value. But, when the ADF and PP tests are again applied to the first differences of the selected indices (see Table 2), they became stationary (the null hypothesis of a unit root is rejected) as the above test statistics are less than the 95 per cent critical value. In order to test for co-integration of the selected indices, the Engle–Granger methodology is used.

7 Descriptive Statistics

Table 3 presents the descriptive statistics obtained from the log of first difference level data (return) like mean, median, maximum, minimum, standard deviation, skewness, kurtosis, Jarque–Bera, probability, sum square deviation of the eight variables BSE, HSI, JCI, KLSE, KOSPI, PSE, SSE and TWI. The average daily closing level price

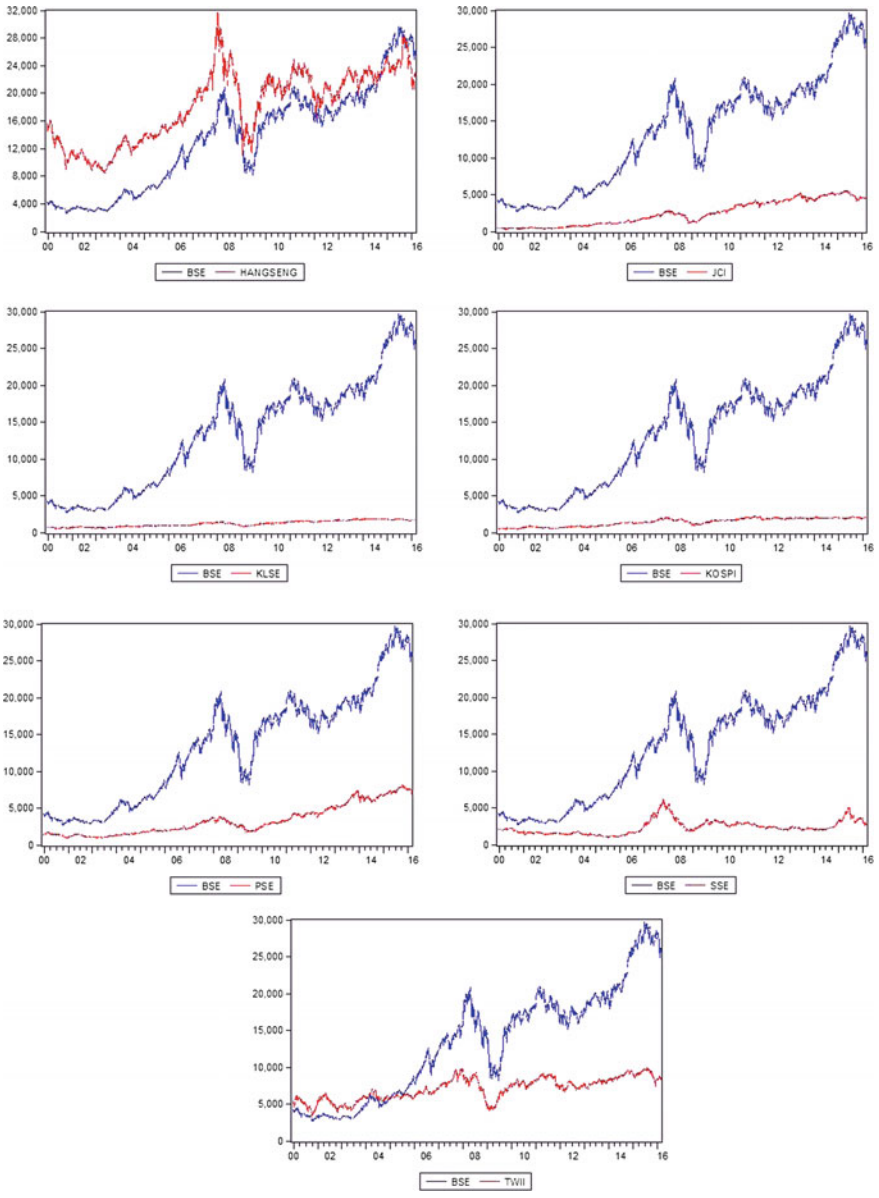

Fig. 2 Daily stock prices

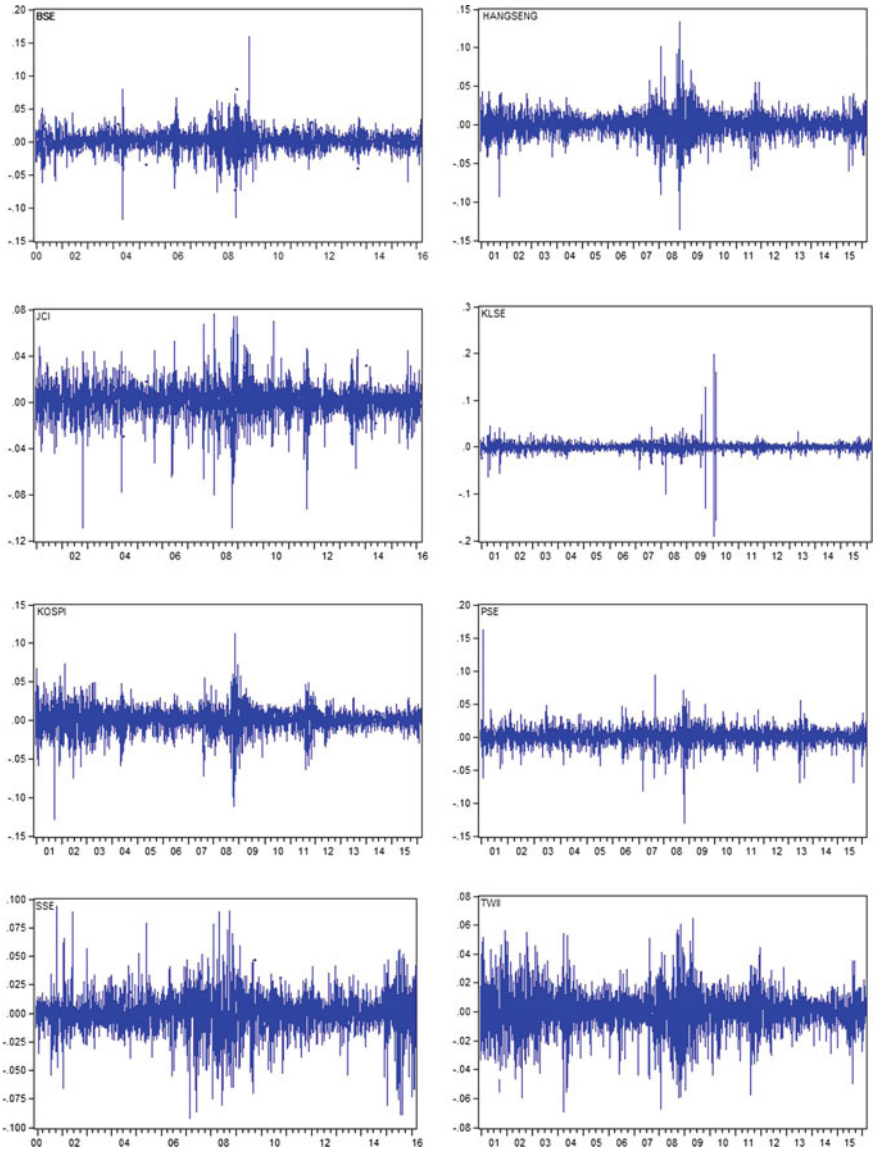

Fig. 3 Daily stock returns

Table 2 ADF and PP test results of level data

ADF and PP test results of level data						
Name of the index	ADF test results			PP test results		
	Computed value	MacKinnon critical value at 5% level	P value	Computed value	MacKinnon critical value at 5% level	P value
BSE	−0.76	−2.86	0.83	−0.78	−2.86	0.82
HSI	−1.38	−2.86	0.59	−2.33	−2.86	0.61
JCI	−1.21	−2.86	0.67	−1.23	−2.86	0.66
KLSE	−0.87	−2.86	0.79	−0.92	−2.86	0.78
KOSPI	−1.89	−2.86	0.34	−1.88	−2.86	0.34
PSE	−0.39	−2.86	0.91	−0.32	−2.86	0.92
SSE	−1.26	−2.86	0.65	−1.34	−2.86	0.62
TWI	−2.02	−2.86	0.27	−2.00	−2.86	0.29
ADF and PP test results of first difference						
Name of the index	ADF test results			PP test results		
	Computed value	MacKinnon critical value at 5% level	P value	Computed value	Critical value at 5% level	P value
BSE	−43.45	−1.94	0.00	−56.15	−1.94	0.00
HSI	−61.94	−1.94	0.00	−61.99	−1.94	0.00
JCI	−54.23	−1.94	0.00	−54.06	−1.94	0.00
KLSE	−47.49	−1.94	0.00	−61.93	−1.94	0.00
KOSPI	−59.39	−1.94	0.00	−59.40	−1.94	0.00
PSE	−53.95	−1.94	0.00	−53.63	−1.94	0.00
SSE	−59.55	−1.94	0.00	−59.62	−1.94	0.00
TWI	−57.45	−1.94	0.00	−57.47	−1.94	0.00

Note Null Hypothesis: There is unit root. Alternative Hypothesis: There is no unit root
Source Compiled from E Views Output

and standard deviation for the eight stock market indices are almost different for the period under study. During the observed period, JCI index shows the highest average rate of return which is followed by BSE. The skewness statistics of daily return is found to be negative for BSE, HSI, JCI, KLSE, KOSPI, PSE, SSE and TWI which indicates these eight indices are negatively skewed. Kurtosis is more than three for all the eight indices during the period suggests that the underlying data is leptokurtic, i.e. squat with short tails about the mean, which indicates that the data is not normally distributed. Application of Jarque–Bera (JB) statistics calculated to test the null hypothesis of normality in the data rejects the normality assumption at

Table 3 Descriptive statistics of log (ln) of first difference level data (Return)

Statistical results	RBSE	RHSI	RJCI	RKLSE	RKOSPI	RPSE	RSSE	RTWI
Mean	0.000504	0.000125	0.000629	0.000226	0.000361	0.000432	8.49E-05	0.000130
Median	0.001040	0.000289	0.001212	0.000442	0.000790	0.000199	0.000548	0.000452
Maximum	0.159900	0.134068	0.076234	0.198605	0.112844	0.161776	0.094008	0.065246
Minimum	−0.118092	−0.135820	−0.109539	−0.190647	−0.128047	−0.130887	−0.092562	−0.069123
Std. Dev.	0.015084	0.015069	0.014056	0.010445	0.014893	0.012968	0.016779	0.013614
Skewness	−0.140915	−0.019191	−0.663266	−0.125091	−0.505292	−0.075745	−0.363478	−0.175160
Kurtosis	10.93892	11.69207	9.550527	111.9338	9.156007	16.01275	7.215176	5.638234
Jarque–Bera	9718.304	11635.25	6879.040	1827460	5993.326	26080.61	2817.610	1090.783
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	1.862429	0.461351	2.325461	0.835730	1.334748	1.598400	0.313754	0.479744
Sum Sq. Dev.	0.840762	0.839094	0.730070	0.403124	0.819565	0.621422	1.040307	0.684795
Observations	3696	3696	3696	3696	3696	3696	3696	3696

Table 4 Pair-wise correlation matrix of Log (ln) of level data

	ln BSE	ln HSI	ln JCI	ln KLSE	ln KOSPI	ln PSE	ln SSE	ln TWI
ln BSE	1	0.9464	0.9725	0.9484	0.9646	0.9294	0.6687	0.8810
ln HSI		1	0.9035	0.9032	0.9191	0.8842	0.7225	0.9109
ln JCI			1	0.9787	0.9688	0.9442	0.6077	0.8659
ln KLSE				1	0.9511	0.9598	0.6040	0.8987
ln KOSPI					1	0.8989	0.6249	0.8972
ln PSE						1	0.5835	0.8371
ln SSE							1	0.6524
ln TWI								1

1% level of significance. The results confirm the well-known fact that daily return of the indices under consideration is not normally distributed and so it is skewed.

8 Correlation

From Table 4, we can see that there significantly high positive correlation among eight Asian emerging countries.

From Table 5, we can see that there significantly low positive correlation among Asian emerging countries but the correlation between Indian market and South Korean market, Hong Kong market and Indonesian market, Indonesian market and Philippines market, Indonesian market and Taiwan market, etc.

Table 5 Pair-wise correlation matrix of Log (ln) of first difference level data (Return)

	RBSE	RHSI	RJCI	RKLSE	RKOSPI	RPSE	RSSE	RTWI
RBSE	1	0.0429	0.02345	0.01440	−0.0313	0.0203	0.0218	0.0319
RHSI		1	−0.0397	0.0539	0.0034	0.0316	0.0322	0.0130
RJCI			1	0.0025	0.0292	−0.01954	0.0068	−0.0003
RKLSE				1	0.0209	0.0453	−0.0028	0.0697
RKOSPI					1	−0.0062	−0.00400	0.0346
RPSE						1	0.02923	−0.0195
RSSE							1	0.0227
RTWI								1

9 Co-integration

In order to test for the co-integration at first we run the regression. The results of Table 6 shows the symptoms of spurious regression as R square values are greater than the lower D-W statistics values. So, new regression model with trend is adopted. If the residuals of the new trend model are stationary, then it would remain no longer spurious. Hence, this model can be used for co-integration purposes. In other words the model is a long-run model. After running the regression of all the trend models as shown in Table 7, the ADF test (EG test) is applied on the residuals. Here, the p-values of EG test statistics are less than the 95 percent critical value so, the null hypothesis of the presence of unit root is rejected except for the (BSE and KLSE), (BSE and TWII) and (TWII and BSE). It means the residuals are stationary, hence all the regression models are co-integrated except the above three cases. As shown in Table 8 the all the R-squared values are less than Durbin-Watson (DW) statistics and these statistics are very close to two. It implies that all the residuals are stationary.

This paper has used daily closing prices of stock market indices from India and seven emerging Asian countries. The ADF test characterized all series as I (1) allowing to proceed with the co-integration analysis. Johansen procedure test for co-integration is used as a benchmark test. From Table 9, it is shown that there is not any evidence of co-integration between India and the other countries like Malaysia, Philippines and China in the sample; however, there is strong evidence of co-integration between India and these countries like Hong Kong, Indonesia, South Korea and Taiwan. The trace test rejects the null hypothesis of no co-integration with a low p-value and the maximum Eigen value test rejects the null of no co-integration with a low p-value. Since the null hypothesis of at least one co-integrating relationship cannot be rejected, it is shown that the stock markets of India with Hong Kong, Indonesia, South Korea and Taiwan do have a long-run relationship.

The open-ended model is assumed for estimation as:

$$f(lnBSE, lnHSI, lnJCI, lnKLSE, lnKOSPI, lnPSE, lnSSE \,\&\, lnTWII) = 0 \quad (4)$$

Table 6 Regression (Spurious) models of level data

Regression model (Spurious)	C(1)	C(2)	p-value	R-squared	Durbin–Watson
ln BSE = C(1) + (C(2) * ln HANGSENG)	−11.79	2.16	0.00	0.89	0.02
ln HANGSENG = C(1) + (C(2) * ln BSE)	5.91	0.41	0.00	0.89	0.02
ln BSE = C(1) + (C(2) * ln JCI)	3.41	0.78	0.00	0.94	0.01
ln JCI = C(1) + (C(2) * ln BSE)	−3.70	1.20	0.00	0.94	0.01
ln BSE = C(1) + (C(2) * ln KLSE)	−4.06	1.89	0.00	0.89	0.01
ln KLSE = C(1) + (C(2) * ln BSE)	2.63	0.47	0.00	0.89	0.01
ln BSE = C(1) + (C(2) * ln KOSPI)	−1.94	1.56	0.00	0.93	0.02
ln KOSPI = C(1) + (C(2) * ln BSE)	1.65	0.59	0.00	0.93	0.02
ln BSE = C(1) + (C(2) * ln PSE)	0.70	1.08	0.00	0.86	0.01
ln PSE = C(1) + (C(2) * ln BSE)	0.51	0.79	0.00	0.86	0.01
ln BSE = C(1) + (C(2) * ln SSE)	−0.37	1.25	0.00	0.45	0.00
ln SSE = C(1) + (C(2) * ln BSE)	4.38	0.35	0.00	0.45	0.00
ln BSE = C(1) + (C(2) * ln TWII)	−14.89	2.73	0.00	0.77	0.01
ln TWII = C(1) + (C(2) * ln BSE)	6.20	0.28	0.00	0.77	0.01

The Johansen Maximum Likelihood procedure is applied to the VAR formed by non-stationary eight variables. In Table 10, Johansen's test and normalized co-integrating vector results are presented. From λ-max and λ-trace statistics, it is investigated that there exists at least four co-integrating vectors can be framed (See bottom of Table 10) for Eq. (4). Next, the significance of each variable in the co-integrating relation is tested by using LR test statistics given by Johansen, which is asymptotically chi-square with one degree of freedom. The variables which are statistically significant can contribute to the long-run equilibrium relationship. From Table 10, it is observed that all the eight variables are co-integrated with four co-integrating vectors. The λ or Eigen value statistic drops roughly for alternative hypothesis of eighth co-integrating vector. Thus, it can be concluded that our models with eight variables are fair representation for long-run relationship. It can also be concluded that Philippines stock market influences both the Indian stock market and Hong Kong stock market positively and significantly where Taiwan stock market influences both the Indonesian stock market and Malaysian stock market positively and significantly.

The study further employs Granger causality test to test the cause and effect relationship between markets under consideration. Granger causality test is conducted by considering the return series of all the indices under consideration. From Table 11, it is clear that the Indian capital market is not getting significantly affected by all of the Asian emerging market economies. Hence, at the time of forecasting the position

Table 7 Regression and trend models of level data

Regression and trend model	C(1)	C(2)	p-value	C(3)	R-squared	Durbin–Watson
ln BSE = C(1) + (C(2) * ln HANGSENG) + (C(3) * TIME)	−3.57	1.25	0.00	0.01	0.96	0.03
ln HANGSENG = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	4.41	0.60	0.00	−0.01	0.92	0.03
ln BSE = C(1) + (C(2) * ln JCI)) + (C(3) * TIME)	3.20	0.82	0.00	−0.01	0.94	0.01
ln JCI = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	0.28	0.70	0.00	0.01	0.96	0.01
ln BSE = C(1) + (C(2) * ln KLSE)) + (C(3) * TIME)	0.01	1.25	0.00	0.01	0.91	0.01
ln KLSE = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	4.51	0.23	0.00	0.01	0.93	0.01
ln BSE = C(1) + (C(2) * ln KOSPI)) + (C(3) * TIME)	1.15	1.07	0.00	0.01	0.95	0.01
ln KOSPI = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	1.87	0.56	0.00	0.01	0.93	0.02
ln BSE = C(1) + (C(2) * ln PSE)) + (C(3) * TIME)	4.68	0.49	0.00	0.01	0.88	0.01
ln PSE = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	4.78	0.26	0.00	0.01	0.92	0.01
ln BSE = C(1) + (C(2) * ln SSE)) + (C(3) * TIME)	4.99	0.42	0.00	0.01	0.90	0.01
ln SSE = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	1.89	0.66	0.00	−0.01	0.49	0.01
ln BSE = C(1) + (C(2) * ln TWII)) + (C(3) * TIME)	−2.10	1.20	0.00	0.01	0.92	0.01
ln TWII = C(1) + (C(2) * ln BSE)) + (C(3) * TIME)	5.46	0.37	0.00	−0.01	0.78	0.02

Table 8 Engle–Granger test for residual of level data

Residual model	Computed value	MacKinnon critical value at 5% Level	p-value	R-squared	Durbin–Watson
u = ln BSE-C(1) − (C(2) * ln HANGSENG) − (C(3) * TIME)	−5.85	−1.94	0.00	0.01	1.97
u = ln HANGSENG − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−6.24	−1.94	0.00	0.01	1.99
u = ln BSE − C(1) − (C(2) * ln JCI)) − (C(3) * TIME)	−3.41	−1.94	0.00	0.01	1.83
u = ln JCI − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−3.06	−1.94	0.00	0.01	1.81
u = ln BSE − C(1) − (C(2) * ln KLSE)) − (C(3) * TIME)	−0.11	−1.94	0.64	0.01	1.93
u = ln KLSE − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−3.58	−1.94	0.00	0.01	2.01
u = ln BSE − C(1) − (C(2) * ln KOSPI)) − (C(3) * TIME)	−4.71	−1.94	0.00	0.01	1.85
u = ln KOSPI − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−4.99	−1.94	0.00	0.01	1.88
u = ln BSE − C(1) − (C(2) * ln PSE)) − (C(3) * TIME)	−2.00	−1.94	0.00	0.01	1.84
u = ln PSE − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−2.64	−1.94	0.01	0.01	1.78
u = ln BSE − C(1) − (C(2) * ln SSE)) − (C(3) * TIME)	−2.24	−1.94	0.02	0.01	1.87
u = ln SSE − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	−2.24	−1.94	0.02	0.01	1.93
u = ln BSE − C(1) − (C(2) * ln TWII)) − (C(3) * TIME)	0.75	−1.94	0.87	0.01	1.85
u = ln TWII − C(1) − (C(2) * ln BSE)) − (C(3) * TIME)	1.16	−1.94	0.93	0.01	1.87

of Indian capital market, the investors should not take into account the positions of the rest of the markets listed above.

Table 9 Johansen–Juselius test results for bivariate data

Country	Null hypothesis	Alternative hypothesis	Eigen value	Trace statistic	Critical value 5%	p-value	Maximum Eigen value statistic	Critical value 5%	p-value
India–Hong Kong	$r = 0$	$r \geq 1$	0.012276	49.07165	25.87211	0.00	45.59169	19.38704	0.00
India–Indonesia	$r = 0$	$r \geq 1$	0.005974	24.35493	25.87211	0.0763	22.11633	19.38704	0.0196
India–Malaysia	$r = 0$	$r \geq 1$	0.004092	18.66785	25.87211	0.3008	15.13487	19.38704	0.1864
India–South Korea	$r = 0$	$r \geq 1$	0.011968	47.74775	25.87211	0.00	44.43896	19.38704	0.00
India–Philippines	$r = 0$	$r \geq 1$	0.003928	19.76932	25.87211	0.2378	14.52617	19.38704	0.2207
India–China	$r = 0$	$r \geq 1$	0.001745	8.970097	25.87211	0.9612	6.447017	19.38704	0.9344
India–Taiwan	$r = 0$	$r \geq 1$	0.008013	33.99802	25.87211	0.0039	29.69430	19.38704	0.0011

Table 10 Johansen's co-integration test for multivariate data

VAR	Null hypothesis	Alternative hypothesis	Eigen value	Trace statistic	Critical value 5%	p-value	Maximum Eigen value statistic	Critical value 5%	p-value
Variable under study: ln BSE, ln HSI, ln JCI, ln KLSE, ln KOSPI, ln PSE, ln SSE, ln TWII									
VAR(4)	$r = 0$*	$r = 1$	0.0289	308.2718	159.5297	0.0000	108.2300	52.3626	0.0000
	$r \leq 1$*	$r = 2$	0.0180	200.0418	125.6154	0.0000	66.94282	46.2314	0.0001
	$r \leq 2$*	$r = 3$	0.0141	133.0990	95.7537	0.0000	52.5069	40.0776	0.0012
	$r \leq 3$*	$r = 4$	0.0102	80.5921	69.8189	0.0054	37.6597	33.8769	0.0168
	$r \leq 4$	$r = 5$	0.0074	42.9324	47.8561	0.1342	27.4185	27.5843	0.0525
	$r \leq 5$	$r = 6$	0.0023	15.5139	29.7971	0.7459	8.8091	21.1316	0.8475
	$r \leq 6$	$r = 7$	0.0011	6.7040	15.4947	0.6123	4.1349	14.2646	0.8448
	$r \leq 7$	$r = 8$	0.0007	2.5690	3.8415	0.1090	2.5690	3.8415	0.1090
LR estimate	ln BSE = − 1.5344 ln KOSPI + 0.0567 ln PSE − 0.0763 ln SSE − 0.3993 ln TWII (0.1705) (0.0991) (0.0900) (0.2712)								
LR estimate	ln HSI = − 0.4541 ln KOSPI + 0.1412 ln PSE − 0.1006 ln SSE − 0.8146 ln TWII (0.0987) (0.0574) (0.0521) (0.1571)								
LR estimate	ln JCI = − 1.5703 ln KOSPI − 0.7908 ln PSE − 0.0368 ln SSE + 1.3504 ln TWII (0.01829) (0.1063) (0.0966) (0.2910)								
LR estimate	ln KLSE = − 0.2831 ln KOSPI − 0.5409 ln PSE − 0.0574 ln SSE + 0.5804 ln TWII (0.0985) (0.0573) (0.0520) (0.1568)								

Notes (i) *indicates significant at 5% levels as computed by MacKinnon-Haug-Michelis (1999), (ii) Figures in parenthesis represent the t-statistics

Table 11 Granger causality test

Pair-wise Granger causality test			
Null hypothesis	Observations	F statistic	p-value
BSE does not Granger Cause HANGSENG	3694	2.32	0.09
HANGSENG does not Granger Cause BSE	3694	1.56	0.21
BSE does not Granger Cause JCI	3694	1.40	0.24
JCI does not Granger Cause BSE	3694	1.74	0.17
BSE does not Granger Cause KLSE	3694	1.54	0.21
KLSE does not Granger Cause BSE	3694	0.89	0.40
BSE does not Granger Cause KOSPI	3694	1.94	0.14
KOSPI does not Granger Cause BSE	3694	2.37	0.09
BSE does not Granger Cause PSE	3694	0.98	0.37
PSE does not Granger Cause BSE	3694	0.60	0.54
BSE does not Granger Cause SSE	3694	0.68	0.50
SSE does not Granger Cause BSE	3694	1.57	0.20
BSE does not Granger Cause TWII	3694	2.36	0.09
TWII does not Granger Cause BSE	3694	3.02	0.04

Source Compiled from E views output

10 Summary and Conclusions

This paper examines the long-run equilibrium relationship of India with selected Asian emerging markets from December 2000 to March 2016. Particularly, study examines if index prices in these eight Asian emerging markets are co-integrated. By using co-integration test, it is observed that there is not any evidence of co-integration between India and the other countries like Malaysia, Philippines and China in the sample; however, there is strong evidence of co-integration between India and these countries like Hong Kong, Indonesia, South Korea and Taiwan. Through multivariate co-integration test, it is concluded that Philippines stock market influences both the Indian stock market and Hong Kong stock market positively and significantly where Taiwan stock market influences both the Indonesian stock market and Malaysian stock market positively and significantly. Through Granger causality test, we find that the Indian capital market is not getting significantly affected by all of the emerging market economies. Hence, the long-run equilibrium relationship implies that the indices are perfectly correlated in the long run and diversification between Indian stock market and other tested Asian emerging stock markets cannot benefit international portfolio investors. However, there can be excess returns in the short run. The present study has further scope for more comprehensive results. It can be extended over a longer period and structural break concept can be implemented. Further, research area can be extended by analysing the economy and stock markets of various developed and developing nations.

The major implication of this study can be for financial stock market, such as emerging Asian stock market in this case, as it should concentrate on investment in Indian stock market for short run because it is not significantly affected by all other Asian emerging markets. But from long-run equilibrium point of view, Philippines stock market influences both the Indian stock market and Hong Kong stock market positively and significantly where Taiwan stock market influences both the Indonesian stock market and Malaysian stock market positively and significantly. Thus, this will lead to increase in liquidity conditions of the markets and the markets may become more predictable and more efficient.

Acknowledgements The authors wish to express their gratitude to the editor and anonymous reviewers for their valuable suggestions and comments which significantly improved the original paper.

References

Adjasi, C., Biekpe, K. D., & Nicholas, B. (2006). Co-integration and dynamic causal links amongst African stock markets. *Investment Management and Financial Innovations, 3*(4), 102–119.

Bangake, C., & Eggoh, J. C. (2010). Finance-growth link in OECD countries: Evidence from panel causality and co-integration tests. *Brussels Economic Review, 53*(3/4), 375–392.

Batori, O. A. (2009). Relationships among European equity markets: Multivariate co-integration and causality evidence across developed and emerging countries. *Journal of International Finance and Economics, 9*(4), 82–96.

Bollerslev, T., & Jubinski, D. (1999). Latent information arrivals and common long-run dependencies. *Journal of Business & Economic Statistics, 17*(1), 9–21.

Bose, S., & Mukerjee, P. (2005). *A study of inter linkages between the Indian stock market and some other emerging and developed markets*. IMF: World Economic Outlook.

Bose, S., & Mukherjee, P. (2006). A study of inter linkages between the Indian stock market and some other emerging and developed markets. Indian Institute of Capital Markets, 9th Capital Markets Conference Paper (pp. 1–13).

Gikas, A. H., Dimitrios, M., & Richard, P. (2006). EMU and European stock market integration. *The Journal of Business, 79*(1), 365–392. http://www.jstor.org/stable/10.1086/497414.

Ivanov, S. I. (2011). The effects of Crisis on the co-integration between the S&P 100 and the S&P 500 indexes. *The International Journal of Finance, 23*(2), 6784–6497.

Jeyanthi, B., Queensly, J., & Pandian, P. (2008). An empirical study of co-integration and correlation among Indian, emerging and developed markets. *The ICFAI Journal of Applied Finance, 14*(11), 35–47.

Mittoo, U. R. (1992). Additional evidence on integration in the canadian stock market. *The Journal of Finance, 47,* 2035–2054. https://doi.org/10.1111/j.1540-6261.1992.tb04696.x.

Tabak, B. M., & Lima, E. J. A. (2002). Causality and co-integration in stock markets: The case of Latin America Banco Central do Brasil. Working Paper Series n. 56.

Tahai, A., Rutledge, R. W., & Karim, K. E. (2004). An examination of financial integration for the group of seven (G7) industrialized countries using an I(2) co-integration model. *Applied Financial Economics, 14*, 327–335.

Wong, W.-K., Agarwal, A., & Du, J. (2005). Financial integration for India stock market, a fractional co-integration approach, Department of Economics, National University of Singapore, Working Paper No. 0501.

Wong, W.-K., Penm, J., Terrel, R. D., Lim, K., & Yann, C. (2004). The relationship between stock markets of major developed countries and asian emerging markets. *Journal of Applied Mathematics and Decision Sciences, 8*(4), 201–218.

Yuhn, K.-H. (1997). Financial integration and market efficiency: Some international evidence from co-integration tests. *International Economic Journal, 11*(2), 103–116.

Optimum Designs for Pharmaceutical Experiments with Relational Constraints on the Mixing Components

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

Abstract In pharmaceutical experiments, the rate of dissolution of a tablet is modeled in terms of the proportions of polymers and diluents used in the tablet. When more than one type of polymer and diluent are used, the total proportions of polymers and diluents in the tablet are generally subject to relational constraints, which give a range of acceptable values for each proportion. This paper considers two models for the mean dissolution rate subject to relational constraints on the polymers and diluents and attempts to find optimum designs for estimating the parameters in the models using the D-optimality criterion.

Keywords Pharmaceutical experiment · Relational constraint
Major and minor components · D-optimality

AMS Subject Classification 62K99 · 62J05

1 Introduction

The modeling of a tablet dissolution is a key area in the field of pharmaceutical research. The two important components affecting mean dissolution time are polymer and diluent, and the mean dissolution time is generally expressed as a function of the proportions of polymers and diluents used in the tablet. The proportions of individual polymers and diluents can vary from 0 to 1 within their respective classes, viz. (a) polymer and (b) diluent. Each class defines a major component (M-component), while each member of a class is called a minor component (m-component). The number of members of a M-component is normally not very large, say two or three (cf. Lewis et al. 2010). Some models for defining the mean dissolution time with

M. Pal · N. K. Mandal
Department of Statistics, University of Calcutta, Kolkata, India

B. K. Sinha (✉)
Indian Statistical Institute, Kolkata, India
e-mail: bikashsinha1946@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_3

major components polymer (A and B) and diluent (D_1, D_2, D_3) have been suggested in Lewis et al. (2010) wherein the mean response is described by a quadratic Scheffé model (cf. Scheffé 1958) in the proportions of the polymers A and B:

$$\eta_x = \sum_{i=1}^{5} \beta_i x_i + \beta_{12} x_1 x_2 \tag{1.1}$$

$$\eta_x = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2, \;\; \beta_{\mathrm{i}} = \sum_{j=3}^{5} \alpha_{ij} x_j, i = 1, 2, \;\; \beta_{12} = \sum_{j=3}^{5} \alpha_{12j} x_j \tag{1.2}$$

where (x_1, x_2) are the mixing proportions of (A, B), and (x_3, x_4, x_5) are the mixing proportions of (D_1, D_2, D_3) satisfying $0 \le x_i \le 1, i = 1, 2, \sum_{i=1}^{5} x_i = 1$.

In other words, the underlying linear models are described as

$$Y_{\boldsymbol{x}} = \eta_{\boldsymbol{x}} + \varepsilon,$$

where $Y_{\boldsymbol{x}}$ refers to the response, and the right-hand side corresponds to the mean response and the error. In this study, the errors are assumed to be uncorrelated with mean zero and common variance σ^2.

For specified fixed proportions of the major components polymer and diluent, Lewis et al. (2010) show that the D-optimal design for parameter estimation in model (1.2) is a 9-point design obtained by multiplying the second-order Scheffé design for the polymers (3 points) and the first-order Scheffé design for diluents (3 points). Generalizing the model to include m and n minor components, respectively, in the major components polymer and diluent, Pal and Mandal (2013) discuss optimum designs for parameter estimation and also for estimating the optimum mixing proportions of minor components in polymer for the case of fixed mixing proportions of polymer and diluents. Model (1.1), however, is appropriate when the proportions of the major components lie within given ranges in [0, 1]. Lewis et al. (2010) indicate that for the case of $m = 2$ and $n = 3$, the exchange algorithm can be used to select nine design points from the 18 design points (extreme points and mid-points of edges) which give the D-optimal design for parameter estimation. They, however, assume equal masses for the designs points. And, their proposed design lacks invariance, though invariance among the minor components of the major components is very apparent.

As a second example, it may be mentioned that in traditional agricultural experiments, we can think of two major components—(i) the class of nutrients in the soil that are required by plants in large quantities, and on which the plants mainly depend for their growth. These nutrients are also usually present in abundance in the soil. Examples of such nutrients are: nitrogen (N), phosphorus (P), potassium (K), magnesium (Mg), calcium (Ca), and hydrogen (H); (ii) the class of nutrients found in the soil that are required by plants in tiny quantities. When they are supplied in large quantities it might be detrimental to the plants. Some examples are: cobalt (Co), iron (Fe), copper (Cu), molybdenum (Mo), zinc (Zn), manganese (Mn), sodium (Na), boron (B), aluminum (Al), and Chlorine (Cl). A mixture of these major components

is what is required for healthy growth of plants. The relational constraints are then naturally called for.

For an updated account of mixture models and methods as also of optimality issues, we refer to a recent monograph by Sinha et al. (2014).

In this paper, we consider the models (1.1) and (1.2) with two major components, M_1 and M_2, having m and n minor components, respectively, and with their proportions in the mixture being subject to specified lower and upper bounds. We discuss optimum designs for estimation of the model parameters, using D-optimality criterion. We exclusively deal with the setup of the first example cited above. The paper is organized as follows. In Sect. 2, we discuss the models and their perspectives. In Sect. 3, the optimum designs for parameter estimation are investigated. Finally, a discussion of our findings is given in Sect. 4.

2 The Models

Consider a mixture of two major components M_1 (polymer) and M_2 (diluent), there being m minor components in M_1 with proportions $\boldsymbol{x}_{(1)} = (x_1, x_2, \ldots, x_m)'$ and n minor components in M_2 with proportions $\boldsymbol{x}_{(2)} = (x_{m+1}, x_{m+2}, \ldots, x_{m+n})'$, where $0 \le x_i \le 1, \sum_{i=1}^{m+n} x_i = 1$. Suppose the proportions of the major component M_1 are required to satisfy $\delta_1 \le \sum_{i=1}^{m} x_i \le \delta_2, \quad 0 < \delta_1 < \delta_2 < 1$.

The mean response $\eta_{\boldsymbol{x}}$ is assumed to be

$$\eta_{\mathrm{x}} = \sum_{i=1}^{m+n} \beta_i x_i + \sum_{i<j=1}^{m} \beta_{ij} x_i x_j, \tag{2.1}$$

and the experimental region is given by

$$\Xi = \{(x_1, x_2, \ldots x_{m+n}) | 0 \le x_i \le 1,\ i = 1, 2, \ldots, m+n, \delta_1 \le \sum_{\mathrm{i}=1}^{\mathrm{m}} x_i \le \delta_2, \quad 0 < \delta_1 < \delta_2 < 1\}. \tag{2.2}$$

We can express the experimental region as $\Xi = \Xi_1 \cap \Xi_2$, where

$$\Xi_1 = \{(x_1, x_2, \ldots x_m) | 0 \le x_i \le 1,\ i = 1, 2, \ldots, m,\ \delta_1 \le \sum_{\mathrm{i}=1}^{\boldsymbol{m}} x_i \le \delta_2\}, \tag{2.3}$$

$$\Xi_2 = \{(x_{m+1}, x_{m+2}, \ldots x_{m+n}) | 0 \le x_i \le 1,\ i = m+1, m+2, \ldots, m+n,$$
$$1 - \delta_2 \le \sum_{i=m+1}^{m+n} x_i \le 1 - \delta_1, 0 < \delta_1 < \delta_2 < 1.\} \tag{2.4}$$

Clearly, Ξ_1 and Ξ_2 denote the experimental regions of M_1 and M_2, respectively.

We revert to a specific design issue studied by Lewis et al. (2010). They started with a specific case of $m = 2$ and $n = 3$ with reference to model (2.1). For estimation of the six model parameters, they confined to suitably defined subclasses of 18 design

points and further reduced the subclass to 6–9 point designs. In the process, they suggested a 9-point design with the desirable features of D-optimality.

In this paper, we take the clue from Lewis et al. (2010) and proceed further.

3 Parameter Estimation and Optimality Considerations

We confine to the mixture model (2.1) and develop strategies for optimal estimation of the model parameters.

3.1 Study of Optimal Mixture Design Model (2.1)

Note that model (2.1) is a general version of model (1.1). Let, $\mathcal{D}$ be the class of all competing continuous designs, for which all the parameters of (2.1) are estimable. We want to find a design ξ in $\mathcal{D}$ that can estimate the parameters with maximum accuracy.

For a continuous design $\xi \in \mathcal{D}$:

$$\delta = \left\{\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_N^*; w_1, w_2, \ldots, w_N\right\}, \tag{3.1}$$

with masses $w_1, w_2, \ldots, w_N$, $w_i > 0$, $\sum w_i = 1$, at points $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots, \boldsymbol{x}_N^*$, the information (moment) matrix (under homoscedastic uncorrelated error model) is given by $M(\xi) = \sum w_i \boldsymbol{f}(x_i^*) f(x_i^*)'$, where

$$\boldsymbol{f}(\boldsymbol{x}) = (x_1, x_2, \ldots, x_m, x_1x_2, x_1x_3, \ldots, x_1x_m, \ldots, x_{m-1}x_m, x_{m+1}, x_{m+2}, \ldots, x_{m+n})' = \left(f_1(\boldsymbol{x}_{(1)})', \boldsymbol{x}_{(2)}{}'\right)',$$

$$\boldsymbol{f}_1(\boldsymbol{x}_{(1)}) = (x_1, x_2, \ldots, x_m, x_1x_2, x_1x_3, \ldots, x_{m-1}x_m)'.$$

Design optimality aims at minimizing some function of $M^{-1}(\xi)$ or maximizing some function of $M(\xi)$. For comparing different designs in $\mathcal{D}$, let us consider the D-optimality criterion, given by:

$$\text{Maximize}\, \varphi_D(\xi), \text{ where } \varphi_D(\xi) = \text{Det.}(M(\xi)). \tag{3.2}$$

The above criterion is invariant with respect to the components in $x_{(1)}$ and $x_{(2)}$.

Theorem 3.1 *There exists an invariant mixture design under model (2.1) having $mn(m+2)$ design points with uniform mass distribution, and in case of $m = 2$ and $n = 3$, this mixture design outperforms the design suggested by Lewis et al. (2010).*

Proof Not to obscure the essential steps of reasoning, we proceed through the following stages.

Stage 1: In this step, we develop a logical analysis for choice of the invariant support points of the ***proposed*** mixture design.

When the mean response has quadratic dependence on the mixing proportions, a reasonable choice of the experimental points would be the extreme points of the design space and mid-points on the edges of the space when the optimality criterion is invariant with respect to the proportions. In case of linear dependence, only the extreme points of the design space seem to be the reasonable choice. In the present situation, the model (2.1) is linear in $x_{(2)}$ and quadratic in $x_{(1)}$, and the extreme points of Ξ_1 and Ξ_2 are, respectively, $(\delta_1, 0, \ldots, 0)$ and $(\delta_2, 0, \ldots, 0)$ and all possible permutations within these, and $(1-\delta_1, 0, \ldots, 0)$ and $(1-\delta_2, 0, \ldots, 0)$ and all possible permutations within these. The mid-points of the edges of Ξ_1 are $(\delta_1/2, \delta_1/2, 0, \ldots, 0)$, $(\delta_2/2, \delta_2/2, 0, \ldots, 0)$, $(\delta_0, 0, 0, \ldots, 0)$, and all possible permutations within these where $\delta_0 = (\delta_1 + \delta_2)/2$.

In view of the above, we may start with a class $\mathcal{D}_1$ of designs ξ with support points given by

(i) $(\delta_i, 0, \ldots, 0; 1-\delta_i, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates each with mass v_i, $i = 1, 2$;
(ii) $(\delta_0, 0, 0, \ldots, 0; 1-\delta_0, 0, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates each with mass v_0;
(iii) $(\delta_i/2, \delta_i/2, 0, \ldots, 0; 1-\delta_i, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates, each with mass w_i, $i = 1, 2$.

We refer to the support points in (i) and (ii) as pure type support points and those in (iii) as mixed type support points (vide Pal and Mandal 2013). For any design $\xi \in \mathcal{D}_1$, there are mn points for each of the types (i) and (ii), and $m(m-1)n$ points of the type (iii), for each $i = 1, 2$, thus giving a total of mn(m+2) design points. The masses assigned to the support points, therefore, satisfy $mn(v_0+v_1+v_2)+C(m, 2)n(w_1+w_2) = 1$, where $C(s, t) = \binom{s}{t}$.

Thus far, we have indicated the nature of invariant mixture design.

Remark 3.1 It may be observed that in the above analysis the number of support points has far exceeded the number of parameters in the model. When we specialize to the case of $m = 2$ and $n = 3$, we will take up this issue and propose further reduction in the choice of the support points. Note that for the same problem, that is to reduce the number of support points, Lewis et al. (2010), for the case of $m=2$ and $n = 3$, used the exchange algorithm. However, the optimum design they derived lack the invariance property though the criterion function used, namely D-optimality criterion, is invariant with respect to the minor components within each of M_1 and M_2. This contradicts the well-established fact that *in an invariant optimality design problem, the support points of the optimum design must be invariant with respect to its components* (cf. Pukelsheim 2006, pp. 331–351, Chap. 13). Further, they have assigned equal masses to the design points.

Remark 3.2 For the sake of completeness, we cite the example considered by Lewis et al. (2010) in which it is assumed that $\delta_1 = 0.1$ and $\delta_2 = 0.5$. They used the exchange algorithm to choose a 9-point design with equal masses which they claim to be the D-optimal design. The 9 points are:

$$\begin{array}{lll} (0.1,\ 0;\ 0.9,\ 0,\ 0), & (0,\ 0.5;\ 0.5,\ 0\,0), & (0.5,\ 0;\ 0,\ 0.5,\ 0), \\ (0,\ 0.5;\ 0,\ 0.5,\ 0), & (0,\ 0.1;\ 0,\ 0,\ 0.9), & (0.5,\ 0;\ 0,\ 0,\ 0.5), \\ (0.25,\ 0.25;\ 0.5,\ 0,\ 0), & (0.25,\ 0.25;\ 0,\ 0,\ 0.5), & (0.05,\ 0.05;\ 0,\ 0.9,\ 0), \end{array}$$

which, as is pointed out earlier, indicates lack of invariance among the components, and determinant of the information matrix is 1.50366×10^{-09}.

Stage 2: Below we propose the following way to reduce the number of support points, while retaining the property of invariance. Soon we specialize only to the case of $m = 2$ and $n = 3$. We consider a class of designs $\mathcal{D}_2$ having the following support points:

(a) all the pure support points in (i),
(b) mixed support points of the form $(\delta/2, \delta/2, 0, \ldots, 0; 1-\delta, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates, for some $\delta \in [\delta_1, \delta_2]$, each with mass w.

The number of support points of a design in $\mathcal{D}_2$ is thus $2mn + C(m, 2)n$.

Remark 3.3 For $m=2$, $n = 3$, $\delta_1 = 0.1$, $\delta_2 = 0.5$, and equal masses for all design points, our calculation shows that the determinant of the information matrix is maximum (4.32709×10^{-09}) for a 15-point design with $\delta = \delta_2$. A comparison of this design with that having 18 support points given by (i) and (iii) with equal masses shows this design to be better as the determinant of the 18-point design is 3.91969×10^{-09}. However, this design may not be a D-optimal design since the number of design points is more than the number of parameters to be estimated, and in that case, the optimum masses allocated to the design points may not be necessarily equal.

We, therefore, proceed to find the D-optimal design within $\mathcal{D}_2$. Specifically, our choice centers around δ and the masses.

For $m=2$, $n = 3$, and any design $\xi \in \mathcal{D}_2$, the information matrix is given by

$$M(\xi) = \begin{bmatrix} 3(\delta_1^2 v_1 + \delta_2^2 v_2) I_2 + \frac{3\delta^2}{4} w \mathbf{1}_2 \mathbf{1}_2' & \frac{3\delta^3}{8} w \mathbf{1}_2 & 0 \\ \frac{3\delta^3}{8} w \mathbf{1}_2' & \frac{3\delta^4}{16} w & 0 \\ 0 & 0 & (2 \sum_{i=1}^{2} (1-\delta_i)^2 v_i + (1-\delta)^2 w) I_3 \end{bmatrix}.$$

Hence,

$$\text{Det.}\,[M(\xi)] = \frac{27}{16}\delta^4 w(\delta_1^2 v_1 + \delta_2^2 v_2)^2[2(1-\delta_1)^2 v_1 + 2(1-\delta_2)^2 v_2 + (1-\delta)^2 w]^3. \tag{2.7}$$

The D-optimal design within $\mathcal{D}_2$ is obtained by determining the optimal values of δ, v_1, v_2 and w that maximize (2.7), subject to $6(v_1 + v_2) + 3w = 1$.

Now, to maintain invariance among the minor components within the major components M_1 and M_2, the minimum number of design points can be $r = mn + C(m, 2)n$. We, therefore, next consider the subclass $\mathcal{D}_3$ of r-point designs with support points as follows:

(a) $(\delta_3, 0, 0, \ldots, 0; 1-\delta_3, 0, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates each with mass v;
(b) $(\delta_4/2, \delta_4/2, 0, \ldots, 0; 1-\delta_4, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates, each with mass w, where $\delta_1 \le \delta_3, \delta_4 \le \delta_2$.

The information matrix of any design ξ with the above support points is given by

$$M(\xi) = \begin{bmatrix} n\delta_3^2 v I_m + \frac{n\delta_4^2 w}{4} 1_m 1_m' & \frac{n\delta_4^3 w}{8} Q & 0 \\ \frac{n\delta_4^3 w}{8} Q' & \frac{n\delta_4^4 w}{16} I_{C(m,2)} & 0 \\ 0 & 0 & \left\{m(1-\delta_3)^2 + C(m,2)(1-\delta_4)^2\right\} I_n \end{bmatrix},$$

$$\text{where } Q^{m\times C(m,2)} = \begin{bmatrix} \overbrace{1\,1\ldots1}^{m-1} & \overbrace{0\,0\ldots0}^{m-2} & \overbrace{0\,0\ldots0}^{m-3} & \ldots & \overbrace{0}^{1} \\ 1\,0\ldots0 & 1\,1\ldots1 & 0\,0\ldots0 & \ldots & 0 \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots & \ldots \\ 0\,0\ldots1 & 0\,0\ldots1 & 0\,0\ldots1 & \ldots & 1 \end{bmatrix}.$$

Hence,

$$\text{Det.}\,M(\xi) = \left\{m(1-\delta_3)^2 v + C(m,2)(1-\delta_4)^2 w\right\}^n \left(\frac{n\delta_4^4 w}{16}\right)^{C(m,2)} |M_{11}| \times |I_{C(m,2)} - \frac{n\delta_4^2 w}{4} Q' M_{11}^{-1} Q|, \tag{3.3}$$

where $M_{11} = n\delta_3^2 v I_m + \frac{n\delta_4^2 w}{4} 1_m 1_m'$.

In particular, for $m=2$, $n=3$, we have $r=9$ and

$$\text{Det.}M(\xi) = \frac{27}{16}\delta_3^4 \delta_4^4 v^2 w[2(1-\delta_3)^2 v + (1-\delta_4)^2 w]^3, \tag{3.4}$$

where $w = 1/3 - 2v,\ v, w > 0$.

For $\delta_3 = \delta_4 = \delta,\ \delta_1 \leq \delta \leq \delta_2$, we get

$$\text{Det.}M(\xi) = \frac{1}{16}\delta^8(1-\delta)^6 v^2 w,$$

which is maximized for $\delta = 4/7$ and $v = w = 1/9$.

Since, for $\delta_3 = \delta_4 = \delta,\ \delta_1 \leq \delta \leq \delta_2$, Det. $M(\xi)$ is a concave function of δ, the optimal value of δ is given by

(a) $\delta = 4/7$ if $\delta_1 \leq 4/7 \leq \delta_2$,
(b) $\delta = \delta_1$ if $\delta_1 \geq 4/7$,
(c) $\delta = \delta_2$ if $\delta_2 \leq 4/7$.

Numerical computation is carried out for $m = 2, n = 3$. Table 1 gives the D-optimal designs within $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ for some combinations of (δ_1, δ_2) when $m = 2, n = 3$.

With this, finally, we have been in a position to settle the claim made in the statement of Theorem 3.1.

3.2 *Study of Optimal Mixture Design for Model (1.2)*

In Sect. 1, we have introduced model (1.2). A general version of this model is taken up below:

$$\left.\begin{aligned} \eta_x &= \sum_{i=1}^{m} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}}^{\boldsymbol{m}} \beta_{ij} x_i x_j, \\ \beta_i &= \sum_{k=m+1}^{m+n} \theta_{ik} x_k,\ i = 1(1)m,\ \ \beta_{ij} = \sum_{k=m+1}^{m+n} \theta_{ijk} x_k,\ i < j = 1(1)m. \end{aligned}\right\} \tag{3.5}$$

Thus, the mean response function can be written as

$$\eta_x = \sum_{i=1}^{m} \sum_{k=m+1}^{m+n} \theta_{ik} x_i x_k + \sum_{i<j=1}^{m} \sum_{k=m+1}^{m+n} \theta_{ijk} x_i x_j x_k, \tag{3.6}$$

The experimental region is, as before, given by $\Xi = \Xi_1 \cap \Xi_2$.

Because of invariance among the components in $x_{(1)}$ and $x_{(2)}$, we once again start with a class $\mathcal{D}_1$ of designs ξ defined in Step 1 of Theorem 3.1.

The objective of the study below is to confine to a subclass $\mathcal{D}_2^*$ [of $\mathcal{D}_1$] of invariant mixture designs and characterize the nature of D-optimal designs for the estimation of the parameters with reference to the model (3.6).

Table 1 D-optimal designs within D_2, D_3, and D_4 for some combinations of (δ_1, δ_2) when $m = 2$, $n = 3$

(δ_1, δ_2)	Subclass	δ_1/δ_3	δ_2/δ_4	δ	v_1	v_2	v	w	Det.
(0.1, 0.4)	$\mathcal{D}_2$	0.1	0.4	0.4	0.02774	0.0116	–	0.0459	1.7308×10^{-8}
	$\mathcal{D}_3$	0.4	0.4	–	–	–	1/9	1/9	1.6384×10^{-8}
	$\mathcal{D}_4$	–	–	0.4	–	–	1/9	1/9	1.6384×10^{-8}
(0.1, 0.5)	$\mathcal{D}_2$	0.1	0.5	0.5	0.0398	0.0876	–	0.0788	3.4684×10^{-8}
	$\mathcal{D}_3$	0.5	0.5	–	–	–	1/9	1/9	2.0931×10^{-8}
	$\mathcal{D}_4$	–	–	0.5	–	–	1/9	1/9	2.0931×10^{-8}
(0.2, 0.6)	$\mathcal{D}_2$	0.2	0.6	0.6	0.07186	0.06288		0.06418	4.6625×10^{-8}
	$\mathcal{D}_3$	0.2896	0.6	–	–	–	0.0702	0.1931	3.0188×10^{-8}
	$\mathcal{D}_4$	–	–	4/7	–	–	1/9	1/9	1.8606×10^{-8}
(0.3, 0.7)	$\mathcal{D}_2$	0.3	0.7	0.6872	0.0644	0.0639	–	0.0770	4.6903×10^{-8}
	$\mathcal{D}_3$	0.7	0.361/6	–	–	–	0.0702	0.1931	4.2936×10^{-8}
	$\mathcal{D}_4$	–	–	4/7	–	–	1/9	1/9	1.8606×10^{-8}
(0.6, 0.8)	$\mathcal{D}_2$	0.6	0.8	0.6	1/9	0	–	1/9	1.6481×10^{-8}
	$\mathcal{D}_3$	0.6	0.6	–	–	–	1/9	1/9	1.6481×10^{-8}
	$\mathcal{D}_4$	–	–	0.6	–	–	1/9	1/9	1.6481×10^{-8}

The information matrix of design $\xi \in \mathcal{D}_1$ is given by

$$M(\xi) = \begin{bmatrix} \sum_{i=0}^{2} \delta_i^2(1-\delta_i)^2 v_i I_{mn} + \sum_{i=1}^{2} \frac{\delta_i^2(1-\delta_i)^2}{4} w_i J_{mn} & \sum_{i=1}^{2} \frac{\delta_i^3(1-\delta_i)^2}{8} w_i P \\ \sum_{i=1}^{2} \frac{\delta_i^3(1-\delta_i)^2}{8} w_i P' & \sum_{i=1}^{2} \frac{\delta_i^4(1-\delta_i)^2}{16} w_i I_{C(m,2)n} \end{bmatrix}, \quad (3.7)$$

where

$$P^{mn\times C(m,2)n} = Diag(\underbrace{Q, Q, \ldots, Q}_{n}),$$

$$Q^{m\times C(m,2)} = \begin{bmatrix} \overbrace{1\,1\ldots1}^{m-1} & \overbrace{0\,0\ldots0}^{m-2} & \overbrace{0\,0\ldots0}^{m-3} & \ldots & \overbrace{0}^{1} \\ 1\,0\ldots0 & 1\,1\ldots1 & 0\,0\ldots0 & \ldots & 0 \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots & \ldots \\ 0\,0\ldots1 & 0\,0\ldots1 & 0\,0\ldots1 & \ldots & 1 \end{bmatrix}. \quad (3.8)$$

For $m = 2, n = 3$,

$$\text{Det.}[M(\xi)] = (a+b)^3[d(a-b) - 2c^2]^3,$$

where

$$a = \sum_{i=0}^{2} \delta_i^2(1-\delta_i)^2 v_i + b,\ b = \sum_{i=1}^{2} \frac{\delta_i^2(1-\delta_i)^2}{4} w_i,$$

$$c = \sum_{i=1}^{2} \frac{\delta_i^3(1-\delta_i)^2}{8} w_i, \qquad d = \sum_{i=1}^{2} \frac{\delta_i^4(1-\delta_i)^2}{16} w_i.$$

As the number of support points is very large compared to the number of parameters to be estimated, we, as before, attempt to reduce it by considering the class $\mathcal{D}_2^*$ of designs with all support points of (i) and (ii) and mixed support points of the type $(\delta/2, \delta/2, 0, \ldots, 0; 1-\delta, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates, for some $\delta \in [\delta_1, \delta_2]$, each with mass w. The information matrix for any $\xi \in \mathcal{D}_2^*$ is given by

$$M(\xi_1) = \begin{bmatrix} \left(\sum_i \delta_i^2(1-\delta_i)^2 v_i\right) I_{mn} + \frac{\delta^2(1-\delta)^2}{4} w P P' & \frac{\delta^3(1-\delta)^2}{8} w P \\ \frac{\delta^3(1-\delta)^2}{8} w P' & \frac{\delta^4(1-\delta)^2}{16} w I_{C(m,2)n} \end{bmatrix},$$

where P is defined in (3.8).

Then,

$$|M(\xi_1)| = \left(\frac{\delta^4(1-\delta)^2}{16}w\right)^{C(m,2)n}\left(\sum_{i=0}^{2}\delta_i^2(1-\delta_i)^2 v_i\right)^{mn}, \quad \delta \in [\delta_1, \delta_2].$$

For given δ_i, v_i, $i=1, 2$ and w, Det. $[M(\xi_1)]$ is a concave function of δ with maximum at $\delta=2/3$. Hence, apart from the pure type design points, the experiment must be conducted at the points $(\delta/2, \delta/2, 0, \ldots, 0; 1-\delta, 0, \ldots, 0)$ and all permutations within the first m coordinates and within the last n coordinates, where

(a) $\delta=2/3$ if $\delta_1 \le 2/3 \le \delta_2$,
(b) $\delta=\delta_1$ if $\delta_1 \ge 2/3$,
(c) $\delta=\delta_2$ if $\delta_2 \le 2/3$.

It is important to note that the choice of δ is independent of m and n.

Remark 3.4 Another way suggested to reduce the number of support points is to consider the class of ***saturated*** designs $\mathcal{D}_3^*$. Let $\delta \in [\delta_1, \delta_2]$ and confine to designs with support points (i) $(\delta, 0, \ldots, 0; 1-\delta, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates each with mass v; (ii) $(\delta/2, \delta/2, 0, \ldots, 0; 1-\delta_i, 0, \ldots, 0)$ and all possible permutations within the first m coordinates and within the last n coordinates, each with mass w, such that $mnv + C(m, 2)nw = 1$. In this case, the information matrix of any design ξ_2 is given by

$$M(\xi_2) = \begin{bmatrix} \delta^2(1-\delta)^2 v I_{mn} + \frac{\delta^2(1-\delta)^2}{4} w P P' & \frac{\delta^3(1-\delta)^2}{8} w P \\ \frac{\delta^3(1-\delta)^2}{8} w P' & \frac{\delta^4(1-\delta)^2}{16} w I_{C(m,2)n} \end{bmatrix}$$

so that

$$\begin{aligned} \text{Det.}\,[M(\xi_2)] &= \left[\frac{\delta^4(1-\delta)^2}{16}w\right]^{C(m,2)n} [\delta^2(1-\delta)^2 v]^{mn} \\ &= \delta^{2(m+2C(m,2))n}(1-\delta)^{2(m+C(m,2))n}\frac{v^{mn}w^{C(m,2)n}}{16^{C(m,2)n}}. \end{aligned}$$

For given v and w, Det.$[M(\xi_2)]$ is a concave function of δ with maximum value at $\delta = \frac{2m}{3m+1}$. Thus, the experiment should be conducted with

(a) $\delta = \frac{2m}{3m+1}$ if $\delta_1 \le \frac{2m}{3m+1} \le \delta_2$,
(b) $\delta = \delta_1$ if $\delta_1 \ge \frac{2m}{3m+1}$,
(c) $\delta = \delta_2$ if $\delta_2 \le \frac{2m}{3m+1}$.

It is noteworthy that the optimum choice of δ is dependent on m but independent of n.

Table 2 D-optimal designs within D_1, D_2^* and D_3^* for some combinations of (δ_1, δ_2) when $m = 2$, $n = 3$

(δ_1, δ_2)		δ_0	δ_1	δ_2	δ	v_0	v_1	v_2	w_1	w_2	v	w	Det.
(0.1, 0.5)	$\mathcal{D}_1$	0.3	0.1	0.5		0.034	0.015	0.053	0.049	0.080			1.08119×10^{-12}
	$\mathcal{D}_2^*$				0.5		0.055	0.055				0.088	1.36482×10^{-12}
	$\mathcal{D}_3^*$	–		–	0.5		–	–			1/9	1/9	4.27219×10^{-12}
(0.3, 0.7)	$\mathcal{D}_1$	0.5	0.3	0.7		0.035	0.055	0.040	0.0012	0.072			2.29991×10^{-12}
	$\mathcal{D}_2^*$		0.3	0.7	2/3		0.055	0.055	–			0.088	3.78612×10^{-12}
	$\mathcal{D}_3^*$	–		–	4/7		–	–			1/9	1/9	8.38711×10^{-12}
(0.6, 0.8)	$\mathcal{D}_1$	0.7	0.6	0.8		0.019	0.039	0.086	0.001	0.044			2.27387×10^{-13}
	$\mathcal{D}_2^*$		0.6	0.8	2/3		0.055	0.055	–			0.088	3.78229×10^{-12}
	$\mathcal{D}_3^*$			–	0.6		–	–			1/9	1/9	8.65667×10^{-12}

A numerical comparison of the D-optimal designs within the three subclasses have been carried out with $m=2$ and $n=2, 3$, and a number of combinations of (δ_1, δ_2), and it is shown in Table 2.

Remark 3.5 It is observed that the D-optimal designs within $\mathcal{D}_3^*$ perform better than those within $\mathcal{D}_1$ and $\mathcal{D}_2^*$.

4 Conclusion

This study is along the lines of Lewis et al (2010) involving mixture designs with major and minor components in the underlying mixture models subject to natural relational constraints. We have endeavored to exploit symmetry and invariance of the model parameters toward identification of D-optimal designs within suitably defined subclasses of admissible mixture designs. Our treatment is based on the layouts of pharmaceutical experiments, though naturally agricultural experiments also provide equally acceptable platforms for applications of these ideas and results.

Acknowledgements The authors thank the anonymous referee for a number of fruitful suggestions, which improved the presentation of our ideas in the paper.

References

1. Lewis, G. A., Mathieu, D., & Phan-Tan-Luu, R. (2010). *Pharmaceutical experimental design*. London, UK: Informa Healthcare.
2. Pal, M., & Mandal, N. K. (2013). Optimum designs for mixtures with relational constraints on the components. *International Journal of Experimental Design and Process Optimization, 3*(3), 276–293.
3. Pukelsheim, F. (2006). *Optimal design of experiments* (Classics in Applied Mathematics). SIAM.
4. Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society B, 20*, 344–360.
5. Sinha, B. K., Mandal, N. K., Pal, M., & Das, P. (2014). *Optimal mixture experiments.* Lecture Notes in Statistics 1028. Springer.

Growth Curve of Socio-economic Development in North-Eastern Tribes

Ratan Dasgupta

Abstract Socio-economic status of north-eastern tribes in India has relevance to formulate developmental policies for the region. The present investigation conducted on north-eastern (NE) tribes via questionnaires and descriptive interviews in a socio-economic survey is related to personal and social events over a long span of time interval focused on NE tribal lifestyle. Investigation by retrospective longitudinal study is of interest in lifestyle and socio-economic assessment for individuals when relevant data on past are not recorded. For tribal individuals selected by stratified random sampling from Tripura, India, we record descriptive-type and question-answer-type response on their past and present standard of living. For each individual and for variables related to development, viz. income, hygiene, food, shelter, clothing, entertainment, education, medical facilities, social environment and human resources, over a time period around a specific time of past/present, we consider measurement of each of these variables in a Likert-type scale to obtain consolidated status scores for comparison over time. Scores on status are then analysed for detecting possible trend in living standard of tribal individuals with progress of time. Over a time range spanning for about last 60 years from 2013, the events reported in this study reveal an upward ascending movement of social change from past to present in terms of increasing composite status score. Growth curves of scores obtained by nonparametric regressions indicate general improvement of tribal welfare status with progress of time. The curves show upward trend in growth, which is faster towards the end. Proliferation rate of scores suggests that much of the changes in a scaled speed of improvement in status are relatively recent phenomena at the year 2013. Composite status scores are based on discrete scores of individual components. Under appropriate assumptions, characterization theorems are obtained from solution of Cauchy equations for discrete random variables relevant in modelling status scores. These include discrete version of bi-exponential and normal distribution, extending the results of Dasgupta (Theory of Probability and Its Applications, 38(3):520–524, 1993).

R. Dasgupta (✉)
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203 B T Road, Kolkata 700108, India
e-mail: rdgupta@isical.ac.in; ratandasgupta@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_4

Keywords Lifestyle status · Likert scale · Jhum cultivation · Proliferation rate Manhattan distance · Cauchy equation

MS Subject Classification 62P25 · 60G20

1 Introduction, Methodology and the Data

Socio-economic status of the tribal individuals is of interest in decision of policy making in north-eastern states of India. Efforts towards economic development are of importance in backward regions. Removing impediments to lasting peace and security is a concern towards stability in the region. The whole of the north-east has been categorized as industrially backward, e.g. see http://planningcommission.nic.in/reports/publications/pub_neregion.pdf.

In order to assess the change in development status of north-eastern tribes over time, individuals are selected by stratified random sampling for detailed interviews via questionnaires, and the responses are initially intended to be quantified in a continuous scale of 1–2. Score 1 represents the poor state, this is the lowest rung of socio-economic condition reported in the study, and 2 represents a comfortable standard of living, the affluent social status reported in the interviews. Quantified individual status is then associated with specific time reference. The status score at intermediate time points are linearly interpolated. Some of the goals in present study are to evaluate the direction of change in status score and obtain rate of change in score over time. Basic issues on socio-economic front requiring appropriate attention for improvement are to be identified, and introduced changes to be monitored.

Stratified random sampling is taken recourse to for selection of villages, and tribals are interviewed in selected villages at the next stage. This sampling procedure has the advantage of penetrating into the population. Each of the selected villages is studied in detail, as far as possible, on tribal individual's socio-economic status; taking into account of accessibility, willingness of tribal villagers to participate in the interviews and limitations of conducting such a survey within some of the hilly regions disturbed due to insurgency.

To protect confidentiality of the respondents, some auxiliary information, which are not of much relevance for the purpose of conducted survey, is usually suppressed or altered in order to hide the identity of individuals in the presented reports, while pertinent information is kept intact for analysis. For a discussion on confidentiality of respondents, see, e.g. Kaiser (2009).

Scores are assigned to answers of questions on different components of development, descriptive part of tribals' opinion is quantified; narrations on socio-economic environment from tribals are recorded in detail. Details of past and present events narrated by tribals during interviews reflect different aspects of the lifestyle scenario. The overall impression is likely to be devoid of personal bias when opinions are aggregated over many individuals. Some of the individuals are aged senior members

of the villages covered in the study, having sharp memory to recall past events. Parts of their statements recorded in the interviews are as follows.

After visiting a selected village near capital Agartala of Tripura on 26/08/2012 and after interviewing some of the local persons therein, it has been found that most of the people prefer nucleus family, they are still living in mud houses, they drink water from well, there are no sanitary latrines, and most of the people are illiterate. One of the problems in the localities is exploitation by money lenders; after lending money to the villagers they charge a huge interest on it, at least 10–15% *per month.* The other major problem is excessive consumption of liquor; this is spoiling health and money of the young people. Many villagers consume country liquor excessively, although limited consumption is common during regular meals, social marriage and in festive celebrations under local tradition. Affected persons IQ are moderate, resulting from a dull state of mind. Proximity to city has made some of the villagers wise and intelligent. There are facilities of ration card. Joint families are breaking down into nucleus family. Ignorance on cause of diseases is prevalent. Villagers like to consume goats' milk, as they believe that goats eat leaves from different plants, which to their belief, turn the milk full of medicinal properties. Without knowing the cause of a disease and in absence of proper medication, they often use intravenous saline injection that may turn fatal in bacterial diseases.

On 29/3/2013, tribal persons interviewed in a village near Agartala reported that even in the year 1952, wild animals including tigers were present in the locality. They used to sell forest resources and products including abundant *muli* bamboo for their livelihood. The then King of Tripura used to come on Elephant to have boat rides on nearby lake. Their social and economical position was better as of now. King provided them necessary attention.

Another interview was taken on 19/9/2011 in a village of North Tripura, within a tea garden. The 55-year-old individual Notor Singh Tripura (name changed) was interviewed, and he informed that their living condition is much better now as compared to what it was 10 years back.

A tribal person, Kabi charan Debvarma (name changed) interviewed on 29/3/2013 was about 85 years old. He narrated the past situation of livelihood and compared that with present time. He now stays in a nucleus family of four members: himself, his wife, his son and his daughter-in-law. He had informal learning from *Balyasiksha*, taught by a village schoolteacher. He narrated that they were ignored by the upper caste individuals of a particular religion, they had to stay in dark at night since there was no electricity; they were very poor, Jhum cultivation was prevalent and to clear land for sowing seeds, they used to put fire in the forest occasionally; which in turn damaged the nearby houses. Paddy, kapas, til, bajra, bamboo shoot, rubber plantation, pineapple, etc., are now grown in place of natural forest. Wild pig, cock, porcupine and rabbits were abundant in forest. Before 1960, people use to hunt deer from forest when these used to come to eat wild gooseberry. Deer meat may still be available discreetly around the area of *Dumboor* falls. Attacks from wild animals including tigers were common in past. Partially injured tigers were killed by villagers, as these are more harmful than healthy one. For this reason, the villagers domesticated dogs to guard their cows. Milk vending, handmade cloth, vegetables, forest products were

the sources of income. Food qualities of many villagers were poor; they used to consume *pesta* yam, wild yam and bamboo shoot from forest in the past. At present their economic condition is good as compared to the past. Many changes occurred which were not there in the past, like use of mobile phones which came in the year 1998, electricity, TV are common now. The place has really changed with new buildings and good transportation. Some villagers now have silk cocoon nursery and have set up a business of selling this silk products. They try to help each other now when in need of money, instead of approaching a money lender.

Earthen pots, etc., were used earlier for cooking, ashes from burned bamboo and banana plants were used as soap substitutes, as these were rich in potash. Now, two square meals are available irrespective of rich/poor demarcation of the past. Rega scheme of employment is also available. Tube wells are available for drinking water. Mosquitoes are present everywhere; as a result malaria disease is also prevalent in the locality. Very few peoples use mosquito nets. Previously, forest land was used for toilet, now brick made sanitary toilets are available. People who were rich in the past are not so rich now in their locality, and the people who were poor in the past have a comfortable standard of living now.

In another village, located on the border of Tripura with Mizoram state, far away from the capital Agartala of Tripura, a group of families were under study on 30/8/2013. There are two persons having a degree of M.A., one having a degree of B.A., and five persons have passed higher secondary examination; the rest ten have just passed secondary level of examination in the village at present. There are inclinations towards tribal cultural activities. Near the village, there are two senior basic and three junior basic schools. In total there are only two joint families. Ration card facility is available. Rega scheme of employment is present with 100 days of work in a year. About 90% of the farmers are engaged in Jhum cultivation that is common among tribals, and the remaining 10% of the farmers are engaged in cultivating potatoes, vegetables, radish, beans in traditional farming, the vegetables are then supplied to nearby town. Fisheries are abundant in tribal areas with financial assistance made available from the government. Rubber and tea plantation on hilly areas have recently developed in which a large number of families are involved, associated with this is a high potential of long-term employment for human resources. Electricity is unavailable for around one and half hours every night. Ring well water is consumed by many families after filtration, and in summer boiled water is consumed. Use of ring well water is not popular among a particular tribe, they mainly use cascade water, water in ring well is spoiled by dumping garbage in it by them, and this is due to lack of proper health education among that tribe.

A major problem in the village is excessive consumption of country liquor; this habit is ruining life of the young. There is a sort of dowry demand from bride's side (in labour and material from the groom). Financial support from government is partly wasted in liquor expenses. Local persons from a particular tribe are staying in rehabilitation camps and misusing the government support. They are staying in the camps on the pretext of the rebel insurgency. Although this trouble has stopped, still some individuals are not interested to return to their own houses. Tribal exploitation

still exists. The extremist problem has stopped about 30–35 years back in the region of 350 families.

Lack of competent teachers is a problem at junior basic school and high school levels. One teacher may sometime have to take many classes for I-V standard and attend to office work as well, due to lack of teachers and office staff in school. Teachers at higher level indicate that weakness of students at ground level thus incurred at the basic study is difficult to rectify at the higher level of education. The villagers had to pay protection money to rebels for years in the past. Many years back, extremists killed a reverend of a missionary church by mistake, instead of their target, head of a second-hand garment business. After that the extremists stopped their activities in this region altogether, following instructions from top rebel leaders.

The above are some of the descriptive responses elicited from tribal villagers selected to participate in the survey, some of the villagers are aged. but have alert state of mind with sharp memory to describe the past events vividly. The sectors that require appropriate attention include proper education at primary level, health, and curb in excessive liquor consumption.

The responses from tribal individuals are converted to a score. Likert scale is commonly used to grade psychometric response. For collective responses to a set of items, the responses are scored along a range. Information obtained from descriptive response and answers to questionnaires during the survey conducted over the time period 19/9/2011–30/8/2013 are transformed to Likert scores on tribal status for analysis.

The referential period of the presented analysis covers a broad range of about 60 years, the described events start from the year 1952 mentioned by respondents in interviews, and the period extends up to 30/08/2013; the last day of interview conducted in the survey. The origin, day 0 is taken at 1/1/1952, and the last date of time point in the analysis is 30/08/2013.

In Sect. 2, we analyse the collected data and indicate possibilities of further study. The results are presented in Sect. 3. Issues requiring attention for improvement of tribal status are identified. Results from analysed data are discussed in Sect. 4. The interview scores are recorded in discrete scale. Some characterization theorems relevant for tribal status score evaluation, involving discrete and continuous random variables, are presented in the Appendix.

2 Analysis of Data and Some Comments

Scores on social status obtained are based on answers on questionnaires and descriptive interviews conducted over a period of time, sampled all over Tripura, only in accessible regions.

Median of scores on social attributes in sample is a robust estimator compared to mean as a measure of central tendency, especially for ordinal variables. For each individual and for variables related to development, like status on income, hygiene, food, shelter, clothing, entertainment, education, medical facilities, social environment etc.,

over a narrow cluster of time points around a specific time, we conveniently consider measurement of each of these in a Likert-type scale [1.0, 1.1, 1.2, $\cdots$, 1.9, 2.0], and then we take the median score as a robust measure of central tendency. For ordinal variables, sample median or Hodges–Lehmann estimator under the assumption of symmetry may be appropriate measures for central tendency. The later estimate has a lower breakdown point 29% compared to 50% of the sample median that is asymptotically normal around population median m with variance $[4nf(m)]^2$, n being the sample size and f is the density function of the corresponding continuous counterpart of the variable that is discretized.

Hodges–Lehmann estimator is a U statistics, and Bahadur–Kiefer representation of quantiles for one-sided convergence is possible from a result of Dasgupta (2015a), under the Lipschitz condition (3.8) assumed therein on the distribution function F; one-sided convergence is useful to have almost sure upper and lower bounds of growth curve. This part of investigation on tribal status score will be undertaken in a subsequent study.

For each fixed individual, the average score, viz. simple/weighted mean of the above-mentioned median or Hodges–Lehmann scores on different characteristics over that small time region, may be assigned to the mean of the small time region specific for that individual. The average score of variables, i.e. averaged median of component responses for an individual, is considered to represent the status of tribal welfare for that individual at that time. Social welfare status at a distant past is usually reported in a way that is conducive in assigning a consolidated score in the range [1, 2] by an experienced interviewer. In some occasions, the interviewees like to assign a consolidated score of the status and/or help the interviewer to arrive at a composite score on status. Variation of the scores over time is of interest in longitudinal growth of tribal development.

Likert scale may be composed of a series of four or more Likert-type items that represent similar questions on social development combined into a single composite score for variables. Likert scale data may be analysed like interval-type data with normal distribution theory applicable, especially when the histogram for sum of scores resembles histogram of a normal distribution.

We obtain a number of consolidated characteristic scores for variables under study, the consolidated scores on each variable here are based on median over a narrow range of time around a fixed time point; median being a robust estimate is not much affected by outliers in that narrow range of time. Mean of these robust (median) scores over different variables assigned to a fixed time represents overall social status at individual level. To explain longitudinal variation, the mean/weighted mean scores for an individual over time are joined by straight lines, when repeated observations are available for an individual in a time range. Next, mean of y values in the graphs representing scores for fixed time in X-axis over different persons provides overall growth variations in a basic form on the entire range of time covered in different interviews, as time points in X-axis varies. Smooth response curves obtained by Fourier smoothing, lowess and spline regression on these mean values of development status over selected time points provide different types of growth curves on tribal status.

Table 1 Lowess, spline and Fourier values of status score of tribal individuals

Day	Mean score	Lowess value	Spline value	Fourier value
0	2.000000	1.409731	1.499543	–
300	1.493443	1.406039	1.485970	1.4934
4140	1.280287	1.370413	1.316346	1.3550
5552	1.291261	1.360211	1.300448	1.3725
9511	1.384557	1.348864	1.384270	1.3468
14700	1.413211	1.396730	1.389821	1.3801
14911	1.364598	1.395798	1.391872	1.3957
17431	1.446921	1.444310	1.411269	1.4294
17810	1.369203	1.454061	1.439211	1.4990
19080	1.553118	1.538547	1.546857	1.5533
21410	1.761346	1.689477	1.761568	1.6226
21747	1.635979	1.705208	1.693431	1.6722
21752	1.793234	1.705427	1.690195	1.7349
21900	1.617391	1.711920	1.594408	1.7591
21960	1.866667	1.714553	1.555575	1.8667
22111	1.400000	1.720858	1.400000	–

The mean score, lowess, spline and Fourier smoothed values of the scores over different time points are given in Table 1. From the mean scores, growth curves are drawn in different figures, corresponding to different nonparametric techniques. Figure legends explain the inherent features of the growth data. These features are represented in different curves over a time span of more than last 60 years, back from the year 2013.

3 Results

The analysis of these data obtained indicates that there is a general trend of socio-economic development on tribal welfare in the last 60 years from 2013. Collected data is represented in Fig. 1. Growth curves obtained from nonparametric lowess smoothing are shown in Fig. 2, and proliferation rates are obtained in Figs. 3 and 4. These confirm improvement of tribal status score over time. In Fig. 5, spline regression for growth curve of status scores is shown, and subsequent proliferation rates are shown in Figs. 6, 7. These reveal similar conclusions of gradual improvement in tribal welfare over time. Proliferation rate is a scaled version of speed of growth, and this is a dimensionless measure; variation of this over time shown in Figs. 3, 4, 6 and 7 suggest that much of the change in tribal welfare status is a relatively recent phenomena. Fourier smoothed growth curve computed by MATLAB is shown in Fig. 8. Figure 8 excludes the first and last time points, as the associated observations

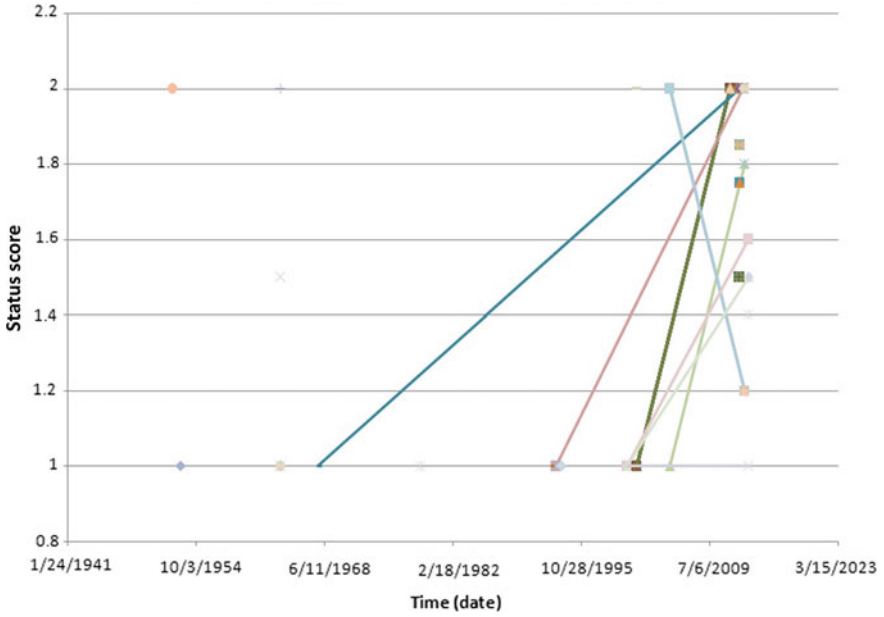

Fig. 1 Growth curve of status for 77 tribal individuals

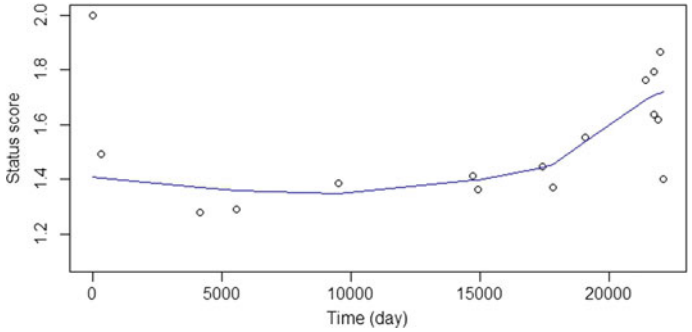

Fig. 2 Lowess growth curve of status

are detected as outliers in Fourier smoothing by MATLAB. The rise in growth is sharp towards the end in Fig. 8. Proliferation rates obtained from Fourier smoothed values are shown in Figs. 9, 10. The features suggest that the status improvements are relatively recent. Characterization results are obtained for bi-exponential and normal distribution in discrete and continuous cases. These are relevant in modelling variables in psychometry, among others. The survey reveals that there are scopes of improvement for tribal welfare in several sectors. For betterment of lifestyle status, attention on health, hygiene and primary education is urgently needed in some regions of Tripura.

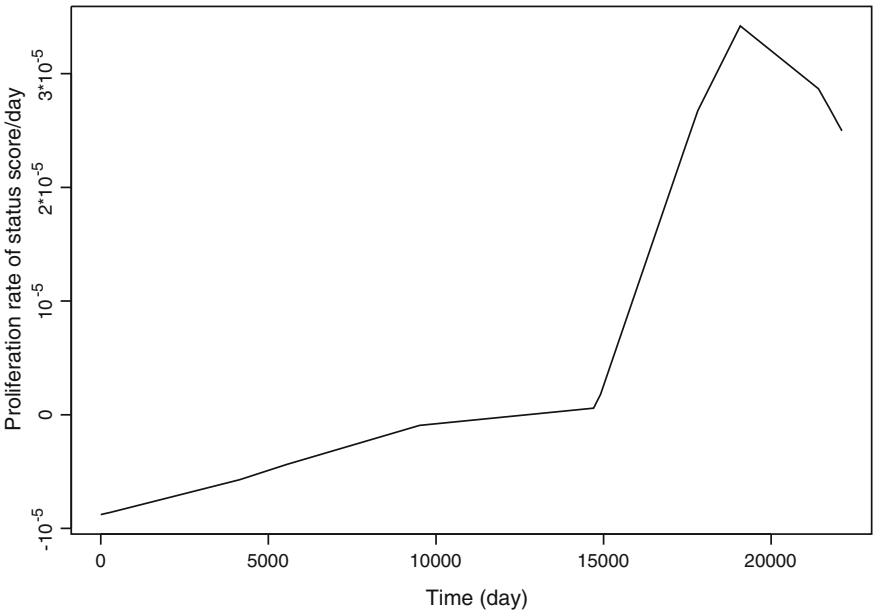

Fig. 3 Proliferation rate of status score from descriptive interview: wt.exp(−0.001 x); spline

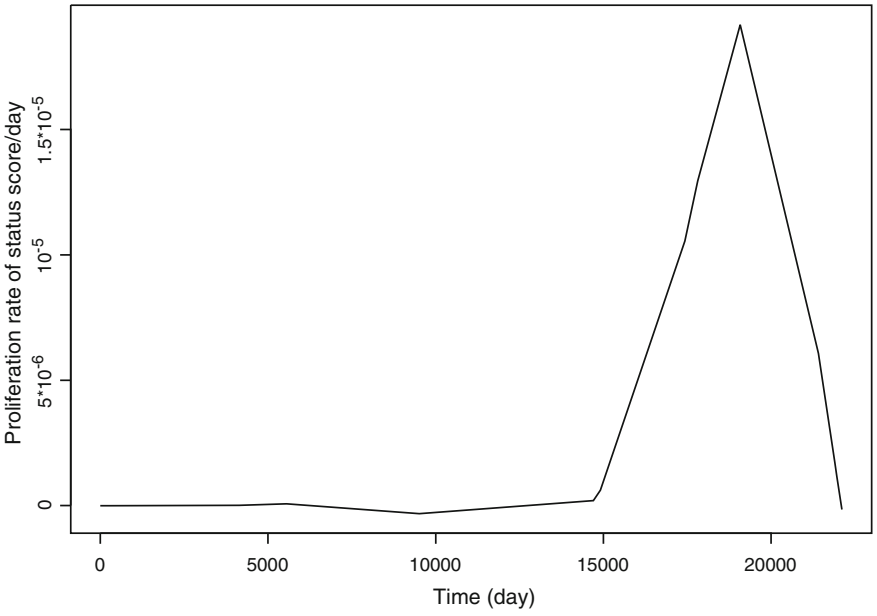

Fig. 4 Proliferation rate of status score from interview with trimmed mean: wt.exp(−0.001 x); spline

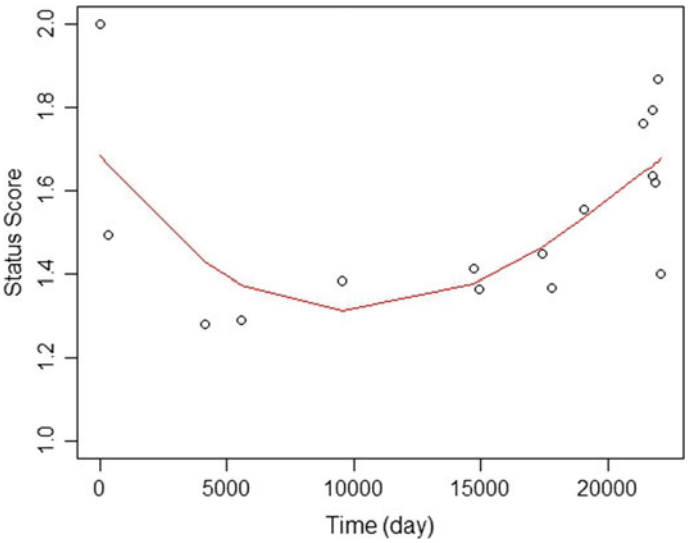

Fig. 5 Spline growth curve of status

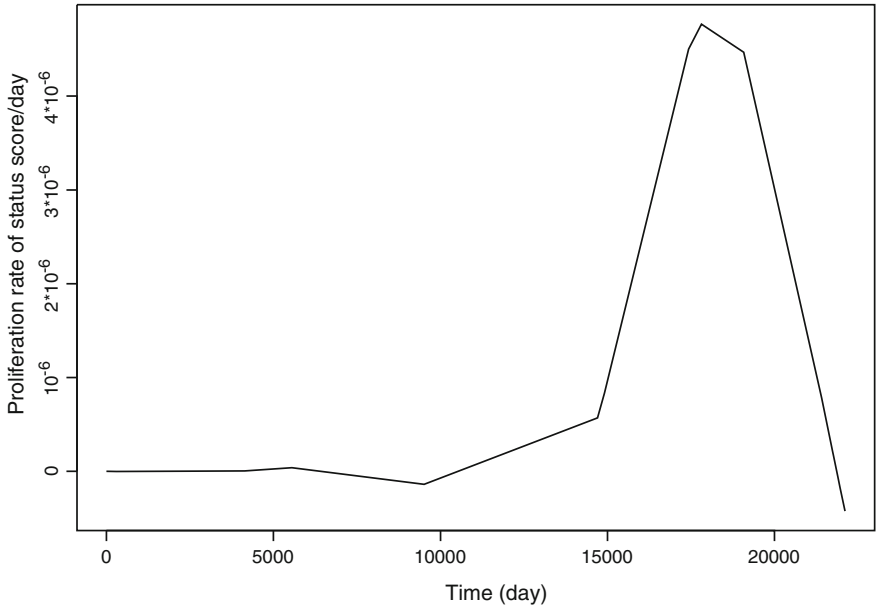

Fig. 6 Proliferation rate of score from interview with trimmed mean: wt.exp(−0.001 x); (Fig. 5)

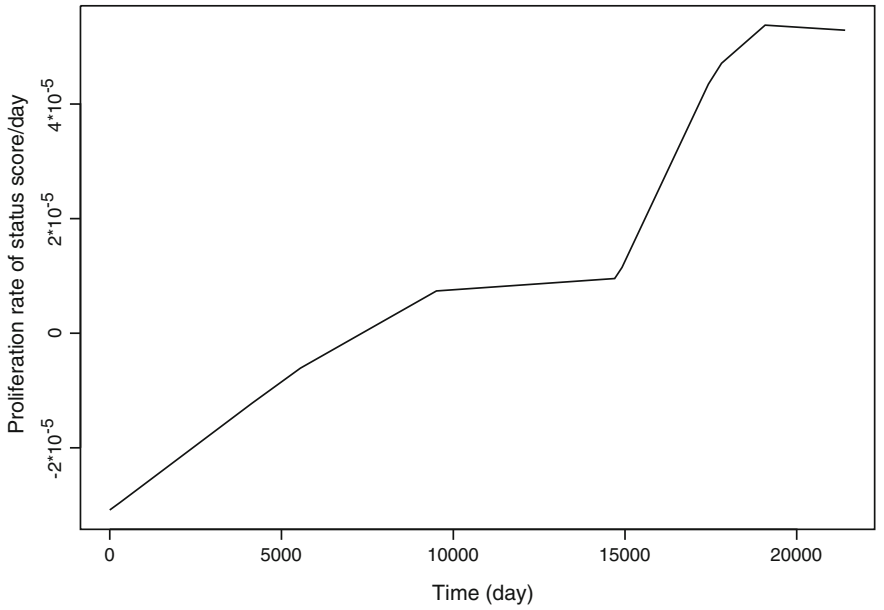

Fig. 7 Proliferation rate of status score from descriptive interview: wt.exp(−0.001 x); (Fig. 5)

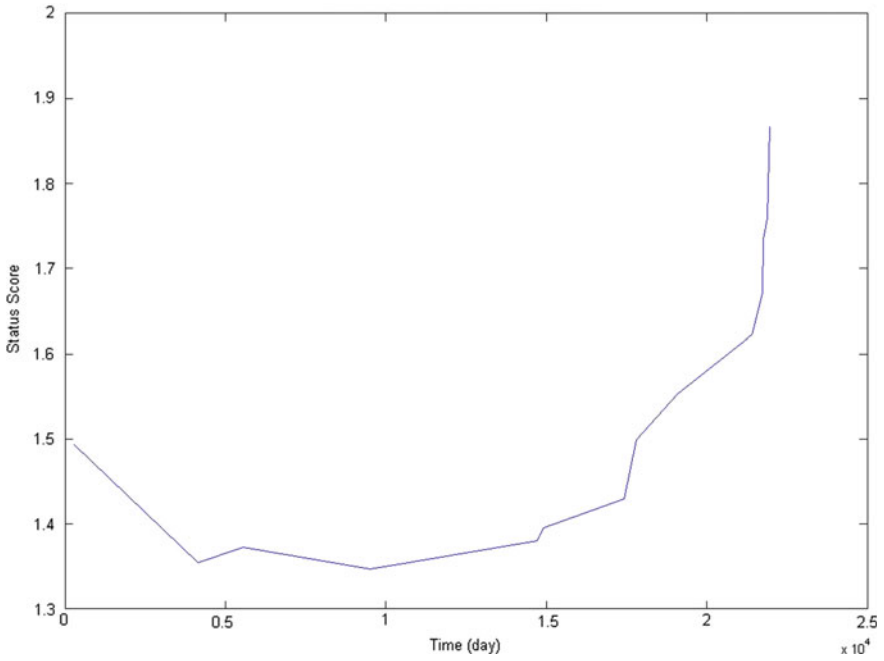

Fig. 8 Growth curve of status score (Fourier smooth)

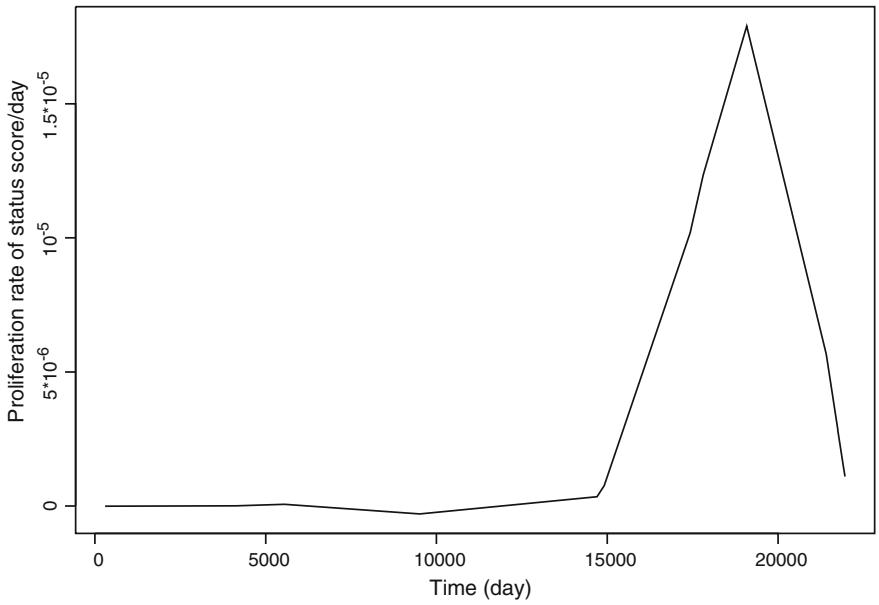

Fig. 9 Proliferation rate of score from interview with trimmed mean: wt.exp(−0.001 x); (Fig. 8)

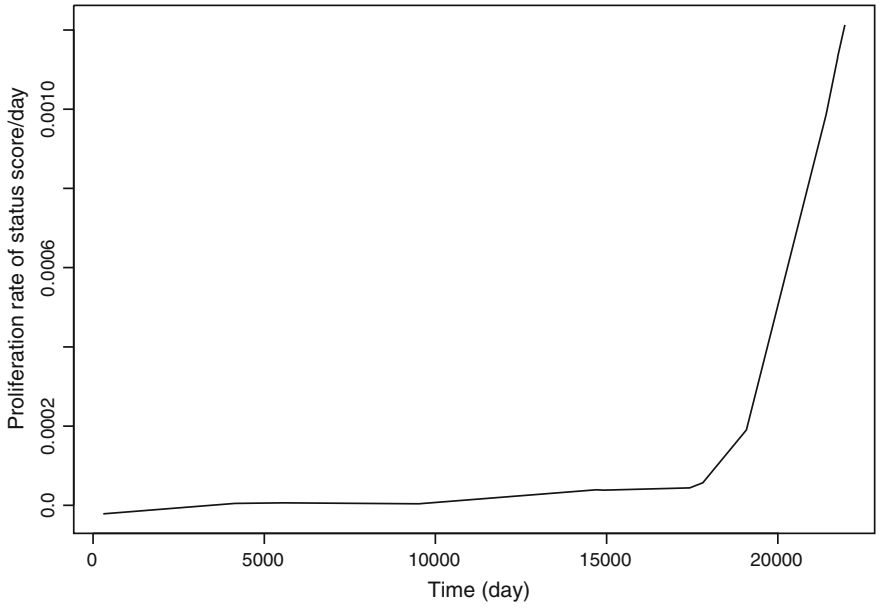

Fig. 10 Proliferation rate of status score from descriptive interview: wt.exp(−0.001 x); (Fig. 8)

Mean scores are computed whenever at least one data is available at a time point, after summing up the scores. We consider linear interpolation of scores to compute sum of scores for individuals, even if there is no recorded data at that time for an individual, but the relevant time point is covered by two adjacent time points, where data for that individual is recorded. Linearly interpolated scores are then added with other scores at a time point and mean score is obtained for that time. Column 2 of Table 1 provides the mean scores of socio-economic status for tribal development for different time points. Columns 3, 4 and 5 represent the nonparametric lowess, spline and Fourier smoothed values, respectively, over time to reveal the growth pattern in scores. Lowess regression attempts to fit linear model by the method of least square to localized subsets of the data. After down-weighting the outliers over several iterations, the method traces a smooth function that describes the deterministic part of the variation present in the data for each point, see, e.g. Cleveland (1981). A cubic spline estimate minimizes residual error of least square plus a roughness penalty measure based on integral of second derivative of the estimate function, see, e.g. Green and silvweman (1994). The technique of Fourier smoothing computes the Fourier Transform (FT) of the data and retransforms it after putting the high frequency noise part to zero in FT, by completely removing the frequency components from a certain frequency and up, see, e.g. http://195.134.76.37/applets/AppletFourAnal/Appl_FourAnal2.html. The first and last values of scores are detected as outliers in MATLAB for Fourier smoothing, so these are not considered in the last column of Table 1, for Fourier technique.

Longitudinal growth curves based on consolidated scores measured on Likert scale are shown in Fig. 1. Some of the large dots in Fig. 1 drawn in Excel represent multiple observations corresponding to more than one individual. In general, an upward trend in status over time is observed.

Mean values of status scores for a time point where at least a single observation in the collected data is available are considered. At those points, score from linear interpolation is also considered for an individual to compute mean. Then the scores are summed up for each such time point, and the mean score is assigned at that time point. Such (x, y) points are shown in Fig. 2. Lowess regression with $f = 0.6$ and three iterations in R yield the nonparametric growth curve of tribal welfare status over a range of more than 60 years. The growth in the curve is prominent especially towards the end.

We compute the proliferation rate based on the growth curve shown in Fig. 2 of lowess regression. Proliferation rate $\frac{d}{dt}\log y = \frac{1}{y}\frac{dy}{dt}$ is a scaled version of velocity $\frac{dy}{dt}$. The measure is independent of the choice of unit used in measuring y. For score y with growth curve computed in Fig. 2, the proliferation rate is obtained in Fig. 3. The curve has an initial slow upward trend till about 15,000 day, and then the curve has a sharp rise, with faster growth till 19,000 day. Beyond this day, the curve shows slight downward tendency towards the end. Computation of rate is based on a technique proposed in Dasgupta (2015b), with exponentially decaying weights $\exp(-0.001x)$ attached to empirical slopes computed from data pairs for time points at distance x with respect to a fixed time point t of interest, where derivative has to

be computed. More weights are assigned to data points near the time t of derivative computation, and less weights are given to distant time points from t. Weighted mean of these empirical slopes at derivative stage and smooth.spline with spar $= 0.0001$ at smoothing stage in SPlus provide proliferation rate at time point t, when divided by y.

There are 16 time points in Table 1 considered in the growth curve construction in Fig. 2. To obtain Fig. 4, we adopt a similar procedure of Fig. 3, but now order the empirical slopes and consider the trimmed mean, the mean of 8th and 9th ordered observations of these empirical slopes at derivative stage. Next, smooth.spline with spar $= 0.0001$ at smoothing stage in SPlus yields proliferation rate at time point t, when divided by y. In contrast to Fig. 3, the rate is comparatively lower in Fig. 4 till the day 15000. Both Figs. 3 and 4 show that the rate curves have sharp rise after 15,000 days till about 19,000 days, indicating the faster change in rates of status score are relatively recent phenomena.

Nonparametric spline regression is used to obtain status growth curve from the score data. The growth curve in Fig. 5 is obtained by smoothing parameter spar $=$ 1.004989, $\lambda = 0.006118$ and with 13 iterations in smooth.spline in R. The equivalent degree of freedom is 3.53227. The curve has a minimum near 10,000 day, and after that there is a gradual sharp rise of growth in status score till the end.

From the growth curve shown in Fig. 5, we obtain the proliferation rates in Fig. 6 following a similar procedure adopted in Fig. 4, where trimmed mean of raw slopes of scores are considered. To some extent, the curve is similar to Figs. 3 and 4. Here the peak of rate is attained at about 17,500 day.

We have seen in earlier figures of proliferation rates that there is downward trend towards the end. When we take weighted mean of all the raw slopes in computation of proliferation rate, the extreme slopes affect the rates. In Fig. 7, last 5 time points (21747, 21752, 21900, 21960, 22111) days are not considered to avoid such problem, and we adopt a similar procedure of computation used in Fig. 3 to get Fig. 7; we consider all the remaining time points to compute the raw slopes and find the weighted sum with monotonically decreasing assigned weight $\exp(-0.001x)$, here x is the distance of the second time point in the set, from present time t; where slope has to be computed for the proliferation rate. The curve in Fig. 7 shows consistently upward trend in the truncated time zone, when upper five values of time are not considered. Proliferation rate curve has several steps in growth over time in Fig. 7. A steep growth is seen after about 15000 day, and sign of stability in rate is seen towards the end.

The curve is drawn by MATLAB with computation of Fourier transform, and then retransforming it, ignoring the noise part. High frequencies from some point and up in FT are put to zero and then retransformed. The first and the last values of scores are detected as outliers in MATLAB for Fourier smoothing, so these are not considered in the Fig. 8. Rise of the curve is sharp towards the end.

Figure 9 on proliferation rate is drawn in a similar manner adopted to draw Fig. 6, considering 14 observations instead of 16 time points; the first and last observations are deleted, as these are detected outliers by MATLAB; the outliers are not shown in Fig. 8 as well.

The curve is drawn in a similar manner like Fig. 7, where weighted sum of raw slopes is considered. Outliers detected by MATLAB in Fourier smoothing are not considered. The rate has a sharp rise towards the end in truncated time zone, indicating changes in status improvement of tribals are relatively recent.

4 Discussions

Social survey of tribal welfare status has relevance in policy making decisions. The present study is conducted on different tribals in north-eastern state Tripura. Survey conducted by stratified random sampling on tribal welfare status of north-eastern tribes reveals a trend of general improvement over time. The analysis covers a time range of about 60 years, from 1952 up to the later part of 2013 (30/08/2013). The origin, day 0 is taken at 1/1/1952, and the last day marked for time in the figures is 30/08/2013; this is the last day of conducted interviews. A part of the survey is retrospective in nature. Scores for answers on questionnaire and general description of past and present events by tribal individuals provide glimpses of tribal lifestyle over years. It appears that much of the improvements in status scores are relatively recent phenomena. There are scopes of improvements in several directions, like convincing individuals on adverse effect from excessive drinking of liquor. Higher qualities of primary education, health and hygiene are to be ensured. Education at primary level should be made as a firm foundation in young mind. For information channelled by education from ground state improves the level of expectations among youths, who aspires for higher income and living standard. Education empowers youth with the mental capacity to devise different ways and means to motivate for improved productivity and enhance living standards. The uprise in status score may be maintained by a proper system of implementation of improvement strategies on the above-mentioned sectors.

Appendix

Psychological trait scores are in general assumed to follow normal distribution. Histogram of sum of scores in Likert scale based on similar type of questions may sometime exhibit pattern of a normal distribution. Here we consider recording the variables in discrete scale and prove some characterization theorems.

In Dasgupta (1993), a characterization of discrete normal distribution based on radial symmetry is obtained, when the support of the distribution is the set of all integers $\mathbf{Z} = \{0, \pm 1, \pm 2, \cdots\}$. In the present paper, we consider a ten-point Likert scale to grade the variables in developmental study.

Characterization theorem for such discrete and bounded variables is possible in a similar manner as described in Dasgupta (1993).We state and prove some results in this direction.

Theorem 1 *Let the joint density of independent and integer-valued random variables* $x_1, x_2, \cdots, x_m$ *for* $x \in Z_n^* = \{0, \pm 1, \pm 2, \cdots, \pm n\}$, *where* n *may take the value* ∞; *be spherically symmetric, i.e. depends only on* $r_m^2 = (x_1^2 + \cdots + x_m^2)$, *and* $P(x_i = 0) > 0, i = 1, 2, \cdots, m$. *Then for* $m \geq 4$ *the form of the p.m.f. is* $p(x) = ce^{-\beta x^2}, x \in Z_n^*$; *for some* $\beta > 0$ *and* c *is such that the total probability is 1.*

Proof of the above is similar to Theorem 1 of Dasgupta (1993). An application of Lagrange theorem and solution of Cauchy equation $f(x_1^2 + x_2^2) = f(x_1^2) + f(x_2^2), x_1, x_2 \in Z_n^*$, lead to discrete normal distribution on the restricted space Z_n^*, in the case n is finite.

Dasgupta (1993) considered not only the Euclidean distance but a different type of distance from the origin, based on biquadrates; $d(x, 0) = ||x|| = (\sum_{i=1}^m x_i^4)^{1/4}$. A characterization theorem for discrete and bounded variable with radial symmetry in biquadrate norm may be obtained in this case.

Theorem 2 *Let the joint density of independent and integer-valued random variables* $x_1, x_2, \cdots, x_m$ *for* $x \in Z_n^* = \{0, \pm 1, \pm 2, \cdots, \pm n\}$, *where* n *may take the value* ∞; *depends only on* $r_m^4 = (x_1^4 + \cdots + x_m^4)$, *and* $P(x_i = 0) > 0, i = 1, 2, \cdots, m$. *Then for* $m \geq 36$ *the form of the p.m.f. is* $p(x) = ce^{-\beta x^4}, x \in Z_n^*$; *for some* $\beta > 0$ *and* c *is such that the total probability is 1.*

Proof of the above uses Dickson's theorem on representation of an integer by sum of 35 biquadrates; see Dasgupta (1993) , Theorem 2 therein. The resulting distribution has a faster fall of tail probability than the discrete version of normal distribution.

Psychological traits measured in Likert scale may require a robust estimate like median rather than sample mean as a measure of central tendency. The relevant distributions where median is the m.l.e. for central tendency are characterized below.

Theorem 3 *Let the joint density of independent random variables* $x_1, x_2, \cdots, x_m$ *with individual support* $(-\infty, \infty)$ *be a function of Manhattan distance, the sum of absolute distance of coordinates from the origin, i.e. joint density depends only on* $r_m = (|x_1| + |x_2| + \cdots + |x_m|)$, *and the density of the variables at origin is positive;* $g_{x_i}(0) > 0, i = 1, 2, \cdots, m$. *Then the marginal density function of each of the coordinate variable is bi-exponential,* $g(x) = ce^{-\beta|x|}, -\infty < x < \infty$ *for some* $\beta > 0$ *and* c *is such that the total probability is 1.*

Proof First consider two independent random variables x_1, x_2 taking values in the real line R. This leads to the Cauchy equation see, e.g. Dasgupta (1993), Rao (1965), p. 158; $f(|x_1| + |x_2|) = f(|x_1|) + f(|x_2|)$, where f is defined in a similar manner. The solution of the above is $f(|x|) = -\beta|x|, |x| \in [0, \infty)$, which leads to bi-exponential distribution, also called the Laplace distribution. The result in higher dimension follows in a similar manner.

Theorem 4 *Let the independent random variables in Theorem 3 be discrete,* $x \in Z^* = [0, \pm 1, \pm 2, \cdots, \pm n)$, *where n may take the value* ∞, *and* $P(x_i = 0) > 0, i = 1, 2, \cdots, m$; *then the form of the p.m.f. is* $p(x) = ce^{-\beta|x|}$, $x \in Z^*$ *for some* $\beta > 0$, *and c is such that the total probability is 1.*

Proof of Theorem 4 for discrete random variable follows the similar steps as those of Theorem 1.

References

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, *35*(1), 54.

Dasgupta, R. (2015a). Growth of tuber crops and almost sure band for quantiles. *Communications in Statistics - Simulation and Computation*. https://doi.org/10.1080/03610918.2014.990097.

Dasgupta, R. (2015b). Rates of convergence in CLT for two sample U-statistics in non iid case and multiphasic growth curve. In R. Dasgupta (Ed.), *Growth curve and structural equation modeling* (Vol. 132, pp. 35–58). Springer Proceedings in Mathematics & Statistics.

Dasgupta, R. (1993). Cauchy equation on discrete domain and some characterization theorems. *Theory of Probability and Its Applications*, *38*(3), 520–524.

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman and Hall.

http://195.134.76.37/applets/AppletFourAnal/Appl_FourAnal2.html

Kaiser, K. (2009). Protecting respondant confidentiality in qualitative research. *Qualitative Health Research*, *19*(11), 1632–1641.

Rao, C. R. (1965). *Linear statistical inference and its applications*. New York: John Wiley.

Interrelationship Between Poverty, Growth, and Inequality in India: A Spatial Approach

Sandip Sarkar and Samarjit Das

Abstract The paper studies on the mechanism of poverty, growth, and inequality in India, in a spatial framework. A balanced panel data set, from five consecutive National Sample Survey Office (NSSO) rounds, has been constructed. The state regions are NSSO stratum and are combinations of different districts of a state of India. We consider a parametric approach that considers not only growth and inequality but also their interactions in the poverty estimation equation. Since the state regions used for this analysis are based on fixed boundaries, and constitution of India allows free movement of citizens, we have also controlled different kinds of spatial dependencies in the model. It has been observed that as a result of increment of poverty of a region, the neighboring state regions' poverty also increases. This is possible due to migration. Our empirical findings also suggest a possible higher number of migrants which are belonging to the class of richer poor. We find as a result of growth poverty reduces but increases due to economic inequality. The policy variables play an important role in the poverty estimation equation and the signs of the coefficient are also appropriate. Several spatially transformed variables that has been incorporated in the poverty estimation equation are found to be statistically significant.

Keywords Poverty · Growth · Inequality · Spatial approach · India

1 Introduction

India has a history of maintaining a sustained growth rate for more than 3%, for the last two decades. Poverty has also been steadily declining in this period. In spite of these two shining aspects of the Indian economy, world's largest number of poor resides in India, following any national or international standards.

S. Sarkar (✉)
Centre for Studies in Social Sciences, Kolkata, India
e-mail: sandip.isi.08@gmail.com

S. Das
Indian Statistical Institute, Kolkata, India

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_5

Our objective in this article is not only to explore the role of growth, but also on the role of distributional effects of income distribution, on poverty reduction. Thus, we also focus on the fact, whether the reduction of poverty is embedded due to unequal incomes. Addressing this problem is not new in the literature, and many theoretical and empirical research has been done in this direction. The central theme of research agenda in this context has been based on estimation of a summary index called growth elasticity of poverty (GEP) (see Ram 2007; Adams 2004; Ravallion and Datt 1998, 2002; Zaman and Khilji 2013). The indicator GEP is important in terms of policy prescription in the sense that it captures the responsiveness of poverty as a result of increase (decrease) of 1% growth. We also focus on the responsiveness of economic inequality on poverty reduction. We refer this parameter as inequality elasticity of poverty (IEP).

Our main contribution in this paper is to incorporate spatial dependencies of the regions in the estimation of these elasticities (i.e., GEP and IEP). So far we have surveyed, all the studies in this area are based on the fact that regions or units of analysis are independent and identically distributed. This might actually be a meaningful assumption in the context of cross-country studies. However, in our context individuals within each units migrate from one part to the other frequently. Further, neighboring regions actually may belong under the same local government, where policies on poverty reduction are same. Further, price of one region may depend to its neighbor. Furthermore, regions closer to developed cities or towns may enjoy certain facilities which actually can play an important role in their poverty reduction. For example, it has been observed that in one of the largest states of India, Uttar Pradesh, the percentage of poor in the western part is 34%, which on the eastern part is much higher (nearly 54%). Since the western part shares a common boundary with Delhi, the development schemes of country's capital might have been trickled down to its neighbor. There are many such observations in this direction, which further motivates us to consider an econometric model with spatial dependencies. Ignoring these dependency would lead to biased and inconsistent estimates of the parameters (see Anselin 2009 for further details).

Our analysis is also new in the sense that instead of considering a state-level analysis, we move to a deeper micro-level analysis of the regions of state. We thus construct a new data set considering the rural and urban regions of different states (state regions) as obtained from National Sample Survey Office Data. These regions are the smallest possible stratum considering the multistage NSSO sampling schemes. Clearly, comparability is not an issue in this regard since the units we consider are independent stratum and the survey design has remained unchanged in this period. It should be noted that consideration of such a micro-level analysis not only increases efficiency merely increasing the number of observations, but also allows us to study on many hidden aspects of heterogeneities within the states, which would have been missed out otherwise. We also include policy variables like education and indoor air pollution in the poverty estimation equation. Since the state regions are based on specific boundaries, we also incorporate spatial dependencies in the model. Ignorance of spatial dependency in the dependent variable in the estimating equation would result to biased and inconsistent estimates of the parameters. Inclusion of the spatial

dependencies, although very important, from the statistical point of view, but has not been done yet in this set up, as per our knowledge. The robustness of the empirical analysis is also tested using three different poverty indices. We begin our analysis by considering a standard Bourguignon (2003) type model, with appropriate panel data specifications. We then incorporate spatial dependence in the model.

The paper is organized as follows. In the next section, we discuss the model. In Sect. 3, we provide a discussion on data. We present the results in Sect. 4.

2 Model

Let P_{it}, Y_{it} and G_{it} represents the poverty, average income, and income inequality of region $i \in \{1, 2, ...N\}$, at time point $t \in \{1, 2, ...T\}$, and p_{it}, y_{it}, and g_{it} as their growth rate.[1] Poverty estimating equation following (Bourguignon 2003) in a panel data context may be written as follows

$$p_{it} = \theta_i + \alpha_1 y_{it} + \alpha_2 y_{it} i_0 + \alpha_3 y_{it}\left(Y_{it}/Z\right) + \beta_1 g_{it} + \beta_2 g_{it} i_0 + \beta_3 g_{it}\left(G_{it}/Z\right) + u_{it} \tag{1}$$

where Z and i_0 are the poverty line and initial income inequality (Gini coefficient). θ_i is the unobserved panel heterogeneity. Consider X as the set of explanatory variables as denoted in Eq. 1. In matrix notation, we can write the model as follows

$$p = X\beta + u \tag{2}$$

where X is the set of all exogenous variables and u is the residual, with usual OLS assumptions. As we mentioned in the introductory section of this paper, one of the central objectives of this paper is measurement of growth elasticity of poverty (GEP) and inequality elasticity of poverty (IEP). GEP (IEP) refers to the responsiveness of poverty reduction, respectively, for increment of 1% growth rate (income inequality). GEP and IEP can be estimated using the following two equations:

$$GEP = \alpha_0 + \alpha_1 i_0 + \alpha_2 (Z/Y) \tag{3}$$

$$IEP = \beta_0 + \beta_1 i_0 + \beta_2 (Z/Y) \tag{4}$$

Bourguignon (2003) also used different poverty estimating equation, including a naive model containing only growth rate of mean income.[2] The rationality of the model in Eq. 1 is that the GEP and IEP do not remain fixed. Clearly, both GEP and IEP depend on the initial Gini coefficients and also the inverse of development

[1] For any variable X, denote the growth rates as x, or $x = \Delta log(X)$.

[2] Bourguignon (2003) assumed that income follows a log-normal distribution. He started with a naive models as $p_{it} = \beta_0 + \beta_1 y_{it} + u_{it}$. A Standard Model, was also proposed may be written as follows $p_{it} = \beta_0 + \beta_1 y_{it} + \beta_2 g_{it} + u_{it}$. It was noticed that R square, increases as one moves from the naive model to the standard model.

indicators, i.e., the ratio of poverty line and mean. In fact, the nonlinearities of the relationship between poverty–inequality–growth relationship are also captured to some extent in Eq. 1.[3] The role of initial conditions on poverty reduction of a society has been empirically established even in the context of India (see Ravallion and Datt 2002, 1998). In Eq. 1, initial income inequality is assumed to be a proxy of initial conditions. The ratio of poverty line and mean income Z/Y_{it} is assumed to be a proxy of inverse development factor.

2.1 Spatial Dependencies

The rationale of spatial econometrics is based on Tobler's first law of geography states that "**Everything is related to everything else, but near things are more related than distant things**." Poverty estimating Eq. 1 is based on the assumption that all the observations are independent and identically distributed (iid). However, since the state regions are based on fixed boundaries, it is possible that they reflect spatial dependencies, where values observed at one location depend on others.

In order to capture spatial dependencies, in terms of an empirical model, one must specify a spatial weight matrix. We consider a contiguous weight matrix, which takes values 1, if two regions are contiguous (neighbors) to each other, else 0. The contiguous matrix $W_N = \{w_{ij}\}$ be a square matrix of spatial weights of size $N \times N$, N is the number of regions, where

$$\begin{aligned} w_{ij} &= 1 \quad (if\ i \neq j \text{ and } i \text{ and } j \text{ are contigous}) \\ &= 0 \quad (else) \end{aligned} \tag{5}$$

Poverty reduction might be modified by incorporating spatial dependencies in the dependent variable and/or in the residuals.

Spatial dependence in the dependent variable: A modified version of Eq. 2 with spatial dependencies of the dependent variable may be written as follows

$$p = \rho W p + X\beta + u \tag{6}$$

The spatial autocorrelation variable is endogenous in the above equation. Thus, OLS estimation given by Eq. 6, would lead to a biased and inconsistent estimate of the parameters. However, consistent estimation of the parameters is possible following a maximum likelihood method of estimation (MLE).[4] In case ρ is statistically

[3] A better way to capture the nonlinearities of the relationship is to adopt a nonparametric estimation equation, similar to Chambers and Dhongde (2011) However, since the state regions are based on fixed boundaries, it is likely to reflect spatial dependencies among each other. Ignoring the spatial dependencies (if it exists) would lead to biased and inconsistent estimates of the parameters. Inclusion of a nonparametric model along with the spatial effects is beyond the scope of the paper.

[4] See Anselin (2009) for further details.

significant, but ignoring the fact we consider a model without the spatial lag variablea and would also give biased and inconsistent estimates.

Spatial dependence in the error terms: It is not always necessary that the spatial dependencies are reflected only in the dependent variable. We also consider a model with spatial dependencies in the residual series and/or in the dependent variable as follows:

$$p = \alpha + \rho W p + X\beta + u + \lambda W_2 u \tag{7}$$

where W_2 denotes the spatial matrix that captures the spatial dependencies of the residual series.[5]

Ignoring the spatial dependencies in the residual series would lead to inconsistent estimation of the standard errors. However, unlike the SAR model even if the spatial dependencies are ignored estimates of the coefficient would be unbiased and consistent. It is possible that the residuals are cross-sectionally dependent, and/or violates the usual assumption of OLS. In order to deal with such situations, one may also use **Driscoll Karry Standard errors** Driscoll and Kraay (1998). It is robust not only to the spatial dependence of the error terms but also handles situations like autocorrelation and heteroskedasticity.

3 Data

The main variable needed for establishing the empirical relationship between growth poverty and inequality is income. However, data on income is not available in India and the present study considers consumption expenditure as a proxy for income. From now on, by income we mean monthly per capita expenditure or MPCE. NSSO conducts a program of quinquennial surveys on consumer expenditure and provides a time series of household consumer expenditure data, which is the prime source of statistical indicators of level of living, social consumption, and well-being, and the inequalities thereof. In this paper, we use NSSO quinquennial surveys on consumer expenditure viz 43rd, 50th, 55th, 61st, and 66th, which provides data for the period of July 1987–June 1988, July 1993–June 1994, July 1993–June 1994, July 1999–June 2000, July 2004–June 2005, and July 2009–June 2010, respectively. In order to collect data on monthly per capita expenditure, one must set a recall period of consumption. We use monthly consumer expenditure on the basis of a mixed recall period data.[6]

[5] Although it is possible to consider a different weight matrices for the dependent variable and residual series, in this case we consider a simple model with a same spatial weight matrix.

[6] In a mixed recall period, data for educational, medical (institutional), clothing, bedding, footwear, and durable goods are collected on a recall period of 365 days. The other items are collected on the basis of a recall period of 30 days. We have used scheduled type 1 data for 66 th round in order to maintain comparability. For details, see NSSO reports.

A balanced panel data set for the five consecutive NSSO rounds is created with rural and urban state regions as the panel variables. A state region is the combination of districts in a state. The number of districts and also states has changed over time. In order to maintain the regional identity over the period, we have to merge more than one state regions in many cases. The number of state regions are 128, of them 64 are rural state regions, and the rest are urban state regions. NSSO considers a multistage sampling survey design, and the rural and urban state regions are the lowest possible stratum.

3.1 Computation of Poverty Growth and Inequality

The first exercise for the poverty estimation of a society is the specification of poverty line. Since poverty lines for the state regions are not available, we use state specific poverty line for the state regions. There has been a change in the methodology of estimating the poverty line, since after the publication of Tendulkar committee report in 2004–05. In order to maintain consistency over periods, we inflate the poverty lines, using consumer price indices for agricultural labor (CPIAL) and consumer price indices for industrial workers, respectively, for rural and urban India. Real MPCE for both rural and urban India is obtained using these price indices. The growth rate of average real MPCE of rural and urban state regions is considered to be the proxy of average growth rate of the society.

3.2 Policy Variables

We now discuss a host of possible variables that serves as controls to our empirical model.

Education: It is possible that a society is able to combat poverty better if the number of literates are higher. Education is also correlated with mean income and inequality. Higher literacy might also lead to higher productivity and consequently higher growth rates. Presence of large number of literates in the society also leads to less corruption. Failing to incorporate the education variable in the estimating equation might have serious consequences of endogenity, resulting in biased coefficients.

NSSO provides data on literacy status of all the individuals coded in different groups viz primary, secondary, higher secondary and graduates and above. We consider the percentage of female adults (aged 15 years or more), having secondary level of education as a proxy of the education variable.[7] We expect a negative coefficient for the education variable in the poverty reduction equation.

[7]We consider the female literacy rates as a proxy of education following Ravallion (Ravallion and Datt 2002)

Air pollution and health hazards through energy consumption (Indoor air pollution) Air pollution might be broadly classified as two different phenomenon namely outdoor phenomenon and indoor phenomenon. The outdoor phenomenon is largely due to the smoke produced by the factories, mostly in the industrial areas. In developing countries, this is often classified as an urban problem. In rural areas, people use bulk of the fuels burned (by mass) as solids, principally wood and coal. Unlike gases and liquids, solid fuels require relatively advanced technology to be pre-mixed with air or otherwise ensure their complete combustion. The airborne emissions of incomplete combustion products, such as carbon monoxide, particulates, and volatile organic compounds, have been extensive. For more details, see Smith (1993) and the references cited therein. The list of health hazards as a result of the indoor pollution that has been documented by Smith is as follows:

(1) Respiratory infections in young children
(2) Adverse pregnancy outcomes for women exposed during pregnancy
(3) Chronic lung diseases and associated heart disease in adults and
(4) Cancer.

Given the data sets, it is not possible to capture the outdoor smoke factor. However, NSSO collects data on principle source of cooking, which might be considered as an indicator of indoor air pollution. We consider the percentage of people effected directly from indoor air pollution as another explanatory variable.[8] A better indoor environment might increase physical abilities of individuals and thus help them combating poverty. Thus, we expect that the coefficient of this variable to be positive.

4 Empirical Results

In this section, we first present some descriptive statistics of the variables that we have used in the analysis. Secondly, we present the Moran's test to have an idea whether the variables indicates the presence of spatial dependence. In the next subsection, we interpret the coefficients of different econometric models. Using the estimation results, we compute GEP and IEP.

4.1 Descriptive Statistics

In Tables 1 and 2, we present the average values of poverty, income inequality, average MPCE, percentage of households having electricity, female literacy rates, percentage of cultivated land in the month of June and July, cultivation and percentage

[8] The percentage of individuals using one of the following coke, coal, firewood and chips, dung cake and charcoal is assumed to be suffer from indoor air pollution.

Table 1 Descriptive statistics: rural India

State names	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Andhra Pardesh	Coastal	32.31	7.62	27.21	983.33	69.34	8.64	8.15	83.49
Andhra Pardesh	Inland	44.86	10.18	25.94	851.52	80.94	7.45	16.51	88.94
Assam	Plains east and west	49.01	10.37	22.22	798.05	34.72	10.80	27.26	91.81
Assam	Hills	52.33	12.09	19.27	732.28	26.72	11.07	53.58	92.83
Bihar	Northen	59.25	15.15	21.38	658.10	5.75	5.88	47.53	87.33
Bihar	Central	63.14	16.16	20.85	636.40	18.38	7.10	35.76	93.73
Gujrat	Eastern+Plains Northen	39.20	9.10	24.73	906.06	81.24	9.82	0.98	83.87
Haryana	Eastern	28.25	6.14	29.97	1299.12	87.46	16.11	34.86	82.30
Haryana	Western	29.13	6.70	28.35	1181.13	84.78	13.11	19.13	84.74
Himachal Pradesh	Himachal	26.64	5.05	28.41	1117.34	95.81	23.38	1.25	80.85
Jammu & Kashmir	Mountains	19.40	3.53	22.28	1085.67	93.81	18.85	0.39	84.52
Karnataka	Coastal ANd Ghat	19.64	3.99	26.14	1030.64	73.20	21.82	1.76	82.32
Karnataka	Inland Eastern	25.37	4.51	21.62	882.72	79.54	13.95	4.78	91.59
Karnataka	Inland Southern	33.85	7.55	24.42	839.42	81.25	10.18	8.10	86.47
Karnataka	Inland Northen	55.45	13.53	22.20	660.35	80.08	8.02	21.73	95.14

(continued)

Table 1 (continued)

State names	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Kerala	Northen	32.80	7.45	30.21	1141.38	68.99	22.35	7.41	88.08
Kerala	Southern	18.58	3.86	35.18	1619.07	79.70	33.84	7.96	80.86
Madhya Pradesh	Vindhya	56.00	14.23	23.49	669.91	54.69	6.93	13.20	98.02
Madhya Pradesh	Central	60.95	16.01	24.61	654.34	66.78	2.93	10.04	97.21
Madhya Pradesh	Malwa	37.36	8.93	28.26	856.60	78.76	2.97	20.29	92.57
Madhya Pradesh	South Central	64.04	18.85	29.79	669.11	64.97	4.62	10.66	96.76
Madhya Pradesh	South Western	66.05	19.22	23.95	604.27	78.95	4.27	12.35	94.45
Madhya Pradesh	Northen	34.73	6.69	23.07	830.83	58.22	4.34	15.57	98.19
Maharastra	Coastal	38.03	9.24	29.24	1016.81	81.60	12.10	2.70	77.82
Maharastra	Inland Western	33.96	7.10	25.92	1029.30	79.69	15.66	8.20	74.92
Maharastra	Inland Northen	55.95	15.82	27.47	827.02	72.27	11.61	4.75	70.75
Maharastra	Inland Central	54.62	16.54	28.51	820.20	76.18	8.19	16.80	65.69
Maharastra	Inland Eastern	57.17	16.03	26.34	806.40	71.10	13.88	11.73	88.16
Maharastra	Eastern	69.05	19.67	24.86	720.21	59.60	10.55	3.83	90.48
Manipur	Plains	49.90	9.22	15.68	921.78	86.93	31.10	55.98	69.51
Manipur	Hills	65.46	14.78	17.58	835.23	64.68	18.11	12.45	93.70

(continued)

Table 1 (continued)

State names	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Meghalaya	Meghalaya	27.25	4.19	19.75	898.12	53.67	9.72	2.93	97.11
Orissa	Coastal	49.05	11.29	23.65	653.96	42.76	10.81	24.85	89.57
Orissa	Southern	78.15	28.23	23.23	462.53	16.74	2.62	8.20	96.80
Orissa	Northen	61.35	17.45	25.81	586.07	24.93	7.01	10.89	94.14
Punjab	Northen	20.29	3.14	28.27	1395.62	94.14	24.52	25.35	69.19
Punjab	Southern	25.67	4.84	27.29	1299.00	93.35	14.01	30.77	75.33
Rajasthan	Western	35.23	6.99	22.49	963.85	46.27	3.02	20.30	95.33
Rajasthan	North Eastern	31.70	6.31	22.25	984.41	59.56	4.38	24.34	94.81
Rajasthan	Southern	51.95	12.23	25.93	875.64	41.80	4.25	3.17	95.04
Rajasthan	South Eastern	37.43	8.04	22.95	935.92	66.76	4.02	8.18	94.56
Sikkim	Sikkim	39.41	7.65	24.30	949.09	91.21	15.75	5.15	67.28
Tamil Nadu	Coastal Northen	48.07	12.59	30.18	812.04	81.84	15.32	6.08	80.40
Tamil Nadu	Coastal	27.74	5.35	25.70	924.83	73.23	12.94	8.35	89.59
Tamil Nadu	Southern	38.56	8.74	25.00	815.82	79.32	13.13	6.50	86.78
Tamil Nadu	Inland	37.64	8.13	30.38	920.75	79.55	11.15	3.64	76.84
Tripura	Tripura	35.74	7.22	21.40	827.76	58.06	7.40	3.60	95.31
Uttar Pradesh	Western	36.14	7.30	25.99	891.73	32.95	8.03	60.58	94.46
Uttar Pradesh	Central	50.66	13.19	25.20	747.90	11.59	6.75	22.15	96.53
Uttar Pradesh	Eastern	54.03	13.10	25.06	738.15	25.74	8.53	60.67	90.79
Uttar Pradesh	Southern	52.81	14.73	30.06	793.53	24.40	4.38	15.40	99.06
West Bengal	Himalayan	42.70	8.80	20.81	766.94	24.56	7.23	3.29	96.17
West Bengal	Eastern plains	50.99	11.34	23.98	736.40	24.92	4.79	9.05	81.40

(continued)

Table 1 (continued)

State names	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
West Bengal	Central Plains	33.84	6.65	23.62	839.78	37.30	7.71	8.18	82.92
West Bengal	Western Plains	43.76	10.07	25.85	796.29	25.95	6.84	5.42	80.27
Arunachal Pradesh	Arunachal Pradesh	41.62	10.61	30.34	1066.05	52.44	13.18	1.60	88.42
Chandigarh	Chandigarh	15.21	2.79	24.88	1619.58	89.43	21.85	21.49	18.14
Delhi	Delhi	10.96	1.72	25.32	1469.08	96.98	33.14	0.22	11.34
Goa	Goa	24.50	5.09	27.71	1524.34	98.05	29.32	1.33	45.44
Mizoram	Mizoram	30.14	5.22	20.03	1087.75	70.00	11.92	2.17	82.30
Pondicheri	Pondicheri	16.37	3.05	30.30	1231.87	78.23	17.79	5.46	74.04
Chattisgarh	Chhattisgarh	61.20	16.13	24.89	649.87	59.34	7.37	1.45	97.37
Uttarakhand	Uttaranchal	30.61	5.27	27.30	1027.59	64.38	14.06	0.64	80.14
Jharkhand	Jharkhand	59.42	15.40	22.75	653.29	22.95	5.28	1.41	96.84

The table contains average values of all the indicators
F. literacy implies female literacy, and C. fuels implies percentage of households using non-combustible cooking fuels

Table 2 Descriptive statistics: urban India

Statenames	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Andhra Pardesh	Coastal	28.31	6.27	36.01	1709.70	88.19	27.00	0.72	35.76
Andhra Pardesh	Inland	27.85	6.11	34.00	1650.88	94.00	32.31	1.27	30.01
Assam	Plains east and west	28.71	6.12	29.93	1438.55	84.54	41.52	0.61	27.61
Assam	Hills	31.80	6.98	31.57	1462.08	79.32	39.40	0.96	37.59
Bihar	Northen	50.63	13.72	30.42	950.79	46.05	24.14	3.70	61.75
Bihar	Central	44.38	11.11	31.02	1053.80	76.67	31.14	4.10	52.45
Gujrat	Eastern+Plains Northen	27.26	5.74	30.44	1564.49	93.84	35.52	0.10	20.43
Haryana	Eastern	22.66	4.97	32.11	1806.76	94.47	39.58	4.24	25.75
Haryana	Western	26.69	6.22	30.01	1568.07	93.59	36.41	2.48	34.05
Himachal Pradesh	Himachal	15.62	3.03	37.74	2137.30	96.35	56.36	0.04	15.36
Jammu & Kashmir	Mountains	9.29	1.51	27.19	1659.49	99.06	47.89	0.04	15.46
Karnataka	Coastal ANd Ghat	25.80	5.53	35.12	1726.58	95.20	43.35	0.14	35.78
Karnataka	Inland Eastern	24.89	4.85	25.95	1390.09	91.77	34.32	1.20	33.56
Karnataka	Inland Southern	13.24	2.63	31.07	1964.11	94.62	44.63	0.90	12.99
Karnataka	Inland Northen	49.79	13.58	30.18	1122.13	88.58	31.35	4.59	52.00

(continued)

Table 2 (continued)

Statenames	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Kerala	Northen	30.30	6.72	37.23	1487.54	86.50	32.56	1.71	70.31
Kerala	Southern	14.29	3.00	39.46	2141.32	90.41	43.73	2.18	52.15
Madhya Pradesh	Vindhya	35.57	8.58	31.43	1157.88	91.10	28.04	1.75	54.39
Madhya Pradesh	Central	36.28	9.05	38.18	1310.27	95.88	35.91	1.64	37.73
Madhya Pradesh	Malwa	23.54	5.30	34.52	1512.02	97.42	33.25	1.70	33.25
Madhya Pradesh	South Central	36.89	9.09	35.05	1225.89	93.48	31.96	1.43	48.91
Madhya Pradesh	South Western	39.58	10.00	30.70	1076.97	95.46	30.96	1.87	41.82
Madhya Pradesh	Northen	35.09	8.56	30.72	1134.02	93.14	27.60	3.23	53.78
Maharastra	Coastal	9.07	1.49	34.13	2315.84	97.82	43.96	0.37	1.65
Maharastra	Inland Western	28.45	6.43	37.41	1832.41	93.64	39.05	0.92	14.01
Maharastra	Inland Northen	45.62	13.05	33.41	1336.43	92.56	33.11	0.75	19.56
Maharastra	Inland Central	55.18	16.74	33.86	1168.85	92.30	24.87	2.21	36.09
Maharastra	Inland Eastern	46.28	13.48	36.13	1374.77	93.03	38.00	2.40	32.01
Maharastra	Eastern	37.13	9.48	28.34	1357.22	89.99	35.03	0.42	31.83
Manipur	Plains	47.36	9.27	18.91	1081.56	94.38	45.09	37.52	39.78
Manipur	Hills	64.25	15.92	15.84	911.65	96.01	22.68	0.19	66.04
Meghalaya	Meghalaya	25.50	4.24	24.47	1503.81	96.17	46.50	0.02	33.39

(continued)

Table 2 (continued)

Statenames	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
Orissa	Coastal	35.81	8.20	34.69	1216.41	78.06	30.48	0.98	51.29
Orissa	Southern	39.33	12.09	35.24	1102.30	68.19	26.11	0.81	59.22
Orissa	Northen	31.15	7.43	31.09	1173.68	78.46	29.74	0.73	56.46
Punjab	Northen	22.28	4.04	31.96	1784.67	97.72	45.93	2.85	16.35
Punjab	Southern	26.03	5.49	32.87	1714.22	97.41	42.69	3.65	26.06
Rajasthan	Western	25.07	4.90	27.92	1336.65	90.84	21.09	1.00	40.31
Rajasthan	North Eastern	28.79	6.07	35.35	1523.98	91.54	29.05	1.74	42.95
Rajasthan	Southern	19.58	3.88	28.24	1556.72	94.35	30.39	0.17	29.85
Rajasthan	South Eastern	29.23	6.72	30.87	1369.06	94.92	26.30	0.58	35.58
Sikkim	Sikkim	22.72	4.33	24.04	1628.45	97.07	34.95	0.06	4.02
Tamil Nadu	Coastal Northen	20.65	4.90	36.38	1742.69	93.00	39.50	1.00	19.26
Tamil Nadu	Coastal	23.92	4.99	31.04	1401.62	88.50	33.62	1.12	38.03
Tamil Nadu	Southern	30.82	7.08	34.87	1349.42	91.91	30.57	1.40	40.38
Tamil Nadu	Inland	22.89	4.28	34.54	1510.18	90.98	29.43	0.81	30.79
Tripura	Tripura	19.39	3.45	29.77	1428.41	88.32	30.23	0.15	51.23
Uttar Pradesh	Western	33.99	8.05	34.77	1301.57	81.97	30.25	5.17	46.79
Uttar Pradesh	Central	35.21	8.85	35.91	1359.13	78.63	38.79	1.57	35.72
Uttar Pradesh	Eastern	41.70	10.22	30.94	1105.41	78.61	28.73	2.96	46.50
Uttar Pradesh	Southern	51.77	14.30	28.53	1006.81	70.92	28.69	1.13	55.09
West Bengal	Himalayan	33.47	8.13	30.56	1270.82	80.04	31.58	0.05	49.52

(continued)

Table 2 (continued)

Statenames	State regions	HCR	PGR	Gini	MPCE	Electricity	F. literacy	Cultivation	C. fuels
West Bengal	Eastern plains	39.12	9.91	32.67	1235.13	71.46	26.75	0.28	47.60
West Bengal	Central Plains	22.90	4.80	36.52	1681.03	84.64	34.68	0.59	36.83
West Bengal	Western Plains	33.50	8.12	33.39	1372.66	73.17	27.73	0.19	45.04
Arunachal Pradesh	Arunachal Pradesh	25.42	5.95	28.39	1489.13	92.94	37.53	0.06	29.63
Chandigarh	Chandigarh	10.75	2.16	40.77	3172.57	95.28	55.07	13.70	4.90
Delhi	Delhi	15.55	3.04	36.34	2467.55	98.37	49.17	0.30	2.65
Goa	Goa	15.37	2.83	33.10	2090.59	96.90	42.70	0.73	10.11
Mizoram	Mizoram	9.68	1.50	22.00	1645.24	96.21	34.20	1.04	23.19
Pondicheri	Pondicheri	13.82	3.06	30.87	1615.69	91.72	37.64	3.61	28.52
Chattisgarh	Chhattisgarh	30.36	7.18	32.66	1301.20	89.47	37.33	0.14	48.14
Uttarakhand	Uttaranchal	22.21	5.00	30.60	1518.17	95.24	43.71	0.04	19.88
Jharkhand	Jharkhand	34.25	8.67	34.51	1337.56	81.06	34.28	0.06	63.48

of households whose chief source of cooking fuels is prone to causing different health hazards and also indoor air pollution, respectively, for rural and urban India.

Less developed states like Bihar, Orissa, Madhya Pradesh, Chattisgarh, and Jharkhand show poor performance in most of the indicators. The poverty indices are also high for these states. The similar pattern is also observed in the highly developed states like Punjab, Haryana, Delhi Chandigarh. These states are also contiguous to each other and perform better in respect to almost indicators. In southern and western states, developed states are Karnataka, Kerala, Tamil Nadu, and Maharastha. They are also contiguous in most of the cases. This pattern indicates the presence of spatial dependence among the variables, including the poverty indices.

Although states seem to exhibit a spatial pattern, however, intra-state inequalities are also inevitable in many cases. For example, the Rural Malwa regions of Madhya Pradesh show average HCR of 37.36%, whereas in the same state the southeastern part exhibits a poverty rate of 66.05%. Similarly, the western part of Uttar Pradesh (contiguous to Delhi) exhibits a poverty rate of 36% part, whereas the other regions show much higher poverty rates.

As expected, the air pollution factor captured by the usage of cooking fuels material shows huge difference between the rural and urban regions. In all cases, more than 80% of the individuals use cooking fuels harmful to health. The only exception being Delhi and Chandigarh where in the rural areas this rate is less than 20%. Percentages of households having access to electricity are better in most of the cases in urban India, but in rural India performance of some states is extremely poor, e.g., Jharkahand, Central Uttar Pradesh, Southern Orissa. The variation of the Gini coefficent has not been extensive.

4.2 Moran's Test

Although the average values reflect a possibility of presence of spatial depencies in the model, we consider a statistical tests for all the variables to check the presence of spatial dependencies if any.

A Moran's test is a crude indicator for the tests of spatial dependence in the data, and the Moran's I can be written as

$$I = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{N^* \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2} \tag{8}$$

where $N^* = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}/N$.

The expected value and its standard error can be derived easily. Moran's test reflects the spatial dependence in the cross-sectional case. If the Moran's I is positive (negative), then it might be concluded that the states' performance is positively (negatively) affected by the neighbor. In Table 3, we report the Moran's I along with the *Probability Values (P Values)* for the test $H_0 : I = 0$. It is seen that the for most of the rounds, all the variables are spatially related.

Table 3 Moran's test

	Round 50		Round 55		Round 61		Round 66	
	I	Pval	I	Pval	I	Pval	I	Pval
Rural India								
$\Delta log(HCR)$	0.07	0.14	0.28	0.00	0.26	0.00	0.10	0.01
$\Delta log(PG)$	0.06	0.18	0.27	0.00	0.30	0.00	0.07	0.07
$\Delta log(MPCE)$	−0.04	0.40	0.13	0.04	0.19	0.01	0.02	0.31
$\Delta log(gini)$	−0.12	0.11	0.14	0.03	0.14	0.02	0.09	0.09
Electricity	0.55	0.00	0.60	0.00	0.54	0.00	0.56	0.00
Education	0.32	0.00	0.40	0.00	0.50	0.00	0.37	0.00
Cultivation	−0.01	0.47	−0.04	0.37	0.00	0.43	−0.07	0.22
Cooking	0.02	0.30	0.05	0.13	0.03	0.23	0.07	0.11
Urban India								
$\Delta log(HCR)$	0.08	0.11	0.05	0.22	0.03	0.29	0.19	0.00
$\Delta log(PG)$	0.06	0.17	0.02	0.31	0.03	0.27	0.20	0.00
$\Delta log(MPCE)$	0.05	0.22	−0.01	0.48	−0.08	0.21	0.03	0.28
$\Delta log(gini)$	−0.02	0.46	0.01	0.36	0.01	0.39	−0.01	0.46
Electricity	0.37	0.00	0.51	0.00	0.40	0.00	0.34	0.00
Education	0.20	0.00	0.07	0.14	0.19	0.01	−0.04	0.40
Cultivation	0.14	0.03	0.13	0.03	0.19	0.01	0.05	0.22
Cooking	0.22	0.00	0.19	0.00	0.08	0.10	0.07	0.13
Rural and Urban India								
$\Delta log(HCR)$	0.07	0.02	0.14	0.00	0.14	0.00	0.14	0.00
$\Delta log(PG)$	0.04	0.11	0.14	0.00	0.16	0.00	0.14	0.00
$\Delta log(MPCE)$	0.01	0.34	0.02	0.25	0.04	0.12	0.07	0.02
$\Delta log(gini)$	−0.04	0.22	0.10	0.00	0.12	0.00	0.05	0.08
Electricity	0.23	0.00	0.28	0.00	0.24	0.00	0.30	0.00
Education	0.00	0.45	0.01	0.30	0.04	0.10	0.03	0.17
Cultivation	0.01	0.28	0.00	0.39	0.01	0.34	0.01	0.32
Cooking	0.05	0.05	0.04	0.09	0.03	0.18	0.00	0.43

4.3 Econometric Results

As mentioned earlier, we generalize the Bourguignon model by incorporating certain policy variables such as cooking fuel consumption and female literacy rates. We also incorporate spatial dependencies in the dependent variable by considering a spatial autoregressive model. The results are reported in Table 4. For the robustness of the analysis, we also consider three different poverty indices viz the head count ratio (HCR), poverty gap (PG), and squared poverty gap (SPG).[9] In order to estimate the

[9]All the three poverty indices belong to the class FGT (Foster et al. 1984) index. HCR is often criticized as naive index that does not consider the inequality among the poor. Both PG and SPG

Table 4 Spatial model

Variables	HCR	PG	SPG
y	−6.41[a] (0.86)	−8.31[a] (1.04)	−9.83[a] (1.36)
$y \times i_0$	7.30[a] (2.31)	10.55[a] (2.80)	12.74[a] (3.64)
$y \times z/Y$	3.20[a] (0.49)	3.01[a] (0.59)	3.10[a] (0.77)
g	4.62[a] (0.66)	5.32[a] (0.80)	5.81[a] (1.03)
$g \times i_0$	−2.67 (1.72)	−2.45 (2.07)	−1.81 (2.70)
$g \times z/Y$	−3.77[a] (0.40)	−3.79[a] (0.48)	−3.94[a] (0.63)
$F.literacy$	−0.05 (0.04)	−0.05 (0.05)	−0.04 (0.06)
$Cooking$	0.07 (0.05)	0.09[c] (0.05)	0.10 (0.07)
ρ	0.20[b] (0.08)	0.12[c] (0.07)	0.10 (0.07)

Notes [a]Estimated results based on a spatial autoregressive model with Driscoll Karry Standard errors

[b]Set of instruments is electricity consumption, average cultivated lands, percentage of households using non-combustible cooking materials, female literacy rates (secondary Level) and MPCE from NSSO employment–unemployment rounds

[c]The notations for the first six variables are similar to Eq. 1. In the parenthesis, we report standard errors. a, b, and c implies significance at 1%, 5%, 10%, respectively (two-tailed test)

standard errors consistently, we use the SAR models with Driscoll Karry Standard errors (DSKSEs) Driscoll and Kraay (1998). DSKSE captures all kinds of cross section and temporal correlation of the residuals. Another option would have been the consideration of spatial dependencies not only in the dependent variable, but also in the error part as in Eq. 7. However, we find insignificant λ in all cases, thus incorporating such models would lead to inconsistent estimation of the standard errors.

From Table 4, it can be observed that the spatial autocorrelation parameter is positive and highly significant for all the cases. As we have mentioned earlier, ignoring this dependency would lead to biased and inconsistent estimates of the parameter. The positivity of the spatial autocorrelation parameter implies that poverty rate of a region is positively related to its neighbor's poverty rate. For example, $\rho = 0.17$ implies poverty of a region and increases by 1.7% if its neighbors' poverty increases by 10%. One possible explanation for the positivity of ρ may be because of labor migration. When poverty of a region increases, poor people migrate to the neighboring regions, and consequently, poverty of that region also increases. The value of the parameter also declines in going from HCR to PG and SPG. This might result from a migration of the richer poor or individuals whose incomes are close to the poverty line. This is the most common type of migration in most developing economies, for further details see Du et al. (2005). If a poor individual enters a society as a migrant, HCR would always increase, independent of the migrant's income. However, since PG and SPG are distributive sensitive, they would depend on the migrant's income. In

take account of the inequality among the poor. Consequently, the poorest of the poor gets more weight in the computation of these indices. SPG is even more general than PG in this regard.

the appendix of the paper, the analytical derivations relate migration and the growth rate of the three poverty indices. The sign and significance of the first six variables match exactly to the earlier papers that have worked with this type of model. The variables are necessary for the computation of growth and inequality elasticities of poverty.

From Table 4, it is also observed that the coefficients of the interaction effects respectively for growth and inequality are of opposite signs. Positivity of the interaction of the growth rate and initial inequality implies higher initial inequality and/or ratio of poverty line and mean income, and reduces the responsiveness of growth on poverty reduction. The negativity of the inequality terms is also intuitively justified. Societies with higher initial inequality and/or poverty line mean ratio reduce the effects of IEP.[10]

The female literacy variable is significant in all the cases except for the choice of SPG as the dependent variable. Role of literates in poverty reduction has been widely documented in Gundlach et al. (2004). Hirway's argument on the problems on identification of the BPL households might also be related to the literacy rates of the society. For example, a poor household might have no information on the BPL cards and do not know how to get in to the lists. If in a society literacy rates are high, the literates might inform the poor people about the BPL cards and also on the procedures to get it. These effects also have been reflected in the poverty estimation equation, resulting the coefficient to be negative. The sign of the cooking variable is positive and significant in all cases. It implies that the state regions where individuals are highly dependent on cooking fuels, also exhibit high degree of poverty.

5 Growth and Inequality Elasticity of Poverty

In this section, our aim is not only the computation of GEP and IEP, but also to interpret the results of the variables that captures the nonlinearities of the poverty–growth and inequality relationships, i.e., first six estimates of Table 4.

In Table 5, we report the predicted elasticities of rural and urban India for different time points, for which NSSO surveys was conducted. As we have mentioned at the beginning of this article, the expected sign of GEP is negative, implying as a result of increase of growth poverty declines. Although the negativity of GEP is a good sign of the economy, the positive values of IEP should not be ignored. The value of the coefficient is substantially higher and in fact close to GEP in many cases. The positivity of IEP implies the adverse effects of inequality and somehow reduces the force of growth to reduce poverty. One such adverse effect of inequality is the interrelationship between income inequality and corruption.[11]

[10]For more details on these coefficients, see Kalwij and Verschoor (2007). The authors have computed the analytical forms of the elasticities are assuming a log-normal income distribution.

[11]Sung and Khagram (2005) have shown that corruption is related to greater inequalities, and the adverse effect is larger in democratic countries. Corruption on the other hand might directly affect

Table 5 Predicted GEP and IEP for rural and urban India

Year	GEP			IEP		
	HCR	PG	SPG	HCR	PG	SPG
Rural India						
1993–94	−1.52(0.52)	−2.69(0.56)	−3.52(0.63)	0.51(0.59)	1.24(0.60)	1.78(0.63)
1999–00	−1.70(0.55)	−2.87(0.56)	−3.70(0.62)	0.72(0.66)	1.46(0.67)	2.00(0.71)
2004–05	−1.88(0.49)	−3.04(0.54)	−3.87(0.61)	0.93(0.54)	1.67(0.55)	2.22(0.57)
2009–10	−2.08(0.48)	−3.22(0.52)	−4.07(0.58)	1.16(0.55)	1.91(0.55)	2.47(0.58)
Urban India						
1993–94	−1.88(0.48)	−2.91(0.51)	−3.68(0.57)	1.14(0.55)	1.89(0.56)	2.48(0.59)
1999–00	−2.19(0.38)	−3.20(0.46)	−3.98(0.53)	1.50(0.39)	2.25(0.39)	2.85(0.41)
2004–05	−2.25(0.41)	−3.25(0.48)	−4.03(0.55)	1.57(0.44)	2.33(0.44)	2.93(0.46)
2009–10	−2.34(0.47)	−3.33(0.49)	−4.12(0.55)	1.67(0.57)	2.43(0.58)	3.03(0.61)

GEP and IEP are predicted from Eqs. 3 and 4. The value of the parameters is from the spatial model with additional endogeneous variable

Set of instruments is electricity consumption, average cultivated lands, percentage of households using non-combustible cooking materials, female literacy rates (secondary Level), and MPCE from NSSO employment unemployment rounds

The notations for the first six variables are similar to Eq. 1. In the parenthesis, we report standard errors. *a*, *b*, and *c* implies significance at 1%, 5%, 10%, respectively (two-tailed test)

It might be readily observed that the absolute values of both GEP and IEP increase with time. This is possible because in this period, mean income has increased substantially leading to decline of Z/Y ratio. It can also be seen that the absolute value of this elasticities increases as we move from HCR to PG and to SPG. This result is intuitively justified if we focus on the axiomatic literature of poverty measurement. Head count ratio is a naive indicator and thus gives equal weight to all the poors. Thus even if income of an individual increases, HCR may remain constant if she is unable to cross the poverty line. Both PG and SPG would decline as a result of such changes in income distribution. This property is also widely known as monotonicity axiom as suggested by the seminal article of Sen (1976). SPG is more general in this regard, since it responds to transfer of income among the poor, widely known as the transfer axiom.

Although low IEP is desirable, in some cases economies with very poor initial condition and low mean income resulted in negative IEP. For example in rural India, we find evidence of negative IEP in three cases: southern regions of Orissa, south-

the effectiveness of many policies against poverty. Government of India considered a programme of targeting the most needy. A measure was developed by which families were categorized as living "**Below the poverty line**." Identified rural families that are below the poverty line are eligible for government support such as subsidized food or electricity and schemes to construct housing and encourage self-employment activities. As pointed by Hirway (2003), "the rich and powerful in a village frequently pressurizes the talati and the sarpanch to include their names in BPL lists." Thus in a society with higher income inequality, instead of the poor households, rich households receive the benefits.

western regions of Madhya Pradesh, and in hilly areas of Manipur. The values of IEP for these regions are, respectively, −0.82, −0.20, and −0.07. As Kalwij and Verschoor (2007) pointed out, IEP is positive unless the region has a low average income. Out of these three regions, the southern part of Orissa is famous for the famine of **Kalahandi**, which took place in the 1980s, in the districts of Kalahandi. This region historically suffers from low growth rate particularly because of the deterioration of the agricultural conditions.[12] Southwestern Madhya Pradesh is also a drought prone region in this neighborhood. If growth rate of poverty is zero, high inequality may reduce poverty for HCR only by transfer of income from the poorest poor to the richer poor, such that latter cross the poverty line.

6 Conclusion

This paper studies on the mechanism of "*Poverty–Growth–Inequality*" in India. We consider Bourguigon (2003) type model, where the poverty estimation equation is derived from a theoretical assumption that income follows log-normal distribution. We construct a balanced panel data set from five consecutive NSSO quinquennial rounds. The panel variable used for the study is rural and urban NSSO state regions, which are lowest possible NSSO stratum. The state regions are combinations of different districts. Since constitution of India allows free movement of citizens, and the state regions are also based on specific boundaries, they might reflect spatial dependencies. In order to capture these dependencies, we consider a spatial autoregressive model, which incorporates spatial dependencies of the dependent variable. We also tried to capture the spatial dependencies of the residuals, but end up with insignificant results. To estimate the standard errors consistently, we use Driscoll Karry Standard errors (Driscoll and Kraay 1998).

For robustness of the analysis, we consider three alternative poverty indices, following the first three indices of the FGT family viz the HCR, PG, and SPG. We consider three alternative model specifications for the choice of three different poverty indices. Our first contribution to the literature is to incorporate additional policy variables in the poverty estimation equation which are female literacy rate and indoor air pollution via cooking. The policy variable coefficient although is smaller in magnitude, but is of appropriate signs. We find that education reduces poverty, whereas indoor air pollution increases it.

Considering a spatial autoregressive model (SAR) model, we find a strong evidence of spatial dependence in the dependent variable. The spatial autoregressive parameter ρ is highly significant and strictly positive in all the cases. Thus, it can be inferred poverty of a region is positively related to its nearby regions. Migration

[12]Historically, these areas are drought prone, with low rainfall over decades. Low agricultural production in this region also has led to different types of aids and supports from the government. This however led to further decline in agricultural production incentives and also agricultural prices. Since rural India is mostly related to agricultural productions, income growth rates also behave accordingly with the deterioration. For further details on the Kalahandi famine, see Pradhan (1993).

might be an important factor for this result. Poor people migrates for the search of jobs and better employment facilities to its neighboring regions. We also find that the value of the parameter declines as we move from HCR to PG and SPG. Higher number of migrants from the richer poor may lead to such result.

We use the coefficients of the SAR model with endogeneity to predict the responsiveness of growth and inequality on poverty, or growth elasticity of poverty (GEP) and income elasticity of poverty (IEP). The GEP estimates are negative and the absolute value is increasing over time. On the other hand, we find that IEP values are positive and are increasing over time. The values of the IEP are sufficiently large and close to the GEP in many cases. Absolute values of both GEP and IEP are found to higher in urban areas compared to the rural areas in most of the cases. However, the gap between the absolute values of the elasticities is in generally higher for the rural areas. Negative IEP has also been observed in some of the cases, resulted because of the low poverty line and mean income ratio.

Acknowledgements We sincerely acknowledge one anonymous referee for comments on an earlier version of this paper. We also acknowledge Manoranjan Pal and Sattwik Santra for helpful comments. This paper is a part of PhD thesis of Sandip Sarkar awarded by the Indian Statistical Institute in 2015. The usual disclaimer applies.

Appendix

In a society, let at time point t, $y_t = (y_1, y_2, ..y_q...y_n) \in \mathbb{R}^n_{++}$, be the income distribution, arranged in ascending order. where n is the number of individuals, and q is the number of poor. The class of FGT index (Foster et al. 1984) for the income distribution y_t, with poverty line say z may be written as follows

$$FGT_\alpha(y_t) = \frac{\sum_{i=1}^{n}(z - y_i)^\alpha I(y_i \leq z)}{nz} \tag{9}$$

For choice of $\alpha = 0, 1, 2$, the poverty indices are HCR, PG, and SPG, respectively. Let hcr_t, pg_t, and spg_t denote the growth rate of the indices. Let x_t denotes the mean income of the poor.

Suppose at time point t+1, r number of individuals migrates from a different region. Also suppose $r = r_1 + r_2$, r_1 denote the number of poor and r_2 denotes the number of non-poor. As a result of migration, let the mean income of the poor changes to x_{t+1}. The following proposition relates the growth rate of HCR and PG.

Proposition $hcr_t \geq pg_t \iff \bar{x}_t \geq \bar{x}_{t+1}$.

Proof It can be shown that $PG_t = HCR_t(1 - \bar{x}_t/z)$. Hence, the growth rate may also be related as

$$1 + pg_t = (1 + hcr_t)\theta \tag{10}$$

where $\theta = \frac{z-\bar{x}_{t+1}}{z-\bar{x}_t}$.

Clearly, the necessary and sufficient condition can be proved $\iff$ $\theta \leq 1$.

The equality holds only when $\theta = 1$ or $x_t = x_{t+1}$. Thus if all the r migrants are non-poor, then also the equality holds.

References

Adams, R. (2004). Economic growth, inequality and poverty: Estimating the growth elasticity of poverty. *World Development*, *32*, 1989–2014.

Anselin, L. (2009). *Spatial econometrics: methods and models*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Bourguignon, F. (2003). The growth elasticity of poverty reduction: Explaining heterogeneity across countries and time periods. In T. Eicher & S. Turnovsky (Eds.), *Inequality and growth: Theory and policy implications* (pp. 3–26). Cambridge: MIT Press.

Chambers, D., & Dhongde, S. (2011). A non-parametric measure of poverty elasticity. *Review of Income and Wealth*, *57*, 683–703.

Driscoll, J. C., & Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, *80*, 549–560.

Du, Y., Park, A., & Wang, S. (2005). Migration and rural poverty in china. *Journal of Comparative Economics*, *33*, 688–709.

Foster, J. E., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, *52*, 761–766.

Gundlach, E., de Navarro, P. J., & Weisert, N. (2004). Education is good for the poor: A note on dollar and kraay. In A. Shorrocks & R. van der Hoeven (Eds.), *Inequality and growth: Theory and policy implications*. Oxford: Oxford University Press.

Hirway, I. (2003). Identification of bpl households for poverty alleviation programmes. *Economic and Political Weekly*, *38*, 4803–4038.

Kalwij, A., & Verschoor, A. (2007). Not by growth alone: The role of the distribution of income in regional diversity in poverty reduction. *European Economic Review*, *51*, 805–829.

Pradhan, J. (1993). Drought in kalahandi-the real story. *Economic and Political Weekly*, *28*.

Ram, R. (2007). Roles of income and equality in poverty reduction: Recent cross-country evidence. *Journal of International Development*, *19*, 919–926.

Ravallion, M., & Datt, G. (1998). Why have some indian states done better than others at reducing rural poverty? *Economica*, *65*, 17–38.

Ravallion, M., & Datt, G. (2002). Why has economic growth been more pro-poor in some states of india than others? *Journal of Development Economics*, *68*, 381–400.

Sen, A. K. (1976). Poverty: An ordinal approach to measurement. *Econometrica*, *44*, 219–231.

Smith, K. R. (1993). Fuel combustion, air pollution exposure, and health: The situation in developing countries. *Annual Review of Energy and the Environment*, *18*, 529–566.

Sung, Y. J., & Khagram, S. (2005). A comparative study of inequality and corruption. *American Sociological Review*, *70*, 136–157.

Zaman, K., & Khilji, B. A. (2013). The relationship between growth-inequality-poverty triangle and pro-poor growth policies in pakistan: The twin disappointments. *Economic Modelling*, *30*, 375–393.

Successional Changes in Some Physicochemical Properties on an Age Series of Overburden Dumps in Raniganj Coalfields, West Bengal, India

Santu Malakar and Hema Gupta (Joshi)

Abstract Opencast mining replaces natural vegetation of an area with huge quantities of overburden dumps called mine spoils. The natural plant succession on these dumps improves the dump features after an interval of time. The present work reports successional changes in some physicochemical properties of an age series of overburden dumps (0, 1, 3, 9, 12, 18, 21 years) formed from open-cast mining in Bansra and Sonepur Bazari colliery of Raniganj Coalfields. Dump pH remained near neutral in all the dumps. As the age of the overburden dumps increased, there was an initial increase in clay, silt, water holding capacity, and moisture content. Organic carbon, calcium, potassium, and electrical conductivity also increased with dump age, but the increase of only two parameters—total nitrogen and mineral nitrogen was significantly related to the dump age. No improvement occurred in dump phosphorus concentration with time. Among exchangeable cations, magnesium and sodium decreased, while calcium and potassium increased with age. The net annual accumulation rate of total nitrogen and organic carbon was 12.59 ppm and 300 ppm, respectively. Growth curve of mean dump parameter indicated that the parameter increased with time up to 9 years but decreased afterward. Time period required by the overburden dumps to reach a nearby native forest soil condition was estimated at approximately 19 years. Anthropogenic disturbances in the intermediate successional stages might be responsible for decline in some dump features at 12th year of succession.

Keywords Overburden dumps · Physicochemical properties · Raniganj coalfield Regression analysis · Fourier smoothing

S. Malakar · H. Gupta (Joshi) (✉)
Department of Botany, Visva-Bharati, Santiniketan 731235, West Bengal, India
e-mail: hemagupta.gupta123@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_6

1 Introduction

Coal is one of the prime sources of energy in India. Opencast mining and underground mining are the two types of methods applied for mining of coal. The Raniganj Coalfield is the birthplace of coal mining in India. In opencast mining, the overlying soil and rock debris along with existing vegetation are removed and deposited into another fresh area, called mine spoil or overburden dump. The spoil is a mixture of disintegrated rocks and rocky soils with coal residues. According to Ghose (2004), every year, opencast mining of coal damages a surface area of about 4 ha in India degrading the original potential and quality of soil (Barpanda et al. 2001). Lacking vegetation cover these dump materials become highly prone to erosion by wind or water (Singh et al. 1996; Gairola and Soni 2010) and spread over the surrounding fertile land and disturb their natural quality (Yaseen et al. 2012).

Natural plant succession on these spoils causes changes in physicochemical characteristic of soil leading to restoration and conservation of biodiversity. Overburden topsoil is usually deficient in major nutrients (Yaseen et al. 2012; Juwarkar et al. 2004), so natural succession of plant species on these dumps takes place at a very slow rate (Singh and Jha 1992; Singh et al. 1996; Wali 1987). For successful revegetation, it is necessary to understand the status of dump characteristics. In the present study, our main objective was to study the successional changes in some physicochemical properties of a chronosequence of overburden dumps and comparing with an adjacent native forest.

2 Materials and Methods

2.1 Study Area

The present study has been carried out in Bansra OCP (1, 9, 12, 18, and 21 year dumps) and Sonepur Bazari OCP (3 year dump) of Raniganj Coalfield (Fig. 1) which falls under Eastern Coalfield Limited (ECL). Bansra OCP lies near to Raniganj city between 23° 37′ 38.75″ N and 23° 38′ 52.66″ N latitudes, and 87° 07′ 36.50″ E and 87° 08′ 52.01″ E longitudes. Sonepur Bazari OCP is located in eastern part of Raniganj Coalfields near Pandaweswar between 23° 40′ 58.74″ N and 23° 41′ 47.64″ N latitudes, and 87° 12′ 55.93″ E and 87° 13′ 57.62″ E longitudes. Bansra OCP is about 10 km from Sonepur Bazari OCP. The climate in general is dry tropical with three prominent seasons, summer (middle of March to middle of June), rain (middle of June to middle of October) and winter (November to February). In summer, average temperature ranges between 38 and 43 °C, may rise to 48 °C. The area receives average annual rainfall between 1240 and 1500 mm (Plates 1, and 2).

Garhjungle, a very old natural forest situated in the Burdwan district between 23°40′ 54.4″ N latitude and 87°40′ 20.2″ E longitude is also studied as a native forest for comparative analysis.

Plate 1 A zero-year overburden dump

Plate 2 A three-year overburden dump

2.2 Overburden Sampling and Analysis

Three overburden samples were collected from 15 cm depth at each of the seven dumps and also from the Garhjungle. The samples were properly packed and brought carefully to laboratory for physical and chemical analysis. Moisture content was estimated in the field moist condition by drying 100 g sample to constant weight at 105 °C in a hot air oven (Buresh 1991). The overburden samples were air dried, cleaned, crushed in mortar and pestle and passed through a 10-mesh (2 mm) sieve before physicochemical analysis. Particle size distribution (soil texture) was analyzed by hydrometer method (Bouyoucos 1927). pH and electrical conductivity were determined in 1:2 soil: water ratio (SYSTRONICS 361 pH meter) and 1:5 soil: water ratio (SYSTRONICS 306 conductivity meter). Water holding capacity and bulk density were determined with keens boxes. Organic C was estimated by following wet digestion method (Walkley and Black 1934). Available K and available Na were estimated using flame photometer (Jackson 1958); calcium and magnesium by titrating

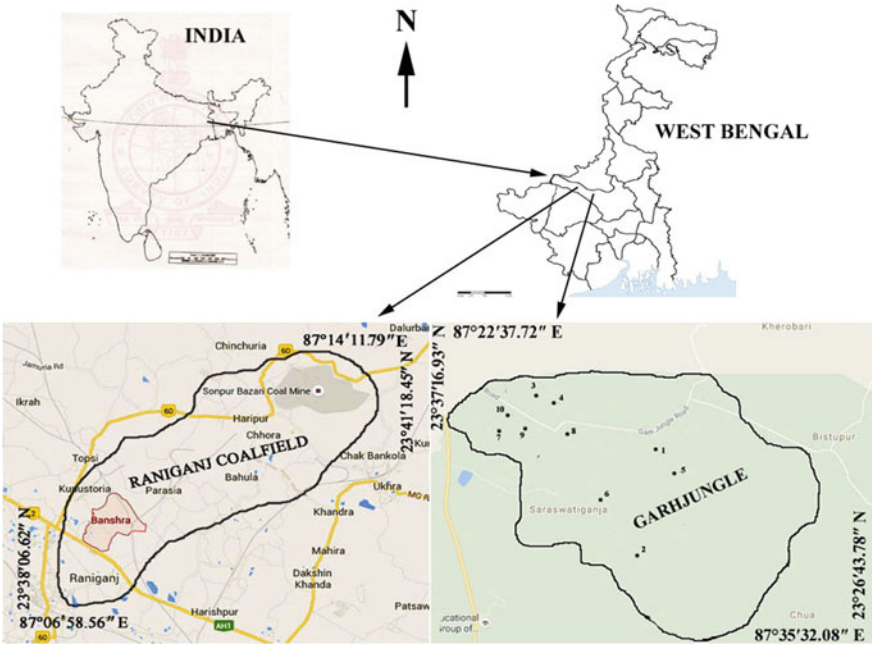

Fig. 1 Location of the study sites. *Source* Survey of India and Google Maps

with EDTA (Baruah and Barthakur 1997). Available P was estimated in $NaHCO_3$ extract of soil by ammonium molybdate–ascorbic acid method (Olsen et al. 1954). Total-N was estimated using Kjeldahl digestion and distillation; mineral-N (NH_4^+–N and NO_3^-–N) was estimated by Kjeldahl distillation of K_2SO_4 extract of soil using Devarda's alloy (Jackson 1958).

2.3 Statistical Analysis

Overburden samples and soil samples were analyzed in triplicates, and result was expressed as mean ± S.E. Correlations were done among physicochemical parameters and also between physicochemical parameters and the dump age. Regression analysis was done to predict the time period required by the overburden dumps to reach the native forest soil condition. Mean dump parameter (Y-axis) was plotted against time (X-axis), and curve smoothing was done with Fourier model using statistical package XLSTAT2016. Natural variability of biological data often masks the underlying true curves. Smoothing is a tool to smooth a time series and make prediction using various methods like moving average smoothing, exponential smoothing, Fourier smoothing. According to Kimball (1974) in Fourier smoothing, the data are smoothed by computing the Fourier transform, setting high-frequency noise com-

ponents of the resulting variance spectrum to zero, and then computing the inverse Fourier transform. The data become dominated by low frequencies of variation which could be separated from the higher-frequency noise. This filtering action makes Fourier transform smoothing more attractive than other methods in extracting a true curve from noisy data (Kimball 1974).

3 Results and Discussions

3.1 Physicochemical Properties of Overburden Dumps

Table 1 displays some physicochemical properties of overburden dumps and the native forest soil. Sand percent was higher in dumps compared to the native forest soil. Silt, clay, and water holding capacity exhibited an increasing trend from 0 to 9 year dumps but decreased from 12 year dump onwards. Similarly sand percent and bulk density decreased up to 9 year dump but again increased from 12 year dump. Increase in clay fraction contributes to the development of soil micropore space reducing the soil bulk density (Ohta and Effendi 1992; Maharana and Patel 2013). Moisture content increased from 6.78% in 0 year dump to 9.13% in 18 year dump but decreased to 7.92% in 21 year dump. Development of vegetation cover is responsible for improvement of moisture content and water holding capacity of the dumps. Electrical conductivity increased from 0 to 21 year dumps. Tripathy et al. (1998) argued higher conductivity to result due to upward migration of salts along with them through cracks or fissures. Dump pH remained near neutral in all the dumps in this study. Previous studies on Raniganj Coalfields have reported increasing trend in pH from acidic to neutral (3.48–6.32) with increase in age of overburden dumps (Biswas et al. 2013); near neutral pH (6.25–6.85) (Yaseen et al. 2012); and alkaline (8.02–8.6) pH (Sadhu et al. 2012). Acidic pH (4.11–5.65) was reported from nearby Jharia Coalfields (Rai et al. 2011). Lovesan et al. (1998) argued that the physicochemical properties of overburden dump materials are site specific and differ due to different geological deposit of rocks.

Total-N ($r = 0.815$) and mineral-N ($r = 0.920$) exhibited increasing trend with the dump age; they are the only parameters showing positive correlation with the dump age. Colonization by leguminous herbs and grasses in early successional stage is responsible for this increase. Available phosphorus remained low in all the dumps as well as in the native forest soil. Among cations calcium and available K increased during succession while available Na and magnesium decreased. Decrease in exchangeable sodium with increase in age of mine spoils is reported earlier (Jha and Singh 1991). Organic C increased from 0.18 to 0.88% during succession from 0 to 21 year dump but the maximum value (1.52%) was attained in the 9 year dump. Trend of increase in organic C with age of mine spoil finds similarity with other reports like Biswas et al. (2013) and Maiti et al. (2002). According to Rai et al. (2011) organic carbon increase in the mine spoils due to gradual accumulation of leaf litter and

Table 1 Physicochemical properties (mean ± SE) of the studied overburden dumps and native forest soil

Soil parameter	Overburden dumps (age in years)							Garhjungle
	0	1	3	9	12	18	21	
Mineral-N (ppm)	22.09 ± 3.64	44.45 ± 4.02	50.77 ± 1.90	63.51 ± 4.31	72.77 ± 9.64	71.46 ± 8.32	88.05 ± 11.33	70.96 ± 3.86
Total-N (ppm)	171.11 ± 20.58	202.22 ± 33.90	388.89 ± 51.00	326.67 ± 40.42	482.22 ± 87.65	497.7833.90	435.56 ± 43.31	482.22 ± 74.20
Moisture (%)	6.78 ± 1.70	9.00 ± 1.99	8.46 ± 1.75	11.78 ± 2.55	8.15 ± 2.16	9.13 ± 1.81	7.92 ± 2.16	5.59 ± 1.39
pH	7.22 ± 0.24	6.84 ± 0.17	7.08 ± 0.22	6.42 ± 0.16	6.90 ± 0.18	6.84 ± 0.29	7.07 ± 0.22	5.67 ± 0.23
Electrical conductivity ($\mu s\ cm^{-1}$)	216.41 ± 30.78	168.39 ± 9.87	229.89 ± 13.67	204.42 ± 40.41	279.09 ± 27.90	194.42 ± 27.46	273.13 ± 18.25	113.34 ± 14.36
Organic C (%)	0.18 ± 0.02	0.32 ± 0.07	0.64 ± 0.15	1.52 ± 0.30	0.54 ± 0.11	1.02 ± 0.16	0.88 ± 0.09	0.56 ± 0.10
Available Na (ppm)	6.33 ± 1.44	8.11 ± 2.55	4.89 ± 1.06	6.56 ± 1.38	10.78 ± 3.74	4 ± 0.24	3.33 ± 0.29	3.33 ± 0.33
Available K (ppm)	5.44 ± 0.41	7.56 ± 0.90	13.56 ± 5.06	8.33 ± 0.85	10.89 ± 1.53	10 ± 1.12	15.22 ± 1.71	10.78 ± 1.65
Available P (ppm)	0.18 ± 0.01	0.16 ± 0.01	0.18 ± 0.01	0.18 ± 0.02	0.19 ± 0.01	0.18 ± 0.02	0.18 ± 0.02	0.18 ± 0.01
Calcium (ppm)	214.13 ± 18.78	285.02 ± 15.01	261.56 ± 21.10	285.51 ± 30.73	195.56 ± 22.40	289.91 ± 32.95	293.82 ± 27.82	61.6 ± 7.41
Magnesium (ppm)	45.47 ± 3.78	68.05 ± 4.84	46.64 ± 5.13	53.68 ± 8.49	41.36 ± 6.64	39.89 ± 3.87	43.71 ± 9.87	14.67 ± 3.47
Water holding capacity (%)	35.20 ± 1.30	41.18 ± 1.12	39.02 ± 1.73	39.4 ± 1.90	36.27 ± 2.04	38.45 ± 4.75	37.74 ± 1.50	27.41 ± 1.13
Bulk density ($g\ cm^{-3}$)	1.13 ± 0.02	1.1 ± 0.03	1.10 ± 0.03	1.12 ± 0.03	1.19 ± 0.03	1.13 ± 0.04	1.15 ± 0.04	1.34 ± 0.03
Clay (%)	19.38 ± 1.39	23.73 ± 1.28	23.11 ± 1.71	29.85 ± 2.62	21.94 ± 1.59	23.00 ± 1.92	20.22 ± 0.93	21.83 ± 0.92
Silt (%)	8.91 ± 0.94	10.02 ± 0.51	12.81 ± 0.94	10.95 ± 0.94	9.11 ± 0.91	10.38 ± 1.19	9.18 ± 0.77	13.66 ± 1.32
Sand (%)	71.71 ± 1.58	66.26 ± 1.37	64.08 ± 2.48	59.2 ± 3.16	68.95 ± 2.4	66.62 ± 2.16	70.61 ± 0.88	64.51 ± 1.37

its decomposition to form humus and vice versa. In a previous study involving a single season overburden collection from similar sites, we have reported increase in electrical conductivity, moisture content, water holding capacity, and organic carbon content with increase in age of overburden dumps (Malakar et al. 2015).

Correlation coefficients (r) indicated significant correlations among some physicochemical parameters. For instance, moisture content was positively related with calcium (0.750), magnesium (0.712), water holding capacity (0.770), and clay (0.764), while negatively with bulk density (−0.739); clay content had positive correlation with organic C (0.745) and negative with sand (−0.906) and available P (−0.722); bulk density had negative correlation with pH (−0.801), calcium (−0.920), magnesium (−0.876) and water holding capacity (−0.944); calcium positively with magnesium (0.802) and water holding capacity (0.949); mineral-N positively with total-N (0.850) and available K (0.737); and pH positively with electrical conductivity (0.746). Positive relationship between clay and organic C has been reported extensively (Maharana and Patel 2013; Roberts et al. 1981; Marrs et al. 1981). Clay acts as absorption sink for organic material (Marshman and Marshall 1981) and protects against decomposition (Dixon 1989; Van Veen and Kuikman 1990) thereby increasing organic carbon level.

3.2 Comparison of Dumps with Native Forest Soil

Comparing 0 year dump with the native forest soil, it was found that the mineral-N was 3.2 times, total-N was 2.8 times, organic C was 3.1 times, available K was 2 times, silt was 1.5 times, clay was 1.1 times, and bulk density was 1.2 times higher in the native forest soil. Parameters like available Na, calcium, magnesium, sand, water holding capacity, moisture, pH, and electrical conductivity were lower in native forest soil than the 0 year dump. Comparison of 21 year dump with the native forest soil indicates that the native forest soil was acidic and has more total-N and more silt and clay than the 21 year dump; however, parameters like sand, moisture, electrical conductivity, organic C, mineral-N, calcium, magnesium, water holding capacity, and bulk density were lower in the native forest soil than 21 year dump. Available P and available Na content were low and same in both. In a recent report from Raniganj Coalfield, exchangeable cations and electrical conductivity increased with dump age and exceeded the native forest soil values, but organic C, N, and P remained lower than native forest soil values even after 21 years of succession (Kumar et al. 2015).

According to Brady (1984), various physicochemical parameters including soil depth, organic matter, and N-concentration increase during soil development. Comparative analysis of spoil features between 0 and 21 year dumps indicate that mineral-N was 4 times, total-N was 2.5 times, organic C was 4.9 times, available K was 2.8 times, calcium was 1.4 times, electrical conductivity was 1.3 times, and water holding capacity was 1.1 times higher in 21 year dump than 0 year. Mineral-N indicated an increase of 65.96 ppm from 0 to 21 year dump, total-N an increase of 264.44 ppm, organic C an increase of 0.7% or 7 mg/g spoil, available K an increase

of 9.77 ppm, and calcium an increase of 79.69 ppm. The net annual accumulation rate was 3.15 ppm for mineral-N, 12.59 ppm for total-N, 0.3 mg/g or 300 ppm for organic C, 0.47 ppm for available K, and 3.8 ppm for calcium. Maharana and Patel (2013) reported net annual accumulation of 200 ppm for organic C and 16.1 ppm for total-N in Mahanadi coalfields, Odisha. Jha and Singh (1991) reported 26.5% increase in total-N in 15 years from Jhingurda Colliery, Madhya Pradesh. Lower values of nitrogen accumulation have been reported from mine spoils of temperate climate (Visser et al. 1983; Jencks et al. 1982). According to Dancer et al. (1977), nitrogen accumulation of 700 kg ha^{-1} is critical for establishment of a substantial self-sustaining ecosystem in disturbed lands.

The absence of phosphorus accumulation in the dumps is in contrast to other reported works where 0.3–0.8 ppm annual accumulation of available P occurred (Maharana and Patel 2013; Jha and Singh 1991). As the area falls under lateritic belt, the leaching processes, removing silica and a large portion of the bases originally present in the rock and soil leaving a residue rich in iron and aluminum and poor in potash and phosphoric acid (Raychaudhuri 1980), might be responsible for this observation. Phosphorus is the soil nutrient that frequently limits tree growth and productivity in soils having laterization as the main pedogenic process (Cleveland et al. 2002). Another reason might be the poor development of mycorrhiza and poor colonization of other phosphorus solubilizing microbes on the overburden dumps.

3.3 Influence of Age on Overburden Dumps

Following the approach of Maharana and Patel (2013), an attempt was made to calculate the time period required for reclamation of 0 year dump to reach the native forest soil condition on basis of changes in soil parameters. Some parameters increase while others decline during succession. Changes in parameters like nitrogen, phosphorus, and organic matter content are considered more important drivers of successional changes in vegetation. However, in this particular analysis, only those parameters which exhibited an increasing trend in succession or showed a positive correlation with age were considered for estimation of mean value. Soil parameters of native forest Garhjungle were considered as unit; proportionate amount of these parameters were calculated for overburden dumps and averaged to obtain the mean dump parameter. Mean dump parameter varied in polynomial trend with the dump age increasing up to 9 years and then decreasing afterward (Fig. 2). Change in dump age explained 97.9% variation in the mean dump parameter. Mean dump parameter approached the value of native forest soil (i.e. 1) initially at approximately 6 years and then at approximately 19 years. It is worthwhile to mention here that Garhjungle is nearly 50 km apart from the dump sites, and has been reported as a disturbed forest on nutrient-poor soils (Ganguli et al. 2016a, b).

An ordered development of dump material occurs in several situations (Roberts et al. 1981; Marrs et al. 1981), however, in some cases, succession may stagnate at a given stage (Schafer and Nielsen 1979). In the present study, many dump features

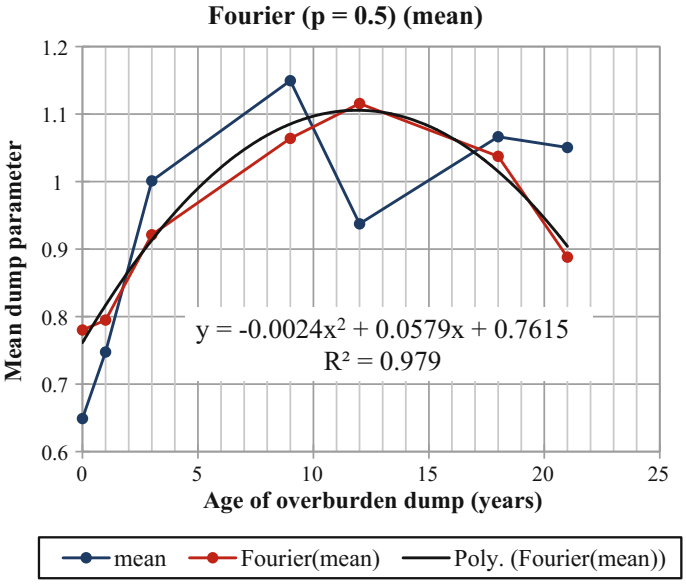

Fig. 2 Regression of mean dump parameters with the age of overburden dumps

like clay, silt, organic C, and moisture content increased initially up to 9 years but decreased at 12th year of succession. This decrease is responsible for the gap between mean curve and Fourier smoothed curve at the fifth data point (Fig. 2). By that time the dumps became revegetated with many herb and shrub species, some small-sized trees also appeared (Malakar et al. 2015). Anthropogenic activity as collection of fuelwood might be responsible for decline in dump parameters 12 year onwards as decline in vegetation cover leads to erosion of soil. Another possibility is the contamination of different dumps with freshly mined spoil.

4 Conclusions

From above results, it is revealed that many dump features improved with age, particularly the nitrogen concentration increased significantly with age. Nitrogen as well as phosphorus is the most important nutrient that influences vegetation or ecosystem development as a whole. No improvement in the phosphorus concentration of dumps occurred in the present study. Many dump parameters improved initially but started declining from the intermediate stage of succession. These observations emphasize the need of proper conservation of the overburden dumps so that soil improvement and natural vegetation succession can proceed without interruption. Further analysis of soil microbial processes like nutrient mineralization and immobilization leading

to ecosystem development will throw more light on improvement of dump features and succession of vegetation.

Acknowledgements Authors thank the Raniganj Coalfields (Eastern Coalfield Limited) for secondary information regarding overburden dumps. We also thank the anonymous reviewer for providing constructive comments to improve the manuscript and Prof Ratan Dasgupta for editorial guidance. First author acknowledges the financial assistance from UGC, New Delhi, in the form of Rajiv Gandhi National Fellowship.

References

Barapanda, P., Singh, S. K., & Pal, B. K. (2001). Utilization of coal mining wastes: An overview. In *National Seminar on Environmental Issues and Waste Management in Mining and Allied Industries*, Regional Engg College, Rourkela, Orissa, India (pp. 177–182).

Baruah, T. C., & Barthakur, H. P. (1997). *A textbook of soil analysis*. New Delhi: Vikas Publishing House Pvt. Ltd.

Biswas, C. K., Mukherjee, A., & Mishra, S. P. (2013). Physico-chemical properties of overburden dumps of different ages at Sonepur Bazari coalmine area, Raniganj, West Bengal (India). *The Ecoscan, 7*(1&2), 57–60.

Bouyoucos, G. J. (1927). The hydrometer as a new method for the mechanical analysis of soils. *Soil Science, 23,* 343–353.

Brady, N. C. (1984). *The nature and properties of soils*. New York: Macmilan.

Buresh, R. J. (1991). Extraction of ammonium, nitrate and nitrite from soil. Collaborative Project IFDC/IRRI.

Cleveland, C. C., Townsend, A. R., & Schmidt, S. K. (2002). Phosphorus limitation of microbial processes in moist tropical forests: Evidence from short-term laboratory incubations and field studies. *Ecosystems, 5,* 680–691.

Dancer, W. S., Handley, J. F., & Bradshaw, A. D. (1977). Nitrogen accumulation in kaolin mining wastes in cornwall. *Plant and Soil, 48,* 303–314.

Dixon, J. B. (1989). *Minerals in soil environments* (2nd ed.). USA: Soil Science Society of America.

Gairola, S. U., & Soni, P. (2010). Role of soil physical properties in ecological succession of restored mine land—A case study. *International Journal of Environmental Science, 1*(4), 475–480.

Ganguli, S., Gupta (Joshi), H., & Bhattacharya, K. (2016a). Vegetation structure and species diversity in Garhjungle sacred forest, West Bengal, India. *International Journal of Environmental & Agriculture Research 2*(9): 72–79.

Ganguli, S., Gupta (Joshi), H., & Bhattacharya, K. (2016b). Soil N-transformation rates in two differently managed dry deciduous forests of West Bengal, India. *World Journal of Research and Review, 3*(4), 45–49.

Ghose, M. K. (2004). Effect of opencast mining on soil fertility. *Journal of Environment and Industrial Research, 63,* 1006–1009.

Jackson, M. L. (1958). *Soil chemical analysis*. Eaglewood, Cliffs, New Jersey: Prentice Hall Inc.

Jencks, E. M., Tryon, E. H., & Contri, M. (1982). Accumulation of nitrogen in mine soils seeded to black locust. *Soil Science Society of America Journal, 46,* 1290–1293.

Jha, A. K., & Singh, J. S. (1991). Spoil characteristics and vegetation development of an age series of mine spoils in a dry tropical environment. *Vegetatio, 97,* 63–76.

Juwarkar, A. A., Jambulkar, H. P., & Singh, S. K. (2004). Appropriate strategies for reclamation and revegetation of coal mine spoil dumps. In *Proceedings of the National Seminar on Environmental Engineering with Special Emphasis on Mining Environment*, Institute of Public Health and Engineers, India (pp. 1–9).

Kimball, B. A. (1974). Smoothing data with Fourier transformations. *Agronomy Journal, 66,* 259–262.

Kumar, S., Maiti, S. K., & Chaudhuri, S. (2015). Soil development in 2–21 years old coalmine reclaimed spoil with trees: A case study from Sonepur-Bazari opencast project, Raniganj Coalfield, India. *Ecological Engineering, 84,* 311–324.

Lovesan, V. J., Kumar, N., & Singh, T. N. (1998). Effect of the bulk density on the growth and biomass of the selected grasses over overburden dumps around coal mining areas. In *Proceedings of the 7*th *National Symposium on Environment*, Dhanbad, Jharkhand, India (pp. 182–185).

Maharana, J. K., & Patel, A. K. (2013). Physico-chemical characterization and mine soil genesis in age series coal mine overburden spoil in chronosequence in a dry tropical environment. *Journal of Phylogen Evolution Biology, 1,* 101. https://doi.org/10.417/2329-9002.1000101.

Maiti, S. K., Karmakar, N. C., & Sinha, I. N. (2002). Studies on some physical parameters aiding biological reclamation of mine spoil dump—A case study from Jharia coalfield. *IME Journal, 41*(6), 20–23.

Malakar, S., Gupta (Joshi), H., & Kumar, M. L. (2015). Species composition and some physico-chemical properties of an age series of overburden dumps in Raniganj Coalfields, West Bengal, India. *International Journal* of *Scientific Research in Environmental Sciences, 3*(7), 0239–0247.

Marrs, R. H., Roberts, R. D., Skeffington, R. A., & Bradshaw, A. D. (1981). Ecosystem development on naturally colonized china clay wastes: II Nutrient Compartmentation. *Journal of Ecology, 69,* 163–169.

Marshman, N. A., & Marshall, K. C. (1981). Bacterial growth on proteins in the presence of clay minerals. *Soil Biology & Biochemistry, 13,* 127–134.

Ohta, S., & Effendi, S. (1992). Ultisol of "lowland *Dipterocarp* forest" in east Kalimantan, Indonesia. *Soil Science and Plant Nutrition, 38,* 197–206.

Olsen, S. R., Cole, C. V., Watanabe, F. S., & Dean, L. A. (1954). Estimation of available phosphorus in soils by extraction with sodium bicarbonate. US Department of Agriculture Circular 939.

Rai, A. K., Paul, B., & Singh, G. (2011). A Study on Physico chemical properties of over burden dump materials from selected coal mining areas of Jharia coalfields, Jharkhand, India. *International Journal of Environmental Science, 1*(6), 1350–1359.

Raychaudhury, S. P. (1980). The occurrence, distribution, classification and management of laterite and lateritic soil. Cah ORSTOM, series. *Pedology, 18*(3–4), 249–252.

Roberts, R. D., Marrs, R. H., Skeffington, R. A., & Bradshaw, A. D. (1981). Ecosystem development on naturally colonized china clay wastes: I. vegetation changes and overall accumulation of organic matter and nutrients. *Journal of Ecology, 69*, 15–11.

Sadhu, K., Adhikari, K., & Gangopadhyay, A. (2012). Effect of mine spoil of Lower Gondwana coal fields: Raniganj coal mines area, India. *International Journal of Environmental Science, 2*(3), 1675–1687.

Schafer, W. M., & Nielsen, G. A. (1979). Soil development and plant succession on 1- to 50-year-old strip mine spoils in southeastern Montana. In M. K. Wali (Ed.), *Ecology and coal resource development* (Vol. 2, pp. 541–549). New York: Pergamon Press.

Singh, J. S., & Jha, A. K. (1992). Restoration of degraded land: An overview. In J. S. Singh (Ed.), Restoration of degraded land: Concepts and strategies (p 17). Rastogi Publications, Meerut, India.

Singh, R. S., Chaulya, S. K., Tewary, B. K., & Dhar, B. B. (1996). Restoration of a coal-mine overburden dumps—A case study. *Coal International* 80–83.

Tripathy, D. P., Singh, G., & Panigrahi, D. C. (1998). *Proceedings of the Seventh National Symposium on Environment*, ISM Dhanbad, Feb 5–7, 205.

Van Veen, J. A., & Kuikman, P. J. (1990). Soil structural aspects of decomposition of organic matter by micro-organisms. *Biogeochemistry, 11,* 213–233.

Visser, S., Griffiths, C. L., & Parkinson. (1983). Effects of surface mining on the microbiology of a prairie site in Alberta, Canada. *Canadian Journal of Soil Science, 63*, 177–189.

Wali, M. K. (1987). The structure, dynamics and rehabilitation of drastically disturbed ecosystems. In T. N. Khoshoo (Ed.), *Perspectives in environmental management* (pp. 163–183). New Delhi: Oxford Publications.

Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining organic carbon in soils: Effect of variations in digestion conditions and of inorganic soil constituents. *Soil Science, 63,* 251–263.

Yaseen, S., Pal, A., Singh, S., & Dar, I. Y. (2012). A Study of physico-chemical characteristics of overburden dump materials from selected Coal mining areas of Raniganj Coal Fields, Jharkhand, India. *Global Journal of Science Frontier Research, 2,* 6–13.

Growth and Nutritional Status of Preschool Children in India

Susmita Bharati, Manoranjan Pal, Soumendu Sen and Premananda Bharati

Abstract The growth study of children is a proxy to health study because it gives valuable information on health and nutritional status of a child. This paper investigates growth and nutritional status of preschool children using data of National Family Health Survey (NFHS-4) collected during 2015–16. The sample size is 205935 children of age 0–59 months. The main objective of the study is to find sex and age-group-wise growth and nutritional status of preschool children in India. We have also investigated the impact of socioeconomic variables on children's nutritional status. Growth study has been done through height and weight, and for nutritional assessment, three age- and sex-specific Z scores, namely weight-for-age, height-for-age, and weight-for-height, have been taken following WHO (2006). The covariates are places of residence, religion, mother's educational status and nutritional status and wealth index of the family. The result shows that there is a positive growth at all the age groups, and maximum increment is seen from 12- to 35-month-old children. Percentage of underweight was maximum among boys in the age range of 9–23 months and among girls in the age range of 12–35 months. It is also seen that large percentages of underweight or stunted children occur from 6 months to second year of life. At present in India, 34.6% boys and 33.3% girls are underweight. Likewise, 38.4% boys and 36.7% girls are stunted. And 21.4% boys and 19.5% girls are wasted. So, boys are suffering from undernutrition a little bit more than girls. Percentages of underweight and stunting have decreased only by 10% when compared with NFHS-3 data which was collected about 10 years back. During this period, the prevalence of wasting has remained almost the same. The findings suggest that reductions in stunting and other forms of undernutrition are possible through proven interventions on mothers' literacy status, mothers' nutrition level, and economic status of the households. It may be recommended that exclusive breastfeeding, safe, appropriate, and high-quality complementary food with micronutrient are essential for proper growth and nutrition of preschool children.

S. Bharati · M. Pal · P. Bharati (✉)
Indian Statistical Institute, 203 B.T. Road, Kolkata, India
e-mail: pbharati@gmail.com

S. Sen
International Institute for Population Sciences, Mumbai, India

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_7

Keywords Growth status · Undernutrition · Socioeconomic condition

1 Introduction

Growth is a proxy to child health, so if the child grows well, it means that the child has a sound health. On the other hand, ill health is often associated with poor growth. Thus, growth study is very important in order to assess health situation of preschool children. The term, 'growth' of a child is often taken to include both size and velocity. Growth standard may thus relate to attainment in size compared to age and may not directly measure growth velocity. It represents a complex interaction of nutritional intake and absorption, which vary within and between populations. So growth study is valuable as it provides good information on the health and nutritional status of a community. It also helps us to improve the health condition of a child. A well-designed growth study can be used as a powerful tool to monitor the health and nutritional status of any community.

In developing countries like India, weight, height, and age are the three main anthropometric measures that are widely used for assessing the growth and nutritional status. The most important indices for assessing the nutritional status of the children in the community, namely weight-for-age, height-for-age, and weight-for-height, are based on these three anthropometric measures (WHO 1986). However, for proper assessment, the sex of the child should also be taken into consideration.

It is known that the percentage of undernourished children is the highest in India among all countries in the world. And still, it is the biggest unresolved problem after the largest food supplementation programs for children introduced through Integrated Child Development Service (ICDS) and midday meal programs. The problem aggravates due to poor child feeding practices and poor access to healthcare system (Dasgupta et al. 2005). One important factor is possibly the unequal intra-household food distribution within the family which has adverse effect on preschool children (NNMB 2006). Some socioeconomic factors like gender, place of residence, parent's educational status (mainly mother's education), religion, and caste have significant impact on the incidence of undernutrition among children (Gragnolati et al. 2005; Som et al. 2006; Bharati et al. 2008a, b).

One may now be curious to know about the present condition of the undernutrition among the preschool children in India so far as these indices are concerned. The present study tries to achieve the same. More precisely, the present study tries to give answer to the following questions.

i. What is the status of growth of 0–59-month-old children measured through weight and height separately for each sex?
ii. What is the nutrition situation of these children by age group, sex, state, and zone?
iii. What is the effect of socioeconomic variables on children's nutritional status?

2 Methodology

We have, in this paper, used the latest data on growth and nutritional status of children. International Institute for Population Sciences (IIPS) coordinated the Fourth National Family Health Survey (NFHS–4), which was carried out in 2015–16 all over India. The NFHS-4 sample covers Indian women of age 15–49 years living in all states and union territories. We have, however, used data of those mothers, who have the last child within the last five years of the data collection. Anthropometric variables, like weight and height of the children, have been considered for growth and nutritional status. The nutritional indices, i.e., 'Z' scores of weight-for-age, height-for-age, and weight-for-height, have been used for nutritional assessment. Z score value '−2' was used as a cutoff point for prevalence estimation (WHO, 2006). Z score is defined as the deviation of the value observed for an individual from the median of the reference population, divided by the standard deviation (SD) of the reference population, i.e.,

$$\text{Z} - \text{score} = \frac{(\text{observed value}) - (\text{median of the reference population})}{(\text{SD of the reference population})}.$$

The classifications of Z score (followed by NCHS/WHO) are below normal if $Z < -2$, normal if $-2 \leq Z < +2$ and above average if $Z \geq +2$.

Weight-for-height (WHZ) index is an indicator of thinness or wasting. Wasting is a short-term malnutrition which arises due to acute starvation, severe disease, famine, etc., but it may result also from a chronic dietary deficiency or disease. Height-for-age (HAZ) is an indicator of stunting which means chronic malnutrition of health, but genetic factor is also related with it. The third index, weight-for-age (WAZ), is primarily a composite index of HAZ and WHZ, i.e., the indicator of both acute and chronic malnutrition. In the young children, low weight-for-age reflects low weight-for-height, but in the later period, it reflects low height-for-age. The covariates are sectors such as rural/urban, religion, mother's educational status, and nutritional status of mother have measured through BMI and wealth index of the family. Here, the representative number of children is 205,935 of ages between 0 and 59 months.

Computation of anthropometric indices was done by using sex, age, weight, and height data on each individual. It was carried out by the SPSS version 18.0. Mother's BMI was computed by using the formula wt/ht^2 (weight is in kg. and height is in meter). A BMI less than 18.5 indicates chronic energy deficiency or undernutrition (WHO 2000).

To draw the relative and effective intervention, the risk of Z score value being less than –2 (i.e., undernourished) has been related with the socioeconomic variables using categorical logistic regression analysis. The nutritional status of children has been considered as dependent variable, and also, the socioeconomic variables are considered as independent variables. Children whose Z scores are below –2 are coded as '1', and those with Z scores –2 or higher are coded as '0'. An estimated odd ratio of '1' indicates that the nature of dependent variable is not different from the reference category. If the estimated odd ratio is >1, the probability of becoming undernourished is more in this category compared to the reference category, and if

Table 1 Mean weight and height of 0–59-month children in India

Age group (mo.)	Weight				Height		
	N	Mean	SD	't'	Mean	SD	't'
00–02	7500	4.09	0.98	–	54.39	4.45	–
03–05	11349	5.84	1.09	111.53**	61.69	4.58	108.29**
06–08	11632	7.03	1.13	80.59**	66.89	4.40	87.73**
09–11	10757	7.68	1.19	41.89**	70.08	4.51	53.59**
12–23	40140	8.96	1.47	82.80**	76.29	5.53	107.33**
24–35	39983	10.86	1.71	168.87**	85.35	5.97	222.64**
36–47	42913	12.43	1.89	125.11**	92.25	6.19	163.10**
48–59	41661	13.95	2.09	110.62**	98.61	6.18	149.55**

it is <1, then it is just opposite to that of '>1' case. It was done by 18.0 version of Statistical Package for Social Science (SPSS). Significance levels of $p<0.01$, 0.05 and 0.1 were considered.

3 Results

The total age group of 0–59-month children has been categorized into eight subgroups such as 0–2, 03–05, 06–08, 09–11, 12–23, 24–35, 36–47, and 48–59 months. Mean and standard deviation of weight and height of 0–59-month children, by age group and sex, are presented in Tables 1, 2a, b. To test the significance of the differences of mean values between two subsequent age groups, 't' test has been used for each subgroup. A positive change been seen for each pair of subsequent age groups, and the maximum increment is seen from 12–23 to 24–35 age group. And all these increments are statistically significant at 1% level.

Tables 3, 4, 5 reflect the incidence of undernutrition by age group for the three different types of nutritional assessment scales, namely weight-for-age, height-for-age, and weight-for-height by sex (WHO 2006). It is seen from Table 3 that undernutrition (through weight-for-age) is uniformly in ascending order and the maximum difference is observed from 9–11 months to 12–23 months among boys. This may be due to weaning. It is the period when the incidence of proper introduction of solid food and major immunization should be completed. So it directly reflects the combination of nutrition and immunization status with other consequences. Among girls, the trend is more or less same, but maximum increment happens between 12–23- and 23–35-month subgroups. It is also noticed that, at the initial point, boys are more underweight than girls, but here at the optimal point, i.e., 48–59-month group, girls are more underweight than boys.

Table 4 shows percentage of nutritional status through height-for-age by sex. Here, it is seen that among boys, percentage of stunting is ascending in nature over

Table 2 Mean weight (a) and height (b) of 0–59-month children in India for boys and girls

Age group (mo.)	Boys				Girls			
	N	Mean	SD	't'	N	Mean	SD	't'
(a) Mean Weight								
00–02	3864	4.21	1.03	–	3636	3.97	0.91	–
03–05	5746	6.08	1.10	83.45**	5603	5.59	1.02	77.51**
06–08	6117	7.29	1.11	52.29**	5515	6.74	1.09	56.89**
09–11	5672	7.94	1.17	30.85**	5085	7.39	1.16	29.93**
12–23	21001	9.22	1.45	61.22**	19139	8.67	1.45	58.29**
24–35	20973	11.09	1.69	122.01**	19010	10.61	1.69	119.91**
36–47	22457	12.65	1.87	91.12**	20456	12.19	1.88	87.39**
48–59	22105	14.18	2.07	81.49**	19556	13.68	2.07	75.62**
(b) Mean Height								
00–02	3864	54.73	4.50	–	3636	54.02	4.37	–
03–05	5746	62.36	4.56	80.83**	5603	60.99	4.49	73.25**
06–08	6117	67.57	4.30	63.99**	5515	66.12	4.38	60.91**
09–11	5672	70.76	4.29	40.17**	5085	69.32	4.31	36.58**
12–23	21001	76.91	5.36	79.73**	19139	75.61	5.63	73.40**
24–35	20973	85.84	5.78	164.10**	19010	84.79	6.14	152.41**
36–47	22457	92.77	6.05	121.78**	20456	91.68	6.29	109.82**
48–59	22105	99.11	6.07	110.23**	19556	98.06	6.24	101.71**

age, though the highest magnitude of stunting (15.3%) occurred between 9–11 and 12–23-month group. And after that, it is zigzag in nature for both boys and girls.

Table 5 shows that the percentage of wasting children decreases with age for both boys and girls. It means that the children's health status is improving with the age. During first two months after birth, 31.5% children of both sexes are at risk of wasting. This may be due to low birth weight or other adverse conditions just after birth. With the adoption of better environmental condition, they cope up with it.

It thus reflects from the above-mentioned tables that greater magnitude of chronic undernutrition starts from 6 months and lasts up to 23 months. One may relate this to stoppage of breastfeeding and weaning practices. Large percentages of chronic undernutrition like underweight or stunting exist till the second year of life, and then, it becomes stable. This trend is same with the third NFHS data (Sen et al. 2011).

Table 6 shows the different categories undernutrition among 0–59-month children in India by states and union territories and by sex. It is seen that 34.6% boys and 33.3% girls are underweight in India. Likewise, 38.4% boys and 36.7% girls are stunted. And 21.4% boys and 19.5% girls are wasted. So, at present, boys are suffering from undernutrition more than that of girls. Among 0–59-month children, percentage of underweight is 34.0, stunted is 37.6, and wasted is 20.5 (not shown in the tables). Zone-wise, the lowest percentage of underweight, stunted, and wasted children are seen in northeast zone for both boys and girls, while the highest percentages of

Table 3 Percentage of underweight, normal, and overweight children of age 0–59 months by age and sex according to weight-for-age?

Age groups (months)	Boys				Girls			
	N	Underweight (< −2SD)	Normal (−2SD − <+2SD)	Overweight (≥ +2SD)	N	Underweight (< −2SD)	Normal (−2SD<+2SD)	Overweight (≥ +2SD)
00–02	3864 (100.0)	26.2	72.2	1.6	3636 (100.0)	20.4	78.4	1.2
03–05	5746 (100.0)	28.5	70.5	1.0	5603 (100.0)	26.1	72.7	1.1
06–08	6117 (100.0)	28.2	71.2	0.7	5515 (100.0)	23.7	75.3	1.0
09–11	5672 (100.0)	31.8	67.5	0.7	5085 (100.0)	27.6	71.8	0.7
12–23	21001 (100%)	35.4	63.8	0.7	19139 (100.0)	30.5	68.8	0.6
24–35	20973 (100.0)	36.4	63.1	0.5	19010 (100.0)	35.2	64.3	0.6
36–47	22457 (100.0)	36.5	63.1	0.4	20456 (100.0)	37.5	62.2	0.3
48–59	22105 (100.0)	35.7	64.0	0.3	19556 (100.0)	38.5	61.3	0.2

Table 4 Percentage of stunted, normal, and above normal children of age 0–59-month by age and sex according to height-for-age

Age groups (months)	Boys				Girls			
	N	Stunted (< −2SD)	Normal (−2SD–<+2SD)	Above normal (≥ +2SD)	N	Stunted (< −2SD)	Normal (−2SD–<+2SD)	Above normal (≥ +2SD)
00–02	3864 (100.0)	22.2	70.3	7.5	3636 (100.0)	17.1	74.1	8.8
03–05	5746 (100.0)	22.4	70.4	7.2	5603 (100.0)	18.8	72.8	8.3
06–08	6117 (100.0)	23.3	69.5	7.2	5515 (100.0)	18.4	74.1	7.5
09–11	5672 (100.0)	28.8	65.9	5.3	5085 (100.0)	24.1	69.4	6.5
12–23	21001 (100%)	44.1	52.5	3.3	19139 (100.0)	38.6	57.5	4.0
24–35	20973 (100.0)	42.4	54.9	2.7	19010 (100.0)	40.3	56.2	3.5
36–47	22457 (100.0)	43.0	55.5	1.5	20456 (100.0)	43.9	54.4	1.7
48–59	22105 (100.0)	38.3	61.0	0.7	19556 (100.0)	40.9	58.5	0.7

Table 5 Percentage of wasted, normal, and overweight children of age 0–59 month by age and sex according to weight-for-height

Age groups (months)	Boys				Girls			
	N	Wasted (< −2SD)	Normal (−2SD–<+2SD)	Overweight (≥ +2SD)	N	Wasted (< −2SD)	Normal (−2SD<+2SD)	Overweight (≥ +2SD)
00–02	3864 (100.0)	31.5	60.7	7.9	3636 (100.0)	31.5	62.3	6.2
03–05	5746 (100.0)	28.8	65.3	6.0	5603 (100.0)	27.8	66.4	5.7
06–08	6117 (100.0)	27.1	68.5	4.4	5515 (100.0)	24.9	71.8	3.3
09–11	5672 (100.0)	26.8	69.8	3.4	5085 (100.0)	24.2	72.5	3.3
12–23	21001 (100.0	22.1	75.6	2.4	19139 (100.0)	19.8	78.0	2.2
24–35	20973 (100.0)	20.5	78.0	1.5	19010 (100.0)	17..9	80.7	1.4
36–47	22457 (100.0)	18.7	79.8	1.5	20456 (100.0)	16.2	82.2	1.6
48–59	22105 (100.0)	17.5	80.8	1.7	19556 (100.0)	16.9	81.3	1.8

underweight and stunted children are seen in central zone. The highest percentage of wasted children are found in west zone for both boys and girls. The states with more than 40% occurrence of underweight children are Jharkhand (48.1% 47.7%, respectively, for boys and girls), Madhya Pradesh (43.9% for boys and 42.0% for girls), Bihar (43.1% for boys and 44.6% for girls), Gujarat (42.3% for boys), Chhattisgarh and Gujarat (40.7% and 42.3% for boys). In case of stunting, the percentages, respectively, for boys and girls, are (47.0, 47.6) in Bihar, (45.6, 43.1) in Jharkhand, (45.5, 41.5) in Meghalaya, (44.8, 44.9) in Uttar Pradesh and (43.8, 41.0) in Dadra and Nagar Haveli. In case of wasting, more than 30% occurrences are found in the state Jharkhand (30.8) and that is for boys only.

Now, we turn to Table 7 to see the incidence of undernutrition by socioeconomic characteristics. It is seen that rural children are more underweight, stunted, and wasted for both boys and girls. Religion-wise, Hindus have the highest percentage of underweight children than those of Muslims, Christians, and other religions, both boys and girls. In case of stunting, Hindus and Muslims have higher percentage than national average, while in the case of wasting, only Hindus are higher than national average for both boys and girls. In the case of mother's education, all categories of undernutrition have an inverse relation with level of mother's education, and the lowest percentages are seen in children of mothers with educational level secondary or more. Mother's health is always positively related with their children's health. Wealth index of the family is directly related with better nutritional status of the children in all categories of nutrition. One may compare the state-wise performances with the all-India levels.

To statistically establish the relationship between different categories of undernutrition with socioeconomy over the 0–59-month children in India, categorical logistic regression has been carried out (Table 8). It is seen that urban children are more underweight, stunted, and wasted than rural children, and these results are statistically significant at 1% level of significance. By religion, Hindus are more underweight and wasted than Muslims, Christians, and other categories, and these results are statistically significant at 1% level of significance. Mother's education, mother's health status, and family wealth index are inversely related with their children's nutritional status, and all these results are statistically significant at 1% level of significance.

4 Discussion

This study reveals the growth and nutritional status among preschool 0–59-month children in India using data from the latest national level survey (NFHS-4) of 2015–16. From growth study, it is seen that there is a positive growth from all the age groups and maximum increment is seen from 12 to 35 months of age. Regarding nutritional status, among boys, maximum underweight occurs between 9 and 23 months and among girls, and it happens during 12–35 months. It is also seen that large percentages of chronic undernutrition like underweight or stunting exist from

Table 6 Zone- and state-wise percentage of undernutrition among 0–59-month children in India

Zones and states	Categories of undernutrition							
	Boys				Girls			
	N	Underweight	Stunted	Wasted	N	Underweight	Stunted	Wasted
Northeast	**15276**	**21.9**	**34.3**	**13.6**	**14510**	**20.0**	**30.9**	**12.4**
Arunachal Pradesh	1832	19.8	31.9	17.7	1734	15.9	27.4	15.1
Assam	4369	29.3	37.1	16.5	4023	26.8	33.9	15.5
Manipur	2532	14.4	30.3	8.0	2360	13.4	28.5	6.1
Meghalaya	1679	30.3	45.5	15.5	1767	28.6	41.5	15.9
Mizoram	2030	13.7	31.0	7.8	1964	13.6	28.4	8.0
Nagaland	1776	18.2	31.8	12.8	1697	16.0	25.6	10.7
Sikkim	472	16.1	32.6	16.3	404	13.1	26.0	14.9
Tripura	586	25.8	25.9	18.9	561	24.8	24.4	16.4
East	**39807**	**40.1**	**43.6**	**21.8**	**35754**	**40.5**	**43.3**	**20.2**
Bihar	10242	43.1	47.0	21.9	9314	44.6	47.6	21.0
Jharkhand	5023	48.1	45.6	30.8	4663	47.7	43.1	29.1
Orissa	4784	36.1	35.2	22.5	4461	34.8	33.8	20.8
West Bengal	2365	32.0	33.0	20.9	2208	33.9	33.9	20.7
Central	**14077**	**43.0**	**41.7**	**27.1**	**12796**	**40.4**	**39.2**	**24.4**
Madhya Pradesh	10229	43.9	42.3	27.4	9086	42.0	40.4	25.3
Chhattisgarh	3848	40.7	40.3	26.3	3710	.36.5	36.2	22.1
Uttar Pradesh	17393	38.3	44.8	19.1	15108	38.4	44.9	16.7
West	**7479**	**39.0**	**37.5**	**27.3**	**6788**	**37.2**	**35.8**	**25.5**
Goa	176	19.9	18.8	21.6	184	26.1	21.2	20.1
Gujarat	3167	42.3	40.3	29.1	2806	39.7	38.7	27.3
Maharashtra	3837	37.8	36.3	26.3	3524	35.9	34.3	24.4
Dadra and NH*	128	35.2	43.8	28.1	122	42.6	41.0	26.2
Daman and Diu	171	26.3	29.8	21.1	152	30.3	28.9	20.4
North	**20836**	**29.0**	**33.7**	**20.0**	**18201**	**26.8**	**31.2**	**18.3**
Haryana	3446	30.6	34.9	22.3	2824	28.4	31.2	21.2
Himachal Pradesh	1258	21.6	26.7	14.8	1112	18.6	22.7	13.8
Jammu and Kashmir	3366	17.1	28.3	11.9	3151	15.9	27.4	10.9
New Delhi	565	26.0	30.6	18.8	499	25.5	27.9	17.4
Chandigarh	87	24.1	37.9	9.2	77	24.7	19.5	10.4
Uttarakhand	2453	27.1	33.6	20.4	2184	26.4	32.2	20.0
Punjab	2361	21.9	25.7	16.8	2042	20.9	24.4	13.9
Rajasthan	7300	38.1	39.5	24.7	6312	35.3	36.8	22.4
South	**10460**	**28.9**	**30.6**	**21.5**	**9951**	**27.0**	**28.8**	**19.2**
Andhra Pradesh	1266	32.9	33.1	19.0	1155	29.5	29.4	15.2
Telangana	1018	30.6	29.2	19.5	871	28.2	28.6	18.6
Karnataka	2991	36.7	38.5	25.9	2871	34.8	36.2	24.0
Kerala	1024	17.2	18.7	16.6	1100	16.8	20.3	16.1
Tamil Nadu	3288	25.8	27.8	21.8	3168	23.4	25.8	18.6
Pondicherry	445	19.1	26.3	16.6	453	22.7	25.8	17.2
Lakshadweep	140	20.7	24.3	17.9	123	22.0	25.2	8.9
Andaman and NH**	288	20.1	27.4	16.7	264	22.3	23.9	15.9
India	**107935**	**34.6**	**38.4**	**34.6**	**98000**	**33.3**	**36.7**	**19.5**

*: Dadra and NH Dadra and Nagar Haveli, **: Andaman and Nicobar Island

Table 7 Percentage of undernutrition among the 0–59-month children in relation to different socioeconomic variables

Socioeconomic variables	Boys				Girls			
	N	% of Under-weight	% of Stunted	% of Wasted	N	% of Under-weight	% of Stunted	% of Wasted
Place of residence								
Rural	81806	28.7	40.5	21.9	74470	35.2	38.9	20.0
Urban	26129	36.5	32.0	19.6	23530	27.4	29.8	18.1
Religion								
Muslim	16428	33.8	39.6	19.6	15246	31.7	37.9	17.2
Hindu	78999	36.6	39.0	22.7	70688	35.7	37.5	20.9
Others	8283	21.3	34.9	13.8	8121	19.6	31.0	12.6
Christians	4225	26.9	31.4	18.7	3945	25.2	29.2	18.1
Mother's education								
Illiterate	32114	45.4	49.7	24.2	30006	44.0	47.6	21.8
Primary	15658	38.9	43.5	22.1	14118	37.4	41.6	19.8
Secondary	49727	29.5	33.0	20.0	44548	28.2	31.5	18.5
Higher	10436	19.2	22.4	17.8	9328	17.4	19.0	16.7
Mother's BMI								
Undernourished	26562	47.7	46.8	27.6	23938	47.2	45.1	25.2
Others	81373	30.3	35.7	19.3	74062	28.8	33.9	17.7
Wealth index of the family								
Poorest	27123	47.9	50.9	26.0	25364	46.6	48.9	23.6
Poorer	25023	38.1	42.8	21.8	23160	36.5	41.0	19.9
Middle	21917	31.1	36.3	20.0	19518	29.7	34.1	18.0
Richer	18361	26.3	29.5	18.6	16728	24.8	27.8	17.3
Richest	15511	20.3	23.3	17.7	13230	18.3	20.8	15.9
Total	107935	34.6	38.4	21.4	98000	33.3	36.7	19.5

6 months to second year of life, and it may happen due to stoppage of breastfeeding and weaning practices.

At present in India, 34.6% boys and 33.3% girls are underweight. Likewise, 38.4% boys and 36.7% girls are stunted. And 21.4% boys and 19.5% girls are wasted. So, boys are suffering a little bit more from undernutrition than girls. It is also seen that at present in India, among preschool children, percentage of underweight is 34.0, stunted 37.6, and wasted 20.5. There are substantive causal relations between mothers' educational and economic status (wealth index) of the household with the nutritional growth of the children.

From the above, it is seen that during ten-year gap, no significant changes have occurred in the status of health and nutrition of preschool children in India. Still now, 30–40% preschool children are undernourished regardless which of the three indices

Table 8 Categorical logistic regressions of weight-for-age, height-for-age, and weight- for-height on different socioeconomic variables among 0–59-month children in India

Independent variables	Boys			Girls		
	Underweight	Stunted	Wasted	Underweight	Stunted	Wasted
	Odd ratios	Odd ratios	Odd ratios	Odd ratios	Odd ratios	Odd ratios
Place of residence						
Rural ®	1.00	1.00	1.00	1.00	1.00	1.00
Urban	1.146***	1.085***	1.069**	1.170***	1.060**	1.090***
Religion						
Muslim®	1.00	1.00	1.00	1.00	1.00	1.00
Hindu	1.139***	0.995	1.183***	1.207***	0.995	1.245***
Others	0.539***	0.907**	0.701***	0.607***	0.824***	0.741***
Christians	0.888**	0.854***	1.024	0.903**	0.817***	1.155**
Mother's education						
Illiterate®	1.00	1.00	1.00	1.00	1.00	1.00
Primary	0.888***	0.869***	0.964	0.875***	0.873***	0.957*
Secondary	0.700***	0.661***	0.931***	0.694***	0.673***	0.961*
Higher	0.512***	0.500***	0.881***	0.486***	0.448***	0.937*
Mother's BMI						
Undernourished®	1.00	1.00	1.00	1.00	1.00	1.00
Others	0.585***	0.764***	0.691***	0.554***	0.755***	0.694***
Wealth index of the family						
Poorest®	1.00	1.00	1.00	1.00	1.00	1.00
Poorer	0.772***	0.816***	0.848***	0.764***	0.822***	0.854***
Middle	0.616***	0.679***	0.782***	0.616***	0.674***	0.772***
Richer	0.518***	0.513***	0.729***	0.513***	0.542***	0.740***
Richest	0.414***	0.434***	0.701***	0.395**	0.428***	0.673***

® Reference category; *** < 0.01: 1% level
** 0.01–0.05: 5% level
* Above 0.05–0.1: 10% level

are used for this purpose. Underweight and stunting have decreased by 10% from NFHS-3, while wasting remains almost the same.

The reason why the rate of reduction of undernutrition in India is very slow like other developing countries (Svedberg 2006; WHO 2007) is not clear. Over the last three decades, the proportion of household expenditure on food items has decreased, partly due to fall of prices of food grains and increase in the access to subsidized food grains. But the questions about adequacy still remain. It is possible that food distribution within the family has not changed much, and the unequal intra-household food distribution may have the adverse effect on preschool children (NNMB 2006).

Thus, it follows that reduction in stunting, and other forms of undernutrition can be achieved through proven interventions on mothers' literacy status, mothers' nutrition level, and economic status of the households. Improvement of mothers' nutritional

status, especially before, during, and after pregnancy, needs for urgent attention. Exclusive breastfeeding, timely safe, appropriate, and high-quality complementary food with micronutrient are essential for children.

References

Bharati, P., Pal, M., & Bharati, S. (2008a). How parent's education and working status affect the nutrition and immunization status of preschool children in India. *Asia Pacific Journal of Tropical Medicine, 1,* 49–60.

Bharati, S., Pal, M., & Bharati, P. (2008b). Determinants of nutritional status of pre-school children in India. *Journal of Biosocial Science, 40,* 801–814.

Dasgupta, M., Lokshin, M., Gragnolati, M., & Ivaschenko, O. (2005). Improving child nutrition outcome in India—Can the integrated child development services program is more effective? World Bank Policy Research Working Paper No. 3647. Washington, D.C. World Bank.

Gragnolati, M., Shekar, S., & Dasgupta, M. (2005). *India's undernourished children: A call for reform and action*. The World Bank, Washington D.C.

NNMB. (2006). National nutrition monitoring bureau. Reports Hyderabad: National Institute of Nutrition 1979–2006. http://www.nfhsindiaorg/nfhs2.html.

Sen, P., Bharati, S., Som, S., Pal, M., & Bharati, P. (2011). Growth and nutritional status of preschool children in India: A study of two recent time periods. *Food and Nutrition Bulletin, 32,* 84–93.

Som, S., Pal, M., Bhattacharya, B., Bharati, S., & Bharati, P. (2006). Socioeconomic differentials in nutritional status of children in the states of West Bengal and Assam, India. *Journal of Biosocial Science, 38,* 625–642.

Svedberg, P. (2006). Declined child malnutrition: A reassessment. *International Journal of Epidemiology, 55,* 1336–1346.

WHO Working Group. (1986). *The use and interpretation of anthropometric indicators of nutritional status*. Geneva: WHO.

World Health Organization. (2000). Obesity: preventing and managing the global epidemic. Report of a WHO Consultation. WHO Technical Report Series, No. 894. Geneva: WHO.

WHO. (2007). *Global data base on child growth and malnutrition*. Geneva: WHO.

Bootstrap of Deviation Probabilities with Applications II

Ratan Dasgupta

Abstract Bootstrap of deviation probabilities is considered for an extended zone than that obtained in Dasgupta (J Multivariate Anal 101(9):2137–2148, 2010), where it is shown that under different moment bounds on the underlying variables, bootstrap approximation for deviation probabilities of standardized sample sum, based on independent random variables, is valid for a wider zone compared to the classical normal tail probability $\Phi(-t)$, $t \to \infty$ approximation. Here, we show that the bootstrap approximation zone may further be extended compared to the earlier results of $t = o(n^{3/8})$, where n is the sample size. When skewness and kurtosis of the random variable X are zero, then by bootstrapping from a modified constructed sample where skewness and kurtosis are near to zero, the bootstrap approximation zone may be extended to $t = o(n^{2/5})$ under the assumption that $Ee^{s|X|^{8/9}} < \infty$, for some $s > 0$. Zero skewness may be achieved by bootstrapping from a symmetrized sample. Reduction in the magnitude of the fourth semi-invariant $\hat{\psi}_4$ in modified sample such that $|\hat{\psi}_4| = o(n^{-3/5})$ suffices for bootstrap approximation to hold in an enlarged zone of $t = o(n^{2/5})$. The results are extended to a triangular array of independent random variables. Construction of modified sample with nearly nil skewness and kurtosis from original sample is explained. Applications of the results include the distinction of growth processes with different levels of contamination mixed with normal process.

Keywords Bootstrap · Large deviation · Semi-invariant

MSC 2000 Classifications Primary 62F40 · Secondary 60F10 · 62E20

R. Dasgupta (✉)
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203 B T Road, Kolkata 700108, India
e-mail: ratandasgupta@gmail.com; rdgupta@isical.ac.in

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_8

1 Introduction

Performance of different statistical tests depends on respective critical regions. To compare test efficiencies, we require finding the probability of critical regions of the test statistics, which can often be approximated by sums of independent random variables. Approximation of the tail probabilities of the standardized sum of independent random variables via bootstrap is thus of interest. Below, we provide a general description of bootstrap.

Consider the iid setup for bootstrap; let $X, X_1, \ldots, X_n$ be iid random variables with distribution function F and $\theta = \theta(F)$ be a parameter of interest. Suppose $T(X_1, \ldots, X_n, F)$ be an estimator or a random variable, possibly depending on the unknown distribution F. Bootstrap procedure approximates the unknown distribution of $T(X_1, \ldots, X_n, F)$ by that of $T(Y_1, \ldots, Y_n, \hat{F})$ where $Y_1, \ldots, Y_n$ denotes a random sample of size n from $\hat{F}$, where $\hat{F} = \hat{F}_n$ is the empirical distribution function that puts equal mass $1/n$ at each of the points $X_1, \ldots, X_n$. Dasgupta (2010) considered bootstrapping deviation probabilities when all the moments of the underlying random variables are finite but the mgf of the random variables may not exist. Nonuniform rates of convergence in central limit theorem for a triangular array of random variables, where variables in each array are independent, see Dasgupta (1989), are a useful tool to obtain these results. With the help of these nonuniform rates, we obtain bootstrap approximation of tail probabilities when mgf of the random variables do not necessarily exist. The variation of normal approximation zone is considered in Dasgupta (1989); similarly, the variation of valid bootstrap approximation zone depends on the stringency of the moment assumption. One may approximate the large deviation probabilities by its bootstrap version up to the range $t = o(n^{1/3})$ when $Ee^{s|X|^{4/5}} < \infty$, for some $s > 0$; this provides a larger zone of bootstrap approximation, going well beyond the normal approximation zone; see Dasgupta (2010). While the limiting distribution in bootstrap sample is the same as that in original sample, bootstrap retains the property of the population from which the original sample is drawn, thus providing additional accuracy in approximation compared to the limiting distribution.

The results are true for a triangular array of independent random variables. The zone was extended up to $t = o(n^{3/8})$ under the assumption of vanishing third moment when $Ee^{s|X|^{6/7}} < \infty$, for some $s > 0$. The normal approximation zones are relatively smaller; it is atmost up to the range $t = o(n^{1/6})$ in general setup and up to the range $t = o(n^{1/4})$, when the third moments are zero; see, e.g., Linnik (1961), Dasgupta (1989). In general, it is not possible to obtain a larger zone than $t = o(n^{3/8})$ for valid bootstrap approximation, vide Remark 4 of Dasgupta (2010). Efficiency measures of statistical tests from a Bayesian viewpoint are explained therein.

The bootstrap approximation zone can be extended further when the kurtosis in population is zero, and when in the sample drawn kurtosis may be made near zero by appropriate modifications. In the following, we outline the procedure to achieve this.

In general, when we are sampling from a symmetric distribution with skewness and kurtosis nil, it may not be so in the sample drawn. Zero skewness may be achieved

by symmetrization of the sample cdf. However, making kurtosis zero is somewhat involved.

Kurtosis in the symmetrized sample $\{x_1, x_2, \ldots, x_n\}$ is $\hat{\mu}_4/\hat{\mu}_2^2 - 3 = (\frac{1}{n}\sum_1^n x_i^4)/(\frac{1}{n}\sum_1^n x_i^2)^2 - 3 = \Delta$ (say) and may be nonzero. Since the parent population has kurtosis zero, $\Delta = \Delta_n \to 0$ as $n \to \infty$. Rate of moment convergences is $O(n^{-1/2})$.

Since the order of an observation is $O(1)$, it appears that adding $1 \leq m < n$ pair of observations $\{\delta, 2\bar{x} - \delta\}$ with existing observations $\{x_1, x_2, \ldots, x_n\}$ will cause the value of $\hat{\mu}_4/\hat{\mu}_2^2$ in modified set of $(m + n)$ observations to increase or decrease for every fixed n depending on small or large choice of δ, where $\bar{x}$ is the sample mean.

One may then select an appropriate value of m and δ from computer simulation so as to make kurtosis nil or nearly nil in the modified sample, from which bootstrap samples may now be generated. Data near the 'middle' or 'peak' of the distribution do not contribute to the kurtosis, whereas extreme observations have an increasing effect. Thus, kurtosis is not a measure of how peak the distribution is; it may be interpreted as a measure of outliers presence. Large values do increase kurtosis, and more values near the central point reduce kurtosis. Recall that sampling from a symmetrized bootstrap will eliminate skewness. The added components like $\{\delta, 2\bar{x} - \delta\}$ maintain the symmetry around $\bar{x}$, the sample mean. Simulations with different sample sizes indicate that when δ is selected from sample observations, kurtosis of modified sample often crosses the origin for different choices of additional points incorporated from sample drawn. Thus, it is possible to modify a given sample drawn from a population with nil skewness and kurtosis, in such a manner that skewness and kurtosis may be made small in a modified sample for bootstrapping, enabling extended zone for normal approximation of the tail probability of standardized sample sum.

The present setup of the investigation is a triangular array of random variables where variables in each array are independent. The setup is fairly common for applications.

In Sect. 2, we deal with the relationship between rates in CLT and large deviation probabilities in view of stringency of moment type assumptions made. Section 3 explains the extension of bootstrap approximation zone when skewness and kurtosis in the population are nil. Appropriate modifications are made in the sample to reduce skewness and kurtosis. We explain the technique of skewness and kurtosis reduction by an example in Sect. 4. The above results may be applied to define efficiency measures of statistical tests from a Bayesian viewpoint with higher-order deviation probabilities and distinction between nearly normal growth processes with different levels of contamination.

2 Rates in CLT and Large Deviation

Consider $X, X_1, X_2, \ldots$ to be a sequence of iid random variables distributed as F with common mean μ and variance σ^2. Let $\overline{X} = \overline{X}_n = \sum_{i=1}^n X_i/n$ be the sample mean and

denote the distribution of standardized sample mean as $F_n(t) = P\{n^{1/2}(\overline{X} - \mu)/\sigma \leq t\}$. Let $\Phi(t) = P(\tau \leq t)$, where $\tau \sim N(0, 1)$, be the cdf of normal deviate. Then $F_n(t) \to \Phi(t)$, weakly. The Berry–Esseen theorem provides an uniform bound with respect to t and does not reflect the role of t, the point of convergence in the rates.

The nonuniform Berry–Esseen bounds provide more accurate convergence rate in CLT explaining the role of t. The rates are useful to investigate the behavior of tail probabilities and the deviation probabilities of standardized sample sums.

The nonuniform rates of convergence are used as basic tools to obtain bootstrap approximation results on large deviation probabilities for standardized sample mean based on a triangular array of independent random variables and to show that bootstrap approximation is more accurate than CLT approximation; see, e.g., Dasgupta (2010). In the following, we follow the same notations of the above-mentioned paper and recapitulate some of the approximation results obtained therein.

Consider $[X_{ni} : 1 \leq i \leq n, n \geq 1]$ to be a triangular array of random variables, where variables in each array are independently distributed. Assume, without loss of generality,

$$EX_{ni} = 0, \quad \forall n \geq 1, \quad 1 \leq i \leq n. \tag{2.1}$$

Denote $S_n = \sum_{i=1}^n X_{ni}$, $s_n^2 = \sum_{i=1}^n EX_{ni}^2$ and $F_n(t) = P(s_n^{-1} S_n \leq t)$. Let

$$\inf_{n\geq 1} n^{-1/2} s_n = c(> 0). \tag{2.2}$$

When the random variables are iid, c of (2.2) equals to the common variance σ^2. The same notation $F_n(t)$ is used for the distribution of standardized sample sum in iid/independent setup with the specification of the context.

When Lindeberg condition is satisfied, one has $F_n \to \Phi$, weakly. To study the speed of convergence, we need to assume the existence of moments slightly higher than two for the random variables X_{ni}. Assume that

$$\sup_{n\geq 1} n^{-1} \sum_{i=1}^{n} EX_{ni}^2 g(X_{ni}) < \infty, \tag{2.3}$$

where $g(x)$ is a nonnegative, continuous, even, nondecreasing function on $[0, \infty)$.

Let $g(x)$ be such that,

$$|x|^k << g(x) << \exp(s|x|), \ \forall k > 0 \tag{2.4}$$

and some $s > 0$, along with $x^{-1} \log g(x)$ is nonincreasing for $x > x_o(\geq 0)$.

Semi-invariants or cumulants of a random variable X are common in the expansion of logarithm of the characteristic function. One may express $\phi(t) = \log E(e^{itX}) = \sum_0^r \kappa_j \frac{(it)^j}{j!} + o(t^r)$, as $t \to 0$, if $E(X^r)$ exists. The coefficients κ_j are named as jth semi-invariant of X. These quantities may be expressed in terms of the moments of the random variable X.

We state below two relevant theorems on nonuniform rates in CLT and on normal approximation zone, viz. Theorems 2.13 and 2.15 of Dasgupta (1989), pp. 158–159. Due to the importance of the theorems in obtaining the present approximation results, we mention these below.

Theorem A *Let* $[X_{ni} : 1 \le i \le n, n \ge 1]$ *be a sequence of independent random variables in a triangular array. Let (2.1)–(2.4) hold. Let* t^* *be the largest value of* $|t|$ *satisfying*

$$1 \le t^2 \le 2(\log|t| + \log g(rs_n t)) \tag{2.5}$$

with $|t| \le \epsilon n^{1/2}$, *where* $\epsilon(>0)$ *is small. Define* $\psi_k = \psi_k(ni)$ *the kth semi-invariant of* $Y_i = Y_{ni} = X_{ni} I(|X_{ni}| \le rs_n t^*)$, $0 < r < 1/2$ *and* $\psi_k^* = \psi_k^*(n) = s_n^{-2} \sum_{i=1}^n \psi_k(ni)$.

Then for the zone (2.5), the following holds.

$$\begin{aligned} &|F_n(t) - \Phi(t)| \\ &\le I_1 + I_2 + I_3 + \sum_{i=1}^n P(|X_{ni}| > rs_n t^*), \end{aligned} \tag{2.6}$$

where

$$I_1 \le bn^{-1/2} \exp\left\{ -\frac{t^2}{2} + \sum_{k=3}^{\infty} \frac{1}{k!} \frac{t^k}{s_n^{k-2}} \psi_k^* + O(t^{*2} g^{-1}(rs_n t^*)) \right\},$$

$$I_2 \le bt^{-1} e^{-t^2/2} \left| \exp\left\{ \sum_{k=3}^{\infty} \frac{(k+1)}{k!} \frac{t^k}{s_n^{k-2}} \psi_k^* + O(t^{*2} g^{-1}(rs_n t^*)) \right\} - 1 \right|$$

$$\times \exp\left[\frac{n^{-1}}{2c} \left\{ \sum_{k=3}^{\infty} \frac{1}{(k-1)!} \frac{t^{k-1}}{s_n^{k-3}} \psi_k^* \right\}^2 + \sum_{k=3}^{\infty} \frac{1}{(k-1)!} \frac{t^k}{s_n^{k-2}} \psi_k^* + O(t^{*2} g^{-1}(rs_n t^*)) \right]$$

and

$$I_3 \le bt^{-1} e^{-t^2/2} \exp\left\{ \frac{t^2}{2} \left(\sum_{k=3}^{\infty} \frac{1}{(k-2)!} \left(\frac{t}{s_n}\right)^{k-3} \psi_k^* \right) + O(t^{*2} g^{-1}(rs_n t^*)) \right\} +$$

$$bt^{-1} e^{-t^2/2} \left| \exp\left\{ \frac{t^2}{2} \left(\sum_{k=3}^{\infty} \frac{1}{(k-2)!} \left(\frac{t}{s_n}\right)^{k-3} \psi_k^* \right) + O(t^{*2} g^{-1}(rs_n t^*)) \right\} - 1 \right|.$$

The tail probabilities $1 - F_n(t)$ and $1 - \Phi(t) = \Phi(-t)$ as $t \to \infty$, are small. It is thus of interest to know when these are of equal order in magnitude, i.e., $\frac{1-F_n(t)}{1-\Phi(t)} \to 1$, as $t \to \infty$. Cramér (1938) proved that if $Ee^{s|X|} < \infty$, for some $s > 0$, then the above holds for $t = o(n^{1/6})$. Extensions made by Linnik (1961) showed that the condition on existence of mgf may be relaxed to $Ee^{s|X|^{1/2}} < \infty$, for some $s > 0$. Further, to

obtain $(1 - F_n(t))/(1 - \Phi(t)) \to 1$ for $t = o(n^{\alpha/(2(2-\alpha))})$, it is enough to assume $Ee^{s|X|^\alpha} < \infty$ for some $s > 0$; see, e.g., Dasgupta (1989). Necessary part of the assumption is also proved therein.

If $|F_n(t) - \Phi(t)| = o(t^{-1}e^{-t^2/2}) = o(1 - \Phi(t))$ holds, we may conclude that $|\frac{1-F_n(t)}{1-\Phi(t)} - 1| = o(1), t \to \infty$.

The following theorem on normal approximation of tail probability is also proved in Dasgupta (1989), as a corollary of Theorem A.

Theorem B *Under the assumptions of Theorem A along with*

$$'\psi_l^*(n) = s_n^{-2} \sum_{i=1}^{n} {}'\psi_l(ni) = o(s_n^{l-2}|t_n|^{-l}) \tag{2.7}$$

for $l = 3, 4, \ldots, k-1$ *where* $'\psi_l(ni)$ *is the* l*th semi-invariant of* X_{ni}*, and for a sequence* $\{t_n\}$ *satisfying*

(i) $t_n = o(n^{\frac{k-2}{2k}})$

(ii) $t_n^2 - (\log|t_n| + \log g(rs_nt_n)) \to -\infty;\ 0 < r < 1/2$ *the following holds.*

$1 - F_n(t_n) \sim \Phi(-t_n)$, $F_n(t_n) \sim \Phi(-t_n)$, *as* $t_n \to \infty$.

If (i) *is more stringent than* (ii)*, e.g., for* $g(x) = \exp(s|X_{ni}|^{(k-2)/(k-1)})$ *for some* $s > 0$*, then (2.7) is equivalent to* $'\psi_l^*(n) = o(n^{-1-l/k}),\ l = 3, 4, \ldots, k-1$.

To prove the theorem, one has to truncate the random variables in such a manner that the mgf of the truncated variables exists and the values of the mgf depend on the value given in (2.3). One may then use Cramér's auxiliary distribution function to prove Theorem A. The absolute value of the random variables is truncated at $rs_nt^* = O_e(n^{1/2}t^*)$, where O_e denotes the exact order. The truncation has negligible effect on the semi-invariants as

$$|'\psi_l(ni) - \psi_l(ni)| = O(s_nt^*)^{-3} = O(n^{-3/2}/t^{*3}) = O(n^{-3/2}),\ l \geq 3, t^* > 1 \tag{2.8}$$

Therefore, we shall use the same notation for the semi-invariants of the truncated and non-truncated random variables, up to the approximation order $o(1)$. For the iid random variables, one has $s_n^2 = n\sigma^2$ and $'\psi_l^*(n) = {}'\psi_l(ni) = {}'\psi_l$.

Below, we prove results on bootstrapping large deviation probabilities on an extended zone $t = o(n^{2/5})$, when the population is symmetric with fourth semi-invariant zero.

3 Bootstrapping Large Deviation

Let us consider a sample of size n from distribution F and denote $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ to be the observed ordered statistics. Notations F_n^*, P^* represent the bootstrap characteristics fixing the given sample; see also Efron (1979).

We now state the first theorem of this section. This modifies Theorem 1 of Dasgupta (2010), toward a larger bootstrap zone under certain assumptions.

Theorem 1 *Let the assumptions (2.1)–(2.4) be satisfied for a sequence of iid random variables* $X, X_1, X_2, X_3, \ldots$ *distributed as* F *with variance* σ^2. *Also, let the function* $\log(x^2 g(x))$ *be subadditive on* $[x_o, \infty)$, *for some* $x_o > 0$, *i.e.,* $\log[(x+y)^2 g(x+y)] \leq \log(x^2 g(x)) + \log(y^2 g(y))$, $x, y \in [x_o, \infty)$.

Then, there exists a constant $r(> 0)$, *such that for a sequence* $\{t_n\}$ *satisfying simultaneously (1) and (2), where*

(1) $t_n = o(n^{1/3})$; *and* $t_n = o(n^{3/8})$ *if* $E_F X^3 = 0$, *and the bootstrap sample is drawn from symmetrized version* $\hat{F}^s$ *of* $\hat{F}$, *the empirical distribution function; and* $t_n = o(n^{2/5})$ *if* $E_F X^3 = 0$, *kurtosis* $\mu_4/\mu_2^2 - 3 = 0$ *and the bootstrap sample is drawn from symmetrized version* $\hat{F}^s$ *of* $\hat{F}$; *the empirical distribution function and additional sample points are incorporated to make the sample kurtosis nearly zero, i.e.,* $|\hat{\psi}_4| = o(n^{-3/5})$; *and*

(2) $t_n^2 - 2(\log t_n + \log g(r\sigma n^{1/2} t_n)) \leq M,\ M > 0,\ 0 < r < 1/2,$

the following holds for the sequence t_n. $1 - F_n^*(t_n) \sim 1 - F_n(t_n)$, $F_n^*(-t_n) \sim F_n(-t_n)$ *in probability as* $t_n \to \infty$.

Proof of the theorem. We explain the essential modifications required in the similar steps in proof followed in Dasgupta (2010). When simulation provides values of m and δ such that skewness and kurtosis are nearly zero in a modified sample from which the bootstrap sample has to be drawn, then the approximation zone may be extended further. As mentioned earlier with appropriate modification of the sample distribution function via symmetrization and adding some more elements in sample, one may obtain sample skewness and kurtosis to be zero, when these are zero in the population. Thus, we may consider a third term in the expansion (3.1) of Dasgupta (2010) as $|\psi_4 - \hat{\psi}_4| = 0$ and concentrate on the difference $|\psi_5 - \hat{\psi}_5| = O(n^{-1/2})$ a.s. The expression then can be written as

$$\frac{1 - F_n^*(t)}{1 - F_n(t)} = O_p(e^{(\frac{t^3}{n^{1/2}}(\psi_3 - \hat{\psi}_3) + \frac{t^4}{n}(\psi_4 - \hat{\psi}_4) + \frac{t^5}{n^{3/2}}(\psi_5 - \hat{\psi}_5))(1 + o(1))}) \tag{3.1}$$

provided $nP(|X| > r\sigma n^{1/2} t^*)/\Phi(-t) = o(1)$ and $nP^*(|Z - \overline{x}| > r\hat{\sigma} n^{1/2} t^*)/\Phi(-t) = o_p(1)$. The third term in the exponent can be written as $\frac{t^5}{n^{3/2}}(\psi_5 - \hat{\psi}_5) = O(\frac{t^5}{n^2})$. Thus, the second term in the exponent is of same order as the third, if $\frac{n}{t}(\psi_4 - \hat{\psi}_4) = O(1)$, i.e., $\hat{\psi}_4 = O(t/n)$, if $\psi_4 = 0$. Letting $t = o(n^{2/5})$ this requires $\hat{\psi}_4 = O(n^{-3/5})$, if $\psi_4 = 0$. The requirement is not very stringent in view of the fact that from sample moment convergence results already $\hat{\psi}_4 = O(n^{-1/2}) = O(n^{-3/6})$, as $\psi_4 = 0$. An

additional reduction of the order like $O(n^{-1/10})$ is required by a little bit of sample modification toward extending the bootstrap zone.

The bootstrap approximation zone may then be extended to $t = o(n^{2/5})$ under the assumption that $Ee^{s|X|^{8/9}} < \infty$.

The results may be extended from iid setup to a triangular array of random variables, where variables in each array are independently distributed. The proof follows the similar steps of Dasgupta (2010) Theorem 2, with similar notations.

Theorem 2 *Let the assumptions (2.1)–(2.4) hold and* $\log(x^2 g(x))$ *is subadditive on* $[x_o, \infty)$, *for some* $x_o > 0$. *Further let* $|n^{-1}\sum_{i=1}^{n} x_{ni}^2 g_1(x_{ni}) - n^{-1}\sum_{i=1}^{n} EX_{ni}^2 g_1(X_{ni})| = o(1)$, *a.s., as* $n \to \infty$ *be satisfied for a triangular array of random variables where variables in each array are independent. Then, there exists a constant* $r(> 0)$, *such that for a sequence* $\{t_n\}$ *satisfying simultaneously (1) and (2), where*

(1) $t_n = o(n^{1/3})$, *and* $t_n = o(n^{3/8})$ *if* $EX_{ni}^3 = 0, \psi_4 = \psi_{4,ni} = 0, \forall n, i$; *i.e., populations are symmetric with kurtosis 0, and the bootstrap sample is drawn from symmetrized version* $\hat{\overline{F}}^s$ *of* $\hat{\overline{F}}$, *and* $t_n = o(n^{2/5})$ *if* $E_F X_{ni}^3 = 0, \forall n, i$ *and the bootstrap sample is drawn from symmetrized version* $\hat{\overline{F}}^s$ *of* $\hat{\overline{F}}$; *the empirical distribution function and additional sample points are incorporated to make the sample kurtosis nearly zero, i.e.,* $|\hat{\psi}_4| = o(n^{-3/5})$; *and*

(2) $t_n^2 - 2(\log t_n + \log g(r s_n t_n)) \to -\infty,\ 0 < r < 1/2;$

the following holds for the sequence t_n.
$1 - F_n^*(t_n) \sim 1 - F_n(t_n),\ F_n^*(-t_n) \sim F_n(-t_n)$ *in probability, as* $t_n \to \infty$, *where* $F_n(t) = P(s_n^{-1} S_n \le t),\ F_n^*(t) = P_{\hat{\overline{F}}}[n^{-1/2}\sum_{i=1}^{n}(z_i - \overline{x}) \le t\hat{\sigma}]$, *and* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Reduction of kurtosis to a further extent of $O(n^{-1/10})$ in sample ensures the validity of bootstrap approximation in an extended zone. Below, we explain by an example the technique of kurtosis reduction.

4 Modification in Sample to Reduce Kurtosis: An Example

The seed value taken for generating random numbers is 1. Total number of observations considered is 5.

First, we have taken 5 iid random observations from N (0, 1) in R. The generated numbers are $-0.8356286, -0.6264538, 0.1836433, 0.3295078, 1.5952806$. We have taken the mean of the five observations and then have ordered the observations. Now, we have ten observations, i.e., 5 for ordered observations and 5 observations from (2 × mean—each observations), so as to symmetrize the sample.

The calculated kurtosis from symmetrized sample is -0.8702354. We recalculated sample kurtosis by taking the fifth observation from the whole set of ten observations along with (2× mean—fifth observation), thus giving two extra observations.

With 12 observations in total, the modified value of kurtosis is −0.4483642, this is lower in magnitude than the previous one.

Next, we recalculated sample kurtosis by taking the sixth observation from the whole set of ten observations along with (2 × mean—sixth observation), thus giving 2 extra observations; with 12 observations in total, the modified value of kurtosis is −0.4468868, and this is lower in magnitude than the previous value.

Then, again we recalculated the sample kurtosis by taking the fifth and sixth observations along with (2 × mean—fifth/sixth observations), which gives four extra observations and thus now we have 14 observations in total.

The calculated sample kurtosis is −0.02784223, which is lower than the previous value in magnitude.

Next, we recalculated the sample kurtosis by taking the fifth observation twice and sixth observation once, along with their symmetrized part, thus having 16 observations in total. The calculated kurtosis is 0.3913355. This crosses from negative value kurtosis to a positive value, indicating the possibility of attaining zero at some intermediate stage. Kurtosis calculated at a previous stage was −0.02784223; this is lowest in magnitude among the values calculated.

We accept the value of kurtosis −0.02784223 in simulation, resulting from 14 observations. These 14 observations constitute the modified sample for bootstrapping. The number of extra observations added after initial symmetrization is 14 − 10 = 4. The suggested procedure may be automated by an appropriate computer program with minimal steps.

5 Applications: Growth Processes with Small and Different Levels of Noise

The above results may be applied to define efficiency measures of statistical tests from a Bayesian viewpoint with probabilities of higher-order deviation $o(n^{2/5})$ under the assumption of nearly zero skewness and kurtosis, so as to validate bootstrap approximation of tail probabilities in an extended zone. Identification of normal processes with *small and different levels of contamination* is possible. In electronic recordings like EEG/ECG, noise is usually associated with signal to contaminate the process. Two such models on stochastic processes are described in Dasgupta (2010). One may proceed in a similar manner to test for equality of population means based on sample mean under normal process with small and different levels of noise. Note that from (3.1) of the present paper, to validate bootstrap in extended zone we need $|\frac{n}{t}(\psi_4 - \hat{\psi}_4)| = O(1)$, $|t| = o(n^{2/5})$, considering both left and right tail probabilities. The above requirement $|\psi_4 - \hat{\psi}_4| = O(n^{-3/5})$ can be made possible for nearly normal processes with different levels of contamination. In Tables 1–3 of Dasgupta (2010), a possible value of $\alpha/2$ may then be considered so as to arrive at the truncation point $n^{2/5.01}$, and to simulate the bootstrap probabilities for different sample sizes under two different models. The bootstrap estimate of the probabilities

of critical regions may be used to quantify the efficiency of using tail probabilities as tests for population means, i.e., critical region probabilities for Bayes tests which use rejection regions based on shrinking tail areas of the sample mean's distribution/large deviation zones.

References

Cramér, H. (1938). Sur un nouveau theoreme limite de la probabilities. Actualites Sci. Indust. No. 736.

Dasgupta, R. (1989). Some further results on nonuniform rates of convergence to normality. *Sankhyā A*, *51*, 144–167.

Dasgupta, R. (2010). Bootstrap of deviation probabilities with applications. *Journal of Multivariate Analysis*, *101*(9), 2137–2148.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Linnik, Y. V. (1961). Limit theorems for sums of independent variables taking into account large deviation, I, Theory of probability and its application (Vol. 6, pp. 3, 131–147).

Recent Advances in the Statistical Analysis of Retrospective Time-to-Event Data

Sedigheh Mirzaei Salehabadi and Debasis Sengupta

Abstract In a cross-sectional observational study on time-to-event, the probability distribution of that time is often estimated from data on current status. Recall data on the time of occurrence of the landmark event can provide more information in this regard. Even so, the subjects may not be able to recall the time precisely. This type of incompleteness is a peculiarity of recall data, which poses a challenge to analysis. Valid likelihood-based procedures for inference have emerged in a number of papers published only recently. In this article, we review these papers and show how one can estimate the time-to-event distribution parametrically or nonparametrically, and also assess the effect of covariates, by using current status data or incompletely recalled data. The methods are illustrated through the analysis of menarcheal data from a recent anthropometric study of adolescent and young adult females in Kolkata, India.

Keywords Current status data · Informative censoring · Interval censoring
Relative risk regression model · Retrospective study · Turnbull estimator

1 Introduction

Time to occurrence of an event is an object of interest in various fields. Observational studies have been carried out to study the time until onset of menarche of females (Bergsten-Brucefors 1976; Chumlea et al. 2003; Mirzaei and Sengupta 2015), breast development of females (Cameron 2002; Aksglaede et al. 2009), dental development of infants (Demirjian et al. 1973; Eveleth and Tanner 1990), birth of the first child of a woman (Allison 1982), beginning of a criminal career (Hosmer and Lemeshow 1999), end of a work career (LeClere 2005), end of a strike (Hosmer and Lemeshow 1999), and so on. In anthropometric studies, the age of passing various

S. M. Salehabadi
E. K. Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA

D. Sengupta (✉)
Applied Statistical Unit, Indian Statistical Institute, Kolkata, India
e-mail: sdebasis@isical.ac.in

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_9

developmental landmarks is examined by their own right, and also as useful covariates for body dimensions used for obtaining growth curves (Salsberry et al. 2009; Vizmanos et al. 2001). One may wish to estimate the probability distribution of the time to occurrence of a particular event in order to compare two populations. Such estimates may also be useful in setting benchmarks for individuals or setting policy objectives. Most of the observational studies on time-to-event are cross-sectional in nature, though there are some instances of study designs for observing a number of individuals continuously or periodically until the occurrence of the landmark event (Korn et al. 1997; McKay et al. 1998).

There are many parametric models for the probability distribution of the time-to-event, viz. exponential, Weibull, lognormal, gamma, Gompertz, log-logistic, Pareto, generalized gamma. Once a parametric model for the time-to-event has been chosen, standard techniques for parametric inference become applicable. However, these techniques are meant for complete data. Cross-sectional time-to-event data may be incomplete in many ways. For example, the time would not be known in the case of individuals who did not experience the event. If one records only the current status of the individual in terms of the happening of the event, the time-to-event is not recorded even for those who have experienced the event. If the interviewed individual is asked to recall the time of occurrence of the event, there may be occasional cases of complete failure to remember. This would result in another form of incompleteness. Yet more complex forms of incompleteness would arise if some individuals are only able to recall a range of time when the event had occurred.

Most of the data arising from these situations can be broadly referred to as censored data. There are modified versions of likelihood-based techniques, which work for censored data. However, the nature of modification depends on the nature of censoring. One has to make certain assumptions about the censoring mechanism in order to be able to specify an appropriate likelihood. A key assumption which is often made is that the mechanism of censoring is independent of the time-to-event. This assumption essentially means that a particularly long or particularly short time-to-event does not have any more or any less chance of being censored, compared to other cases. It can be shown that this assumption can be particularly problematic for data obtained through recall. The event of recall induces a special type of dependent censoring that has been specifically modeled in recently published literature.

This article is intended to provide an up-to-date overview of the methods of inference available to those who aspire to analyze time-to-event data collected from a cross-sectional study, without going deeply into the technical details, which can always be obtained from the original sources cited here. We focus on methods that are based on likelihood. Consequently, many popular methods, such as those based on probit model for the event of menarche before a specific age (Hediger and Stine 1987), are excluded from the purview of our discussion.

The remainder of this article is organized as follows. Section 2 reviews the current status data on time-to-event and likelihood-based inference procedures available for it. Section 3 deals with perfectly recalled time-to-event data and the relevant procedures. Section 4 dwells on parametric and nonparametric inference in the case where some of the time-to-event are not recalled at all. Section 5 shows how one can

incorporate the effect of covariates on time-to-event distribution through regression models for various types of cross-sectional data. In Sect. 6, there is a brief discussion on partial recall and recall error. An illustrative data analysis is reported in Sect. 7. The data analysis is based on a study of menarcheal age of adolescent and young adult females, undertaken by the Indian Statistical Institute, Kolkata. Some concluding remarks are given and areas for future work are identified in Sect. 8.

2 Current Status Data

Current status data, also known as status quo data (Teilmann et al. 2009), consist of the value of a binary status variable that indicates whether or not the landmark event has occurred till the day of observation.

Consider a set of n subjects with the landmark event occurring at times $T_1, \ldots, T_n$, which are samples from a common distribution F with density f and support $[t_{min}, t_{max}]$. Let these subjects be observed at times $S_1, \ldots, S_n$, respectively, chosen from a finite set $\mathcal{S}$. Let, for $i = 1, \ldots, n$, δ_i be the indicator of $T_i \leq S_i$, i.e., the event having had occurred on or before the time of interview.

Current status data arise from the observation consisting only of (S_i, δ_i), $(i = 1, 2, \ldots, n)$. The corresponding likelihood, conditional on the time of interview, is

$$\prod_{i=1}^{n} [F(S_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{1}$$

where $\bar{F} = 1 - F$. If the distribution F is assumed to be a member of a parametric family characterized by the parameter θ (which may be a vector of parameters), then the parametric MLE is obtained by maximizing the above likelihood with respect to θ. There has been considerable interest in the parametric analysis of current status data (Shiboski and Jewell 1992; Sun and Kalbfleisch 1993). For properties of the MLE based on the above likelihood, see Lee and Wang (2003).

It is also possible to estimate the distribution nonparametrically, that is, without assuming any particular functional form of the distribution. Note that if the ith respondent is observed to have experienced the event of interest, then it is known that the time-to-event T_i belongs to the interval $[t_{min}, S_i]$. If the event has not been experienced, then T_i belongs to the interval $[S_i, t_{max}]$. In either case, T_i is known to belong to an interval. This is a special case of interval censoring, sometimes referred to as Case I interval censoring (Sun 2006).

In general, interval censoring refers to the situation where one only knows that the time-to-event lies in a certain window of time; i.e., T_i belongs to an observed interval $[L_i, R_i]$. The case of no censoring $(L_i = R_i)$ can be indicated by the binary variable η_i. When the data contain instances of no censoring $(L_i = R_i)$, censoring from the right $(R_i = t_{max})$, censoring from the left $(L_i = t_{min})$, and censoring from

both sides ($t_{min} < L_i < R_i < t_{max}$), the censoring is called mixed interval censoring (Sun 2006, Chap. 2).

If the censoring mechanism is independent of the time T_i (an assumption that usually holds for current status data), the general likelihood for interval-censored data is

$$\prod_{i=1}^{n} [f(T_i)]^{\eta_i} [\bar{F}(L_i) - \bar{F}(R_i)]^{1-\eta_i}, \tag{2}$$

where f is the probability density function corresponding to the distribution F. Note that in the case of current status data, $\eta_i = 0$ for every i, and $[L_i, R_i]$ is constrained to be either $[t_{min}, S_i]$ or $[S_i, t_{max}]$, so that the likelihood (2) reduces to (1). A nonparametric maximum likelihood estimator (NPMLE) of F for general interval-censored data would be the distribution function that maximizes the above likelihood. This NPMLE was derived by Ayer et al. (1955). Turnbull (1976) worked on it further and gave a computational algorithm. This algorithm consists of partitioning the range $[t_{min}, t_{max}]$ into disjoint subintervals, such that every observed interval $[L_i, R_i]$ can be expressed as a union of these subintervals. There can only be a finite number of such subintervals. Once this partitioning is done, the task of identifying the NPMLE reduces to allocating optimum probabilities to these subintervals so that the total probability is 1 and the above likelihood is maximized. See Keiding et al. (1996) for details of this estimator, generally known as the Turnbull estimator.

The Turnbull estimator has an undesirable characteristic. When a subinterval is of positive length (i.e., left and right end-points do not coincide), the probability allocated to that interval can be distributed in any manner within the interval, without affecting the value of the likelihood. In other words, the NPMLE is not unique. Two different distribution functions that allocate identical probabilities to each subinterval (while distributing the probability within the intervals in different ways) can happen to be NPMLEs. In the case of current status data, every single subinterval is likely to be of positive length. Therefore, the ambiguity about the NPMLE prevails everywhere, except at the boundaries of the subintervals! Practically speaking, the NPMLE specifies a distribution only at a finite number of points and is silent about how they should be interpolated to obtain the full description of a distribution function.

A desirable property of an estimator is that when the sample size is increased, it should be probabilistically very close to the quantity being estimated. This property is called consistency. Consistency of an estimator, under appropriate conditions, needs to be established for it to be credible. This holds for estimators of single parameters, vector parameters, and even functions. In particular, when a distribution function is estimated by a function computed from the data, it should converge to the true distribution function, under appropriate conditions, as the sample size goes to infinity. In the case of an NPMLE of a distribution function obtained from interval-censored data, this requirement poses a conceptual problem, since the NPMLE is only the set of values of a function at a few points and not a fully specified function. Gentleman

and Geyer (1994) brought in the requisite formalism to establish the consistency of the NPMLE of F from independently interval-censored data.

Various methods of inference for interval-censored data have been explained in books such as Sun (2006), Kalbfleisch and Prentice (2002), and Lee and Wang (2003).

3 Perfectly Recalled Time Data

In some cross-sectional studies, a subject is asked to recall the time of occurrence of the landmark event, in case it has already taken place. Such retrospective data are usually incomplete (Roberts 1994; Padez 2003). The subject may not be able to recall the time at all or may be able to specify only a range for the requisite time. Even if there is no difficulty of recall (which may happen, for instance, if there is a formal record of the time of occurrence), there would be incompleteness in the data in respect of those individuals who did not experience the event yet. In this section, we only consider the latter situation, where there is no problem of recall and the only incompleteness arises from the possible nonoccurrence of the event at the time of data collection.

Going by the notations used in the previous section, the observable quantities in this situation are T_i when $\delta_i = 1$ and S_i when $\delta_i = 0$. The censoring involved here is from the right, in the sense that the time-to-event is longer than the time of observation (censoring time). This is a special case of interval censoring, with $R_i = L_i = T_i$ when $\delta_i = 1$ and $L_i = S_i$, $R_i = t_{max}$ when $\delta_i = 0$. The simplified form of the likelihood (2) is

$$\prod_{i=1}^{n} [f(T_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{3}$$

Assuming S_i is random and independent of T_i, we essentially have randomly right-censored data, which has been dealt with extensively in the literature. Usual large sample properties of many parametric likelihood-based techniques have been shown to hold for randomly right-censored data, under appropriate conditions (Lawless 2003). Modifications of goodness-of-fit tests for randomly right-censored data have also been proposed (Lawless 2003, Chap. 10). If one does not assume the functional form of the distribution, the above likelihood can be maximized with respect to the function $\bar{F}$ to obtain the nonparametric MLE. This NPMLE happens to be the well-known Kaplan–Meier estimator. For properties of this estimator, two-sample tests, and other related procedures, see Kalbfleisch and Prentice (2002), Hosmer et al. (2008), Lawless (2003), and Klein and Moeschberger (2003).

4 Recalled Time Data with Occasional Failure to Recall

Let us now consider the situation where a subject may not be able to recall the time of the event of interest. Non-recall necessarily means that the time-to-event T_i can

have any value smaller than the time till observation (S_i), which corresponds to left censoring. Here, we ignore the possibility of the subject recalling an approximate date and regard such occurrence as a non-recall event. Thus, the entire data set would consist of only three types of cases: complete data arising from the cases of perfect recall, left-censored data arising from the case of non-recall, and right-censored data arising from the cases where the event did not take place yet. These cases can be described by the binary variable δ_i, which indicates whether the event happened till the time of observation ($T_i \leq S_i$), and another binary variable ε_i, which indicates whether the time of the event is recalled at all (assuming that it has happened). Specifically, the three cases correspond to $\delta_i \epsilon_i = 1$, $\delta_i(1-\epsilon_i) = 1$, and $\delta_i = 0$.

Such a data set can be readily seen to be a special case of interval-censored data, discussed in Sect. 2, where the likelihood (2) reduces to

$$\prod_{i=1}^{n} \left[(f(T_i))^{\varepsilon_i} (F(S_i))^{1-\varepsilon_i} \right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{4}$$

However, this likelihood and the related procedures are applicable only when the censoring mechanism is independent of the time-to-event. Incompleteness in recall data in a cross-sectional study occurs in such a way that this assumption is violated. This is because of the fact that memory often fades with time. Between two persons interviewed at the same age, the one with earlier occurrence of the event of interest has less chance of recalling the time. Mirzaei et al. (2014) and Mirzaei and Sengupta (2016) have shown that the use of the likelihood (2) can lead to biased estimation, both in the parametric and the nonparametric cases, though there are instances when the NPMLE (Turnbull estimator) has been used for studying the distribution of age at reaching a developmental landmark by using recall data (see, e.g., Aksglaede et al. 2009).

In some existing models and methods for dependent censoring (see, e.g., Finkelstein et al. 2002; Scharfstein and Robins 2002), censoring is assumed to occur through duration variables that have the same origin of measurements as that of the duration of interest. Since this assumption does not hold here, these methods are not applicable. Mirzaei et al. (2014) took into account the special type of incompleteness arising from recall data by new modeling. They recognized that the non-recall probability may depend on the observation time and the time-to-event, and modeled it as a function π of the time elapsed since the occurrence of the event till the time of observation,

$$P(\varepsilon_i = 0 | S_i = s, T_i = t) = \pi(s - t),$$

where $s > t > 0$. The three types of data mentioned above would lead to different contributions to the likelihood. By putting these cases together, the likelihood according to this model can be shown to be

$$\prod_{i=1}^{n} \left[\left(\int_0^{S_i} f(u)\pi(S_i - u)du \right)^{1-\varepsilon_i} [f(T_i)(1 - \pi(S_i - T_i))]^{\varepsilon_i} \right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{5}$$

When π is a constant, the likelihood (5) becomes a constant multiple of the independent interval censoring likelihood (4). As a further special case, if $\pi = 1$, it reduces to the current status likelihood (1). On the other hand, when $\pi = 0$, the likelihood reduces to the perfect recall likelihood (3). Thus, the model that leads to the likelihood (5) is more general than the models for independent censoring.

Mirzaei et al. (2014) assumed parametric forms of the functions π and F, established consistency and asymptotic normality of the MLE under the above model, subject to suitable regularity conditions. They also suggested a graphical method of guessing the functional form of the non-recall probability π. Mirzaei and Sengupta (2016) allowed the distribution function to be arbitrary and eliminated the integral in the likelihood (5) by assuming a piecewise constant form of π:

$$\pi(x) = \begin{cases} b_1 & \text{if } x_1 < x \le x_2, \\ b_2 & \text{if } x_2 < x \le x_3, \\ \vdots & \\ b_k & \text{if } x_k < x < \infty, \end{cases} \tag{6}$$

where $0 = x_1 < x_2 < \cdots < x_k$; $0 < b_1, b_2, \ldots, b_k \le 1$. They derived the NPMLE of F obtained by maximizing the resulting likelihood, which can be obtained through a self-consistency algorithm. Significantly, they showed that when the sample size is large, the NPMLE tends to have probability concentrated only on the distinct times of exactly recalled events. Accordingly, they proposed an approximate NPMLE (AMLE), which is computationally much simpler and is asymptotically equivalent to the NPMLE. The AMLE is obtained by maximizing the approximate likelihood, written in terms of the probabilities $q_1, \ldots, q_m$ attached to the exactly recalled event times $t_1, \ldots, t_m$, as the product of weighted sums

$$\prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} q_j \right). \tag{7}$$

The weights α_{ij} are computable from the data as linear functions of $b_1, \ldots, b_k$, which may be regarded as nuisance parameters while maximizing (7) with respect to $q_1, \ldots, q_m$. Mirzaei and Sengupta (2016) discussed how the variance of the AMLE can be estimated. They showed that both the NPMLE and the AMLE are consistent estimators of the underlying distribution under general conditions.

The two-sample problem for data of this type has not been addressed yet. A solution under the restriction of proportional hazards may be obtained by considering the Cox regression model with a single binary covariate, discussed in the next section.

5 Regression

All the likelihoods presented in the three foregoing sections are based on the assumption that the underlying time-to-event for all the individuals are independent and have a common distribution F with density f. If each individual has a different distribution, the same likelihoods continue to hold after F and f in the factors are replaced by their individual-specific versions: F_i and f_i, respectively.

A parametric regression model provides a functional description of the distribution of T_i given the covariate vector Z_i in terms of the distribution parameters θ and the regression parameters β. Specifically, $F_i(t|Z_i)$ can be written as $F_0(t|Z_i, \theta, \beta)$, where F_0 is a known 'baseline distribution.' This substitution reduces the problem of obtaining the MLEs of the regression parameters as another optimization problem with β and θ (and possibly the parameters of the function π) as optimizing variables. This problem is conceptually similar to parametric estimation. Standard procedures (see, e.g., Lee andWang 2003) with appropriate modification of asymptotic results are applicable.

In recent years, semiparametric regression models have gained popularity. These models deal with covariates parametrically, while keeping a nonparametric flavor as far as the baseline distribution is concerned. They make fewer assumptions than a completely parametric model, but more assumptions than a model that would assign a different time-to-event distribution to every case. This amounts to expressing $F_i(t|Z_i)$ as $F_0(t|Z_i, \beta)$, where F_0 is a completely unspecified distribution function. Examples of semiparametric regression models are Cox's relative risk model (Cox 1972), the accelerated failure time (AFT) model (Wei 1992), the additive hazard regression model (Klein and Moeschberger 2003), the proportional odds model (Dabrowska 1988), and so on. A summary of the methods available for randomly right-censored data may be found in Hosmer et al. (2008). For current status data, Huang (1996) provided consistent estimators of covariate effects under Cox's proportional hazards regression model. See Huang and Wellner (1997), for a review of various methods for other regression models, with special emphasis on current status data. See Sun (2006) for an updated summary of regression models and methods for general interval-censored data under the assumption of independent censoring.

Mirzaei and Sengupta (2015) considered regression under Cox's model for the special type of dependent censoring arising from recall data with the possibility of non-recall. When this model is combined with the likelihood (5), the resulting likelihood becomes

$$\prod_{i=1}^{n}[\bar{F}_i(S_i|Z_i)]^{1-\delta_i}\left[\{f_i(T_i|Z_i)(1-\pi(S_i-T_i))\}^{\varepsilon_i}\left(\int_0^{S_i} f_i(u|Z_i)\pi(S_i-u)du\right)^{1-\varepsilon_i}\right]^{\delta_i}, \tag{8}$$

where

$$\bar{F}_i(t|Z_i) = [\bar{F}_0(t)]^{\exp(\beta Z_i)}, \tag{9}$$

$f_i(t|Z_i)$ being the derivative of $F_i(t|Z_i)$. The above likelihood is meant to be maximized with respect to β, the (possibly vector) parameter used to describe the function π and the unspecified function F_0. Mirzaei and Sengupta (2015) simplified the optimization problem by (a) removing integrals through a piecewise constant form of π and (b) restricting probability allocations of the baseline distribution to the distinct times of precisely recalled events. Their simulation results show that chi-square tests of significance, obtained from the likelihood in the conventional manner after disregarding the nonparametric nature of the likelihood and the approximations involved, produce reasonably reliable p values. An R program for fitting this model is available from the authors on request.

6 Imperfect or Erroneous Recall

As mentioned in Sects. 2 and 4, it is possible that respondents may recall only a range of time for the event of interest. Mirzaei et al. (2016) found in the case of a menarcheal data set (partially analyzed in the next section) that, rather than remembering a range of ages for the age at menarche, respondents often remember a range of calendar dates for the occurrence of the event. Thus, the different types of partial recall can be grouped into recalling the month of occurrence, the year of occurrence, and so on, apart from the scenario of no recall at all. They proposed a multinomial logistic model for the recall probabilities and were able to extend the parametric method reported in Sect. 4 to this situation. Work on nonparametric estimation and extension of the Cox regression model is in progress.

It should be noted that the recalled time-to-event can sometimes be erroneous (see Beckett et al. 2001). Grouping of the cases of partial recall as above might reduce the impact of recall error somewhat, but would not address the issue specifically. There have been some attempts to incorporate this fact into the modeling through latent variables (see, e.g., Rabe-Hesketh et al. 2001). However, adapting such modeling to recall data would require further research.

7 Data Analysis

The data we use here are based on an anthropometric study conducted by the Indian Statistical Institute in and around the city of Kolkata, India, from 2005 to 2011 (Dasgupta 2015, p.108). A total of 2195 randomly selected females, aged between 7 and 21 years, were surveyed. The subjects were interviewed on or around their birthdays. The data set contains age, some physical information of each individual, menarcheal status, age at menarche (if recalled), and some socioeconomic information. For this data set, the landmark event is the onset of menarche. Among the 2195 cases in the data set, 775 individuals did not have menarche, 443 individuals recalled the exact date of the onset of menarche, 276 and 209 individuals recalled the

Table 1 Estimated parameters and median age at menarche from different methods for real data

Estimator	Estimate (standard error)			Median	95% Confidence interval of median
	α	β	η		
Current status MLE	10.74 (0.320)	12.17 (0.005)		11.76	(11.62, 11.90)
Interval censored MLE	11.80 (0.061)	12.65 (0.001)		12.25	(12.20, 12.30)
Binary recall MLE	10.19 (0.090)	12.21 (0.001)	3.47 (0.140)	11.78	(11.72, 11.84)

calendar month and the calendar year of the onset, respectively, and 492 individuals could not recall any range of dates. Thus, the data are interval-censored. A major goal of this study was to estimate the distribution of the age at onset of menarche and the dependence of age at menarche on socioeconomic variables. For simplicity, we dichotomize the recalled information; i.e., we club the cases of partial and no recall and refer to them as cases of no exact recall.

To illustrate the parametric approach, we used the Weibull model for menarcheal age and the exponential model with scale parameter η for non-recall probability. We compared the performance of MLEs based on the current status likelihood (1) (described here as current status MLE), the likelihood (4) based on interval-censored data with noninformative censoring (described here as interval-censored MLE), and the likelihood (5) based on binary recall information when the censoring mechanism is recognized as informative (described here as binary recall MLE). Computation of MLEs in all the cases is done through numerical optimization of likelihood using the quasi-Newton method (see Nocedal and Wright 2006). Table 1 gives a summary of the findings. The interval-censored MLE of the median is somewhat different from the other two MLEs, which is possibly because of the bias of the former. The binary recall MLE has a narrower confidence interval for the median than the current status MLE.

As another illustration, in Fig. 1, we compare graphically the closeness of the parametric estimator of the time-to-event distribution with the AMLE (see Sect. 4), for the menarcheal data set when a piecewise constant model of π with $k = 8$ is used for the non-recall probability. The jump points of the piecewise constant function are assumed to be evenly distributed over the range 0–13 years (maximum possible separation between menarcheal age and age at observation in the sample). The two estimators are somewhat close to one another.

The age at menarche can potentially be affected by diet and physical activities. These factors can be related to more easily measured socioeconomic variables such as parents' education and monthly family expenditure (Khan et al. 1996; Padez 2003; Aryeetey et al. 2011). We considered the monthly family expenditure in Indian rupees (indexed with respect to 2008 as base year) and a couple of binary variables indicating

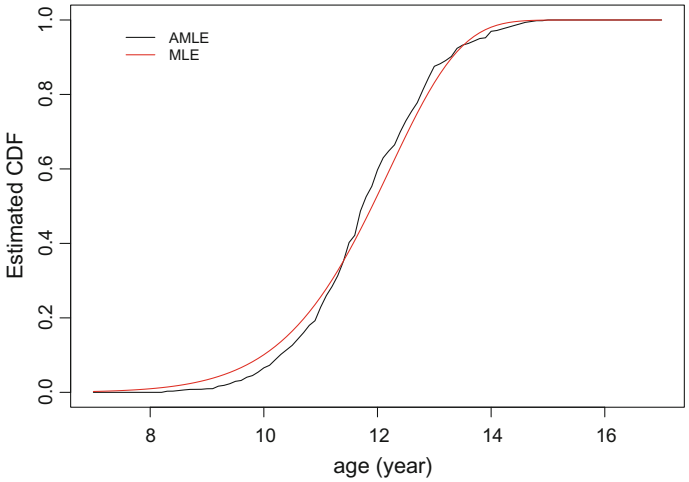

Fig. 1 Comparison of MLE and AMLE of menarcheal age distribution

Table 2 Estimated regression coefficients and their p values

Covariates	Estimated value	p value
Whether father passed high school	0.091	0.0036
Whether mother passed high school	0.249	0.0061
Monthly family expenditure	0.0002	0.0047

whether the father or the mother of the respondent had passed high school. The present analysis concerns a subset of the original data, consisting of respondents who came from a nuclear family and were the only child of their respective parents. Among the total of 673 respondents, 241 individuals did not have menarche, 147 individuals had menarche and recalled the date of its onset, and the remaining 285 individuals had menarche but could not recall the date. The median of monthly family expenditure was Rupees 7808. The fathers of 492 respondents and the mothers of 420 respondents had passed high school.

The estimated regression coefficients and the corresponding p values are reported in Table 2. All the coefficients are found to be significant at the 1% level. The p value of the combined hypothesis of insignificance of all the three regression coefficients is 0.00093.

We now consider four hypothetical subjects with covariate profiles Z described below.

$Z = (0, 0, 7808)$: Monthly family income is Rupees 7808 (median income of the group), neither parent passed high school.

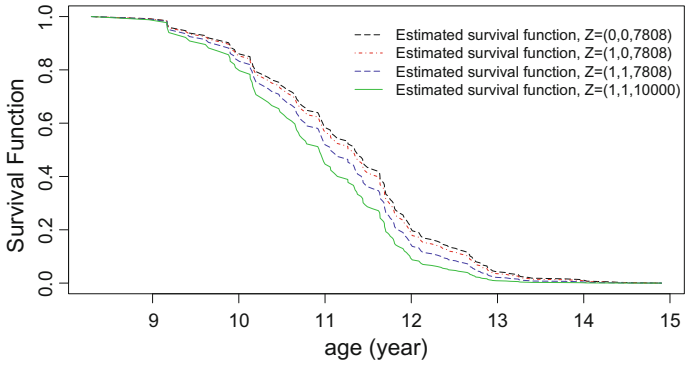

Fig. 2 Estimated survival function in different cases

$Z = (1, 0, 7808)$: Monthly family income is Rupees 7808, only the father passed high school.

$Z = (1, 1, 7808)$: Monthly family income is Rupees 7808, both the parents passed high school.

$Z = (1, 1, 10000)$: Monthly family income is Rupees 10000, both the parents passed high school.

A comparative plot of the estimated survival functions of these four subjects is given in Fig. 2. Father's status of having passed high school is found to be associated with earlier maturation. The mothers having the same qualification are seen to have an even greater impact in the form of earlier maturation. A 28% higher monthly family expenditure is also found to have a considerable impact on the survival function of the age at menarche.

8 Concluding Remarks

Cross-sectional time-to-event data obtained from recall have been found to be surprisingly complex in terms of the nature of incompleteness. Many interesting questions have been answered in recent years through careful modeling, and many more remain to be answered. We have indicated in Sect. 5 how the Cox regression model can be fitted in the case of recall data with the possibility of non-recall. Fitting of other regression models and adapting such models to partial recall data remain to be explored. Further challenges include handling of recall error and of random effects (frailty).

Acknowledgements This research was partially sponsored by the project 'Physical growth, body composition and nutritional status of the Bengal school aged children, adolescents, and young adults of Calcutta, India: Effects of socioeconomic factors on secular trends,' funded by the Neys-Van Hoogstraten Foundation of the Netherlands. The authors thank Professor Parasmani Dasgupta, leader of the project, for making the data available for this research. Also, the first author thanks

Dr. Bibhas Chakrobarty for his financial support through the Duke-NUS start-up grant R-913-200-074-263, the NIH grant 1 R01 DE023072-01 and the Singapore Ministry of Education grant MOE2015-T2-2-056.

References

Aksglaede, L., Sorensen, K., Petersen, J. H., Skakkebak, N. E., & Juul, A. (2009). Recent decline in age at breast development: The Copenhagen puberty study. *Pediatrics*, *123*, 932–939.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98.

Aryeetey, R., Ashinyo, A., & Adjuik, M. (2011). Age at menarche among basic level school girls in Medina, Accra. *African Journal of Reproductive Health*, *103*, 103–110.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, *26*, 647–647.

Beckett, M., DaVanzo, J., Sastry, N., Panis, C., & Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *Journal of Human Resources*, *36*, 593–625.

Bergsten-Brucefors, A. (1976). A note on the accuracy of recalled age at menarche. *Annals of Human Biology*, *3*, 71–73.

Cameron, N. (2002). *Human growth and development*. Academic Press.

Chumlea, W. C., Schubert, C. M., Roche, A. F., Kulin, H. E., Lee, P. A., Himes, J. H., et al. (2003). Age at menarche and racial comparisons in us girls. *Pediatrics*, *11*, 110–113.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Dabrowska, D. M., & Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, *83*, 744–749.

Dasgupta, P. (2015). Physical growth, body composition and nutritional status of Bengali school aged children, adolescents and young adults of Calcutta, India: Effects of socioeconomic factors on secular trends. (in collaboration with M. Nubé, D. Sengupta & M. de Onis). http://www.neys-vanhoogstraten.nl/wp-content/uploads/2015/06/Academic-Report-ID-158.pdf

Demirjian, A., Goldstien, H., & Tanner, J. M. (1973). A new system of dental age assessment. *Annals of Human Biology*, *45*, 211–227.

Eveleth, P. B., & Tanner, J. M. (1990). *Worldwide variation in human growth* (2nd ed.). Cambridge University Press.

Finkelstien, D. M., Goggines, W. B., & Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring. *Biometrics*, *58*, 298–304.

Gentleman, R., & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, *81*, 618–623.

Hediger, M. L., & Stine, R. A. (1987). Age at menarche based on recall data. *Annals of Human Biology*, *14*, 133–142.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. John Wiley.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis* (2nd ed.). Hoboken: John Wiley.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*, *24*, 540–568.

Huang, J., & Wellner, J. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: John Wiley.

Keiding, N., Begtrup, K., Scheike, T. H., & Hasibeder, G. (1996). Estimation from current status data in continuous time. *Lifetime Data Analalysis*, *2*, 119–129.
Khan, A. D., & Schroeder, D. G., Reynaldo, M., Haas, J. D., & Rivera, J. (1996). Early childhood determinants of age at menarche in rural Guatemala. *American Journal of Human Biology*, *8*, 717–723.
Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. New York: Springer-Verlag.
Korn, E. L., Graubard, B. I., & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology*, *145*, 72–80.
Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). New York: John Wiley.
LeClere, M. J. (2005). Modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, *4*, 5–12.
Lee, E. T., & Wang, J. W. (2003). *Statistical methods for survival data analysis*. John Wiley.
McKay, H. A., Bailey, D. B., Mirwald, R. L., Davison, K. S., & Faulkner, R. A. (1998). Peak bone mineral accrual and age at menarche in adolescent girls: A 6-year longitudinal study. *Journal of Pediatrics*, *13*, 682–687.
Mirzaei, Salehabadi S., & Sengupta, D. (2015). Regression under Coxs model for recall-based time-to-event data in observational studies. *Computational Statistics & Data Analysis*, *92*, 134–147.
Mirzaei, Salehabadi S., & Sengupta, D. (2016). Nonparametric estimation of time-to-event distribution based on recall data in observational studies. *Lifetime Data Analysis*, *22*, 473–503.
Mirzaei, Salehabadi S., Sengupta, D., & Das, R. (2014). Parametric estimation of menarcheal age distribution based on recall data. *Scandinavian Journal of Statistics*, *42*, 290–305.
Mirzaei, S. S., Sengupta, D., & Ghosal, R. (2016). Estimation of menarcheal age distribution from imperfectly recalled data. Applied Statistical Unit, Technical Report No. ASU/2016/4, Indian Statistical Institute. http://www.isical.ac.in/asu/TR/TechRepASU201604.pdf
Nocedal, J., & Wright, S. J. (2006). *Numerical optimization*. New York: Springer.
Padez, C. (2003). Age at menarche of schoolgirls in Maputo, Mozambique. *Annals of Human Biology*, *30*, 487–495.
Rabe-Hesketh, S., Yang, S., & Pickles, A. (2001). Multilevel models for censored and latent responses. *Statistical Methods in Medical Research*, *10*, 409–427.
Roberts, D. F. (1994). Secular trends in growth and maturation in British girls. *American Journal of Human Biology*, *6*, 13–18.
Salsberry, P. J., Reagan, P. B., & Pajer, K. (2009). Growth differences by age of menarche in African American and white girls. *Nursing Research*, *58*, 382–390.
Scharfstein, D., & Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, *89*, 617–634.
Shiboski, S. C., & Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, *87*, 360–372.
Sun, J., & Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, *88*, 1449–1454.
Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.
Teilmann, G., Petersen, J. H., Gormsen, M., Damgaard, K., Skakkebaek, N. E., & Jensen, T. K. (2009). Early puberty in internationally adopted girls: Hormonal and clinical markers of puberty in 276 girls examined biannually over two years. *Hormone Research Paediatrics*, *72*, 236–246.
Turnbull, Bruce W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, *38*, 290–295.
Vizmanos, B., Marti-Henneberg, C., Clivillé, R., Moreno, A., & Fernández-Ballart, J. (2001). Age of pubertal onset affects the intensity and duration of pubertal growth peak but not final height. *American Journal of Human Biology*, *13*, 409–416.
Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis (with discussion). *Statistics in Medicine*, *11*, 1871–1879.

Mathematical Aptitude and Family Income in North-Eastern Tribes

Ratan Dasgupta

Abstract Aptitude test scores have applications in predicting success of students on future career, appropriate choice of subjects in higher studies, etc. Scholastic Aptitude Test (SAT) scores are known to be positively associated with family income. SAT has mathematical aptitude as one of the test components. It is of interest to investigate for possible relationship of mathematical aptitude and family income among the tribals in Northeast India for proper choice in career selection. We study mathematical aptitude and its relationship with family income among North-eastern tribal students on several occasions. The cross-sectional survey is conducted relatively recently within the time range of about 6 months during 6 March 2017 to 2 September 2017 in Agartala, Tripura, after a region-wise stratification of academic institutions at the first stage. The students are then selected by simple random sampling from different academic institutions on several time points. In the lowest rung of economic condition with poor monthly income, the aptitude scores are seen to be low. A slowly increasing trend in scores is seen with family income by nonparametric LOWESS regression, especially towards higher-income groups. Association of variables by Spearman rank correlation and Kendal's τ indicate little or no relationship of income and aptitude scores, in contrast to LOWESS growth curve. Correlation between LOWESS predicted score and log (family income) is 0.7635 among data from 154 students, indicating a positive association of the variables. Intra-class correlation in collected data indicates lack of cohesion in aptitude scores within groups, and the same holds for the aggregate of all students over five groups in the conducted survey, implying that the scores do not resemble much with each other.

Keywords Mathematical aptitude score · Proliferation rate · Scholastic Aptitude Test (SAT) · LOWESS regression · Rank correlation · Intra-class correlation (ICC) · Velocity of scores

MS subject Classification 62P25 · 60G20

R. Dasgupta (✉)
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203 B T Road, Kolkata 700108, India
e-mail: rdgupta@isical.ac.in; ratandasgupta@gmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_10

1 Introduction and the Methodology

Aptitude tests are useful in guiding individuals for selection of appropriate vocation. Scholastic Aptitude Test (SAT) is one such test in use for admission in different institutions. The SAT scores are positively associated with family income; SAT has mathematical aptitude as one of the components in test. The present cross-sectional study is conducted in different schools of Agartala, capital of Tripura. We investigate the relationship of mathematics aptitude scores with family income for North-eastern tribes in selected sample of tribal students coming from all over Tripura. Some of the selected tribal students stay in hostels, rented house, relatives' places, while studying in the capital for education. Academic institutions selected from the capital Agartala of Tripura are stratified by local regions. Simple random sampling is adopted for selecting students from different institutions in five occasions during the time period 6 March 2017 to 2 September 2017. The test administered consists of questions on basic mathematics, numbers, logical reasoning, etc. The scores are expressed in percentage for analysis.

Nonparametric LOWESS regression shows a positive association of the variables. A slight increasing trend of scores is seen with family income. Proliferation rate of aptitude scores on family income is computed by different techniques. These reveal a scaled change in velocity of scores on income at different levels, identifying sensitive regions of income on aptitude scores.

Scholastic Aptitude Test (SAT) is reported to have correlation with family income. There are adjustments planned for levelling the test for wide range of families; see e.g. https://www.washingtonpost.com/news/wonk/wp/2014/03/05/these-four-charts-show-how-the-sat-favors-the-rich-educated-families/?noredirect=on\&utm_term=.ec6a198888b4.

Relationship of scores with family income is also investigated by several non-parametric tests, for example by Spearman rank correlation, Kendall's τ. The cohesiveness of the scores is examined by intra-class correlation.

It appears that the mathematics aptitude score may be influenced by family income to an extent via family environment, but the association seems weak, except for high-income groups identified in the scatter diagram. In the next section, we explain the results obtained from analysis of collected data. A minimum level of income with lowest proliferation rate of score is identified. The results are elucidated in different figures and legends. In Sect. 3, we further discuss the topics related to aptitude.

2 Results

The aptitude scores are expressed in percentage. The percentage scores with multiplicities are represented in Fig. 1. The minimum-width band containing scores of all the tribal students interviewed over different schools is shown in Fig. 2. The band becomes wider towards the completion time of the survey conducted,

indicating increase in score variability at the end of interviews with inclusion of more institutions.

Data scatter of 154 individuals shown in Fig. 3 indicates that for persons from the lowest rung of economic condition with low monthly income, the scores are low. The LOWESS growth curve of score versus income with $f = 2/3$ and three iterations in Fig. 3 shows an upward trend for a slight increase in income from the lowest rung. In an environment where minimal economic requirement of the individuals are fulfilled, efforts towards more education may arise. The curve shows slightly downward trend afterwards with increase in income. Economic contentment may sometimes lead to lack of efforts towards further scholastic improvement. The tendency is seen to be more prominent when the data is analysed with logarithmic transformation (to the base 10) of income. An increasing trend in scores is observed with further improvement of economic condition till the end of growth curve. Urge to improve on social status in terms of higher educational aspiration is common for affluent society members, when the atmosphere is conducive.

Income distributions are skew in general. A lognormal distribution may sometimes explain a skew income distribution. Histogram of logarithm of incomes in the present case indicates possibility of a symmetric distribution in Fig. 4. However, the distribution is seen to be non-normal in Fig. 5 for quantile-quantile plot of normal distribution. The relationship in LOWESS growth curve becomes smoother compared to Fig. 3, when score versus logarithm of income is considered in Fig. 6. The curve in Fig. 6 has a prominent downward trend around the middle portion.

Pearson correlation coefficient between original variables is 0.1357, and the value is lowered to 0.1040 after log transformation is made to income. Log transformation may not be suitable to further linearise the relationship of the variables from this viewpoint. However, correlation between LOWESS predicted score and log (family income) is 0.7635 in the data from 154 students. This indicates possibility of a positive association of the variables.

We compute the proliferation rate $\frac{d}{dx} \log y$ of aptitude score y on family income x in Fig. 7 by adopting a similar technique of [1]. Proliferation rates are useful to identify the sensitive regions of income where rate of change in aptitude scores is critical. The curve shows a minimum proliferation rate at income $x = 1380$ (Rs). A slight modification of the above technique based on median instead of mean provides a relatively smooth proliferation rate curve shown in Fig. 8. Here, again the curve shows a minimum at income level $x = 1380$ (Rs).

Spearman rank correlation assesses monotonic relationships between the variables. The correlation is highest when the two variables are monotonically related, even though the relationship may not be linear. Spearman rank correlation computed in R over the two variables in different groups shows insignificant relationship of mathematical aptitude score and income, except for a small group of 10 individuals. After merging all the groups with 154 individual altogether, the Spearman correlation 0.0926 remains insignificant.

A nonparametric measure of association, Kendal's tau computes the correlation between signs of the differences of attributes to check whether increase in one coordinate has positive or negative sign change in other coordinate. Computed value of

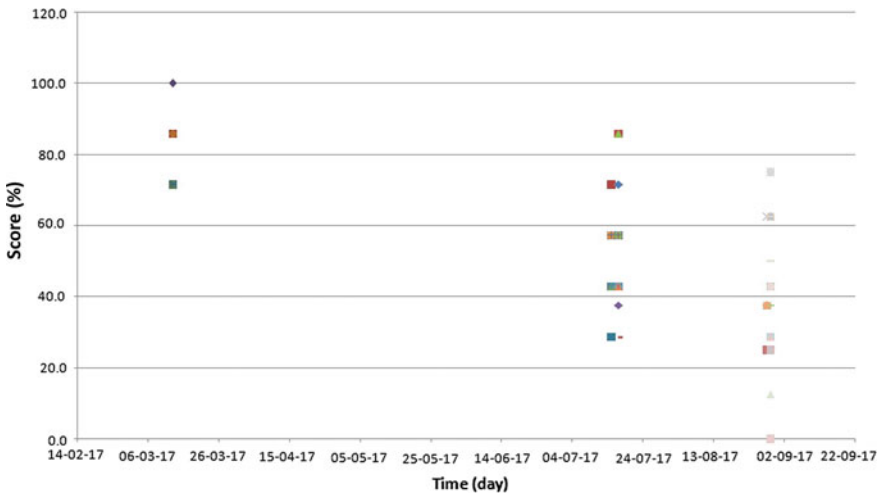

Fig. 1 Mathematical aptitude score in cross-sectional study

τ is 0.0446 for the aggregated group of 154 students. The p value of significance is 0.4735. Thus, the variables are seen to be non-associated by these nonparametric tests.

However, the nonparametric LOWESS regression of the variables explains the variation of two variables reducing the noise part. A slight increasing trend of mathematics score with income is observed towards higher-income group.

Intra-class correlation (ICC) represents cohesion of the variate values; the ICC computed in R is insignificant in all the five groups of individuals, and this remains insignificant in the combined group as well. In the total of 154 individuals, the value is 0.1679, the value is low, indicating units in a group do not resemble strong enough to each other; there are lots of variations among mathematical aptitude scores.

The results are further explained in different figures with legends.

Aptitude scores of interviewed individuals over time are shown in Fig. 1. The conducted study during the time period 6 March 2017 to 2 September 2017 is based on mathematics aptitude questions posed to selected tribal students by simple random sample in different institutions of Agartala, coming from different parts of Tripura. Scores of 154 students selected are recorded in percentage. The first school where the interviews were conducted is a English-medium residential minority high school of tribal students admitted from all over the Tripura State. Other institutions under study are for tribal students of general category in religion. The institutions were selected by stratified random sampling at the first stage. At the subsequent stage, simple random sample of students is considered.

The minimum-width band containing scores of all the tribal students interviewed over different schools is shown in Fig. 2. The band starting from the date 6 March 2017 becomes wider towards the end of interview period, indicating an increase in variability of scores from tribal students in schools of general category in religion.

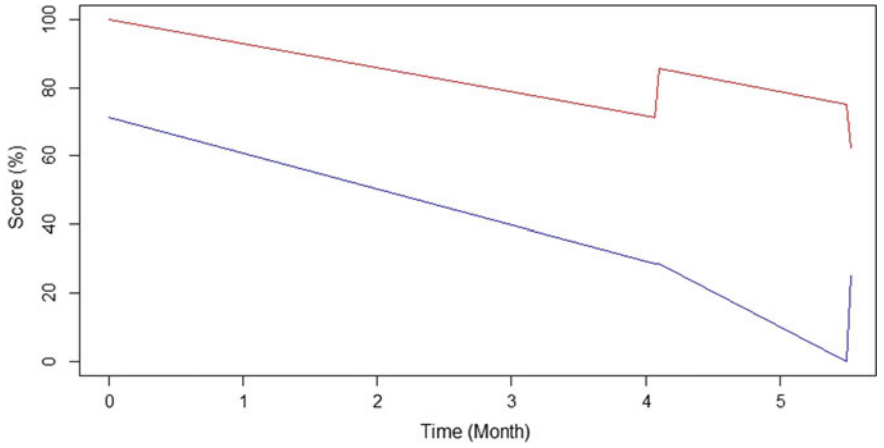

Fig. 2 Upper and lower bands of math scores in (%) of 154 individuals

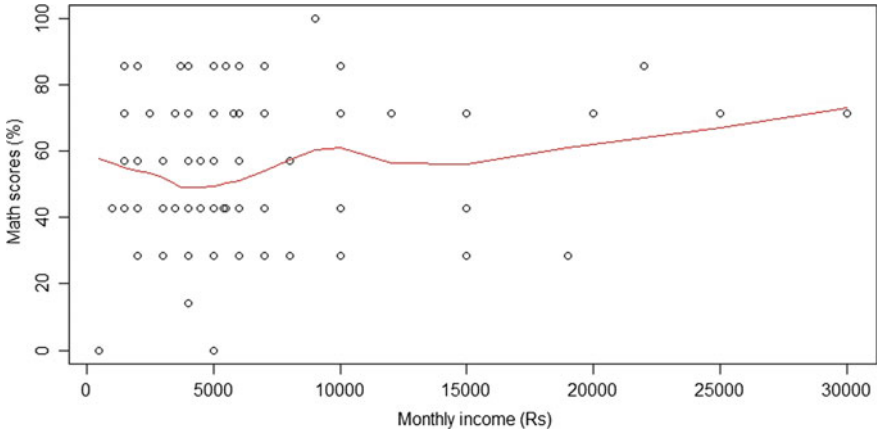

Fig. 3 Scatter diagram of family income versus aptitude score and LOWESS growth curve

The scatter diagram of income versus aptitude score shows slightly increasing trend for data points towards the top right corner. The same is prominent in LOWESS regression with $f = 2/3$ after three iterations. Some initial variations in the curve are also seen.

Logarithm (to the base 10) of family income per month for 154 individuals seems to be symmetric from the histogram shown in Fig. 4. The distribution may not be normal. Distribution of log(family income) is more peaked at the centre compared to a normal deviate.

In the normal Q-Q plot shown in Fig. 5, a distinct departure from normality for log(monthly income) is seen. This feature is also present in the histogram of

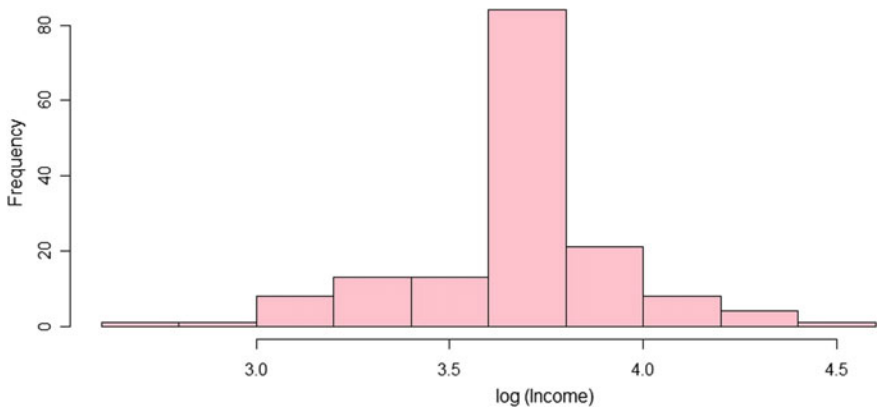

Fig. 4 Histogram of log(family income)

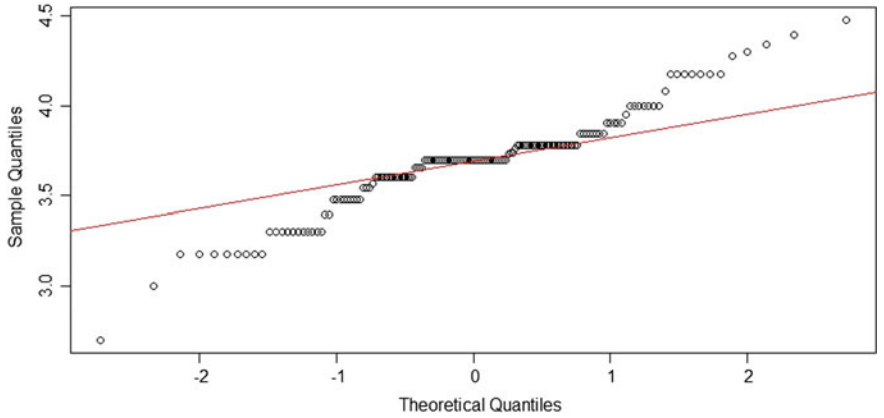

Fig. 5 Normal Q-Q plot for log(monthly income)

the variable in Fig. 4. Sometimes, a skew income distribution become normal after logarithmic transformation, unlike the present case.

A slight upward trend is seen in the LOWESS curve of log(family income) per month versus math aptitude score. The logarithm is taken to the base 10. The curve is obtained by $f = 2/3$ after three iterations. A cluster of the points are more in the middle portion of the graph, as family income is seen to be nearly symmetric after log transformation. The growth of the curve is slightly dampened in the middle region. Growth of score appears to be more for very low- and very high-income group categories.

Proliferation rate $\frac{d}{dx} \log_e y$ is related to the velocity of the growth curve. Proliferation rate $\frac{d}{dx} \log_e y = (\log_e 10) \frac{d}{dx} \log_{10} y$ is a scaled version of velocity $\frac{dy}{dx}$. The logarithm is taken to the base 10 in Fig. 6. So, the graph has to be scaled by a factor $\log_e 10$, to obtain proliferation rates. The measure is independent of the choice of

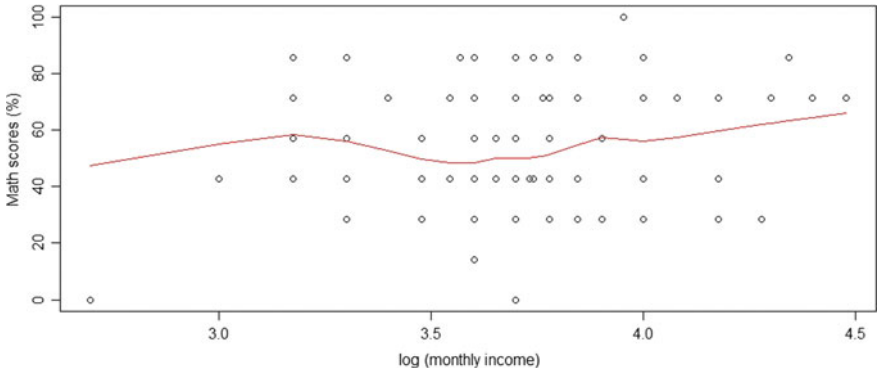

Fig. 6 Scatter diagram of log(family income) versus aptitude score and LOWESS growth curve

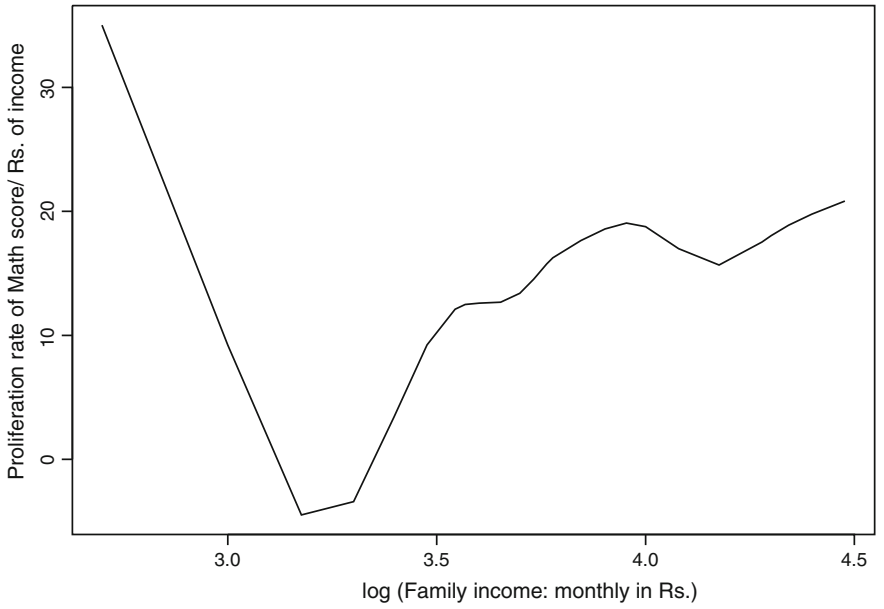

Fig. 7 Proliferation rate of math score, cross-sectional study 2017: wt. exp(−0.001 x); (Fig. 6)

unit used in measuring y. The curve has an initial sharp downward trend till about monthly family income of Rs. 1380, and then, the curve has a sharp rise with minor oscillations till the end. Computation of rate is based on a technique proposed in [1], with exponentially decaying weights $\exp(-0.001x)$ attached to empirical slopes computed from data pairs. Weighted mean of these empirical slopes at derivative stage and smooth.spline with spar $= 0.0001$ at smoothing stage in SPlus provide proliferation rate at income y.

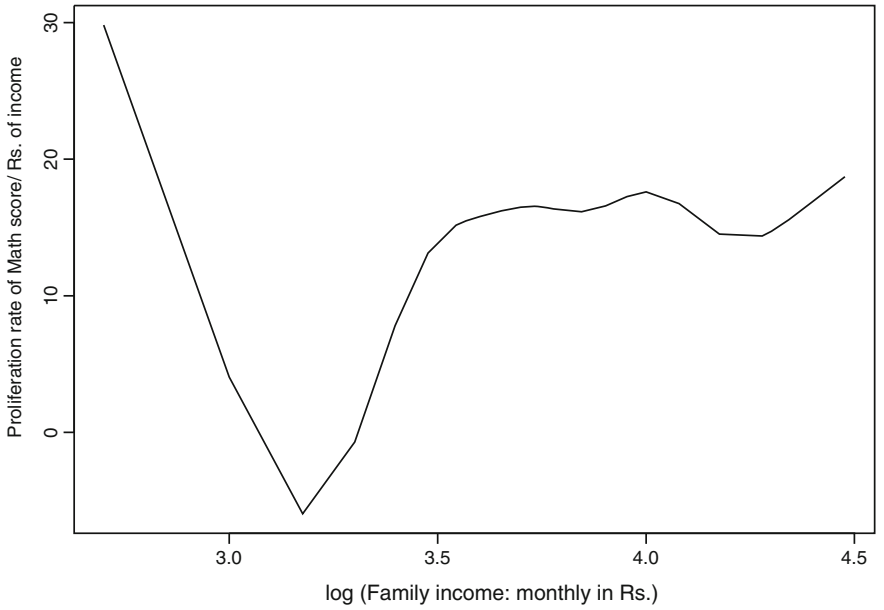

Fig. 8 Proliferation rate of math score 2017: with trimmed mean: wt. exp(−0.001 x); (Fig. 6)

The minimum of rate is observed at $\log_{10}(x) = 3.14$, i.e. when the monthly family income x is about Rs. $10^{3.14} = 1380$.

There are 26 distinct points considered in the growth curve construction in Fig. 6. To obtain Fig. 8, we adopt a similar procedure of Fig. 7, but now order the empirical slopes and consider the trimmed mean, the mean of 13th and 14th ordered observations of these empirical slopes at derivative stage. Next, smooth.spline with spar = 0.0001 at smoothing stage in SPlus yields proliferation rate at income y. Here again, the graph has to be scaled by a factor $\log_e 10$. In contrast to Fig. 7, the rate graph is comparatively smooth in Fig. 8. Both Figs. 7 and 8 show that the rate curves have sharp fall till income of about Rs. 1380. Then, gradual rise in rates with little oscillations towards higher income is seen. The curve has sharper rise from the initial minimum value in Fig. 8, compared to rise in Fig. 7.

3 Discussions

Aptitude is an inherent quality or trait, not quite skills and achievement that may be gained through a learning process. Outstanding performance may not always be attributed to talent; see e.g. [2]. Mathematical aptitude is believed to be an inborn quality; aptitude scores of North-eastern tribal students may be influenced by family environment, including family income.

The curves in Figs. 3 and 6 show a slight increasing trend of aptitude scores with family income. Aptitude scores like SAT scores are dependent on family income.

Aptitude scores can successfully be applied in selection of higher studies and future career of students, and in assessing abilities to perform specific tasks. Stenberg et al. [4] suggest that success in a core business subject is partially dependent on students' mathematical aptitude. Growth of mathematics aptitude scores in tribal students shows a weak increasing trend at higher-income group.

References

1. Dasgupta, R. (2015). Rates of convergence in CLT for two sample U-statistics in non iid case and multiphasic growth curve. In R. Dasgupta (Ed.), *Growth curve and structural equation modeling* (Vol. 132, pp. 35–58). Springer Proceedings in Mathematics & Statistics.
2. Gladwell, M. (2008). *Outliers: The story of success*. New York: Little, Brown & Co.
3. https://www.washingtonpost.com/news/wonk/wp/2014/03/05/these-four-charts-show-how-the-sat-favors-the-rich-educated-families/?noredirect=on&utm_term=.ec6a198888b4
4. Stenberg, L. C., Varua, M. E., & Yong, J. (2010). *Mathematics aptitude, attitude, secondary schools and student success in quantitative methods for business subject in an Australian Catholic University Experience*. In 39th Australian Conference of Economists. http://researchonline.nd.edu.au/bus_conference/26.

Folklore Versus Genetics: A Mitochondrial DNA Investigation About the Origin and Antiquity of the Adi Sub-tribes of Arunachal Pradesh, India

S. Krithika and T. S. Vasulu

Abstract Mitochondrial DNA (mtDNA) polymorphisms of five sub-tribes of Adi tribe of Arunachal Pradesh (Northeast India) were examined with the aim of investigating their extent of genetic variation; genetic relationships (maternal lineage) and population structure (fission–fusion), especially changes in genetic structure as a result of migration and settlement in the recent past history as described in rich folklore tradition. Samples from Panggi, Komkar, Padam, Minyong, and Pasi sub-tribes were analyzed for mtDNA hypervariable regions I and II, where Panggi and Komkar were sampled from Upper Siang district, Padam and Minyong samples were collected from villages of East Siang district, and Pasi was sampled from both East and Upper Siang districts. Macrohaplogroup M shows the highest frequency among Pasi (77.77%) and Padam (71.43%). While all the studied Komkar samples belong to haplogroup N, 60% of the Panggi samples belong to N. Gene diversity (1.000) and nucleotide diversity (0.2072–0.2989) values are high among the sub-tribes. Mean pair-wise differences for Adi sub-groups are found to vary between 8.523 (±4.134276) among Komkar and 11.3187 (±5.467784) among Minyong. *AMOVA* results indicate a fair degree of genetic differentiation among the Adi sub-tribes (F_{ST}: 0.13328). Phylogenetic and principal component analyses (PCA) depict a close cluster of Panggi, Minyong, and Pasi and distant location of Padam and Komkar groups. The size, shape, and pattern of the mismatch distribution vary in each of the

S. Krithika
Department of Clinical and Experimental Epilepsy, Institute of Neurology, University College (UCL), London, UK
e-mail: krithisundar81@gmail.com

T. S. Vasulu (✉)
Biological Anthropology Unit, Indian Statistical Institute, Kolkata 700108, India
e-mail: vasulu@gmail.com; vasulut@hotmail.com

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_11

sub-groups and significantly differ from the theoretical distribution. While Pasi and Minyong tend to show unimodal distribution, Komkar exhibits a bimodal tendency and Panggi depicts a multimodal distribution. The variation of the mismatch distribution curve, among the Adi sub-groups, reflects changes in their demographic size in the recent past that possibly had influenced their mtDNA profiles. The results based on mtDNA is in agreement with Adi folklore accounts of their historical warfare conflicts and tribal feuds which resulted in their fission–fusion population structure among the Adi regional populations.

"Who are we? The answer to this question is not only one of the tasks but the task of Science"
Erwin Schrodinger

1 Introduction

Who are we? An intriguing question to investigate about the peopling of India, its origin, antiquity of ethnically and linguistically diverse populations. This question has been one of the debatable topics of investigation, primarily, in the fields of archaeology, history/prehistory, and linguistics (e.g., Blinkhorn and Petraglia 2017; Vishnupriya et al. 2017; Fuller 2006, 2007); however, with the recent developments in molecular genetics and theoretical population genetic models, we were able to get more clarity and better insight into the antiquity and origin of ethnically, culturally and linguistically diverse castes and tribes, viz. Austro-Asiatic, Dravidian, Indo-European, Andaman and Nicobar and Tibeto-Burman.

The puzzling question of antiquity can be investigated at different dimensions (or levels) of diversity—at national level, at regional level and with respect to (or concerning to) a specific regional caste, tribe, etc. Indeed, there has been studies about the antiquity and the diversity of Indian or South Asian populations based on molecular genetic markers, whole genome sequencing (Mayukh Mondal et al. 2017; Basu et al. 2016; Arunkumar et al. 2012; Mujumdar 2010; Krithika et al. 2013, 2009; Reich et al. 2009; Sahoo et al. 2006; Kivisild et al. 2003 etc.) and to a certain extent at the regional level (e.g., Mondal et al. 2017; Arun Kumar et al. 2015; Chaubey et al. 2015; Tamang et al. 2012; Niraj et al. 2012; Thangaraj et al. 2010; Kumar et al. 2007; Krithika et al. 2006a, b; Kraaijenbrink et al. 2014).

In general, the question of antiquity can be better investigated in case of populations which preserves, over generations, either historical records or tradition or cultural attributes (e.g., myths, folklore) that describe the origin and antiquity of a population. Several local castes and tribes maintain folklore tradition as part of their cultural rituals, etc., that describe their origin and antiquity which can guide us to investigate through molecular genetic markers. Such studies among a few specific populations have indeed given us an insight into the origin and antiquity, e.g., Roop Khand (Harney et al. 2017), Bill tribe (Chaube et al. 2015), Roman Zypsies (Niraj Rai et al. 2012), Siddis (Shah et al. 2011), among Arunachal tribes (Krithika et al.

2006b), Adi tribes (Krithika et al. 2013), among Himalayan tribes (Kraaijenbrink et al. 2014; Hackinger et al. 2016). In this connection, the Tibeto-Burman-speaking populations in northeastern region, with their recent migration and settlement history, diverse traditional lifestyle, and folklore tradition provide opportunities to investigate the antiquity and origin.

Northeast India is the abode of rich regional diversity of populations that can be noticed in their population structure and demographic parameters, such as variation in population size, patriarchal and matriarchal marriage systems; different languages, such as Indo-European, Austro-Asiatic, and Tibeto-Burman, culture, food habits, and occupations ranging from hunting–gathering, shifting culture to settled agriculture (Dani 1960; Elwin 1959; Bhasin and Walter 2001). The extent of diversity of this region can be ascribed to influx of migrations of diverse populations and their intermingling in different parts of the region (Rapson 1955; Ruhlen 1991). Among the different language families representing Northeast India, a majority the tribes speak Tibeto-Burman language and show wide variations in their cultural traits, living standards: Some are hunters–gatherers, shifting and settled agriculture; and some of them live in isolation with least contact with others, and population structure variables like size, distribution, marriage patterns and degree of endogamy (Majumdar 1980; Singh 1998). Apart from our understanding the regional diversity, these populations also help us to understand the extent of relationship with the east and southeastern populations to whom they are culturally, linguistically, and physically related (Elwin 1959; Rapson 1955; Ruhlen 1991; Majumdar 1980; Singh 1998).

Previous anthropological and genetic studies pertaining to the Tibeto-Burman groups were few and limited to some local tribes (Das et al. 1980; Roychoudhury 1981; Walter et al. 1986; Deka et al. 1988; Roychoudhury 1992; Cordaux et al. 2003; Kashyap et al. 2004; Kumar et al. 2004a). However, there are hardly any publications on tribes from Arunachal Pradesh, except a few recent studies by Indian Statistical Institute, Kolkata (Krithika et al. 2006; Maji et al. 2007; Krithika et al. 2007). In this regard, the Tibeto-Burman-speaking populations inhabiting the easternmost tip of Northeast India, Arunachal Pradesh, (sharing the international border between India and Bhutan, Tibet, Myanmar) were hardly dealt with and hence there exist a dearth of population genetic studies in this region, except a few in the recent past. In this regard, it is worthwhile to state that Arunachal Pradesh offers unique opportunities for population genetics studies, as a majority of these tribes (about 20) still live in relative isolation and continue to maintain their traditional way of life in different parts of the state—despite some changes in their lifestyle activities of some of the tribes living near urban areas as a result of developmental activities during the last three decades. And as communication and other facilities have improved over the recent times, some tribes inhabiting the low-level regions, adjoining to the Assam, have had contacts and interactions with other populations as a result of weekly markets, migration (Arunachal Pradesh Human Development Report 2005 (2006)). Of the 20 major tribes, Adi is one major tribal cluster confined to central river valley regions of the state.

Like several other major tribes, Adi tribal cluster consists of several localized sub-tribes in a neighborhood area abutting the Siam and Subansiri river valleys of

the southeastern extension of the Himalayan mountains. About a few decades ago, the tribe used to practice hunting–gathering subsistence economy along with shifting cultivation or zoom cultivation. Interesting aspect of the Adi tribal culture is their folklore tradition that they continue to maintain over generations from their ancestors and narrate and celebrate during their annual ritual ceremonies. The folklore tales and songs narrate and describe about their original abode, dispersal, and migration, over a period of five hundred years, from northeastern Himalayan region toward southern lower Himalayan mountain valley down the Brahmaputra river and its tributaries of Siam and its several branches. The folklore also tells details about their subdivision and warfare and feuds over time and as a result the division and formation of sub-tribes divided or cleavaged or fissioned out along the clinal and related families. This is typical of fission-and-fusion process of population structure which has been studied among similar such tribes in other parts of world, who lived as hunters–gatherers, for example, South American tribes (Yanomama, Makritare Indians, etc.) in Asia and Africa. The warfare and internal tribal feuds also lead drastic demographic changes in population size, marriage pattern, and formation of either new sub-tribes and/or clubbing of small tribes as a larger single sub-tribe that migrate and live in another nearby river valley. From the population genetic point of view, the fission–fusion population structure of population subdivision and formation of sub-tribes provide opportunities to investigate the microevolutionary process. This also aids in investigating their origin, genetic affinity, antiquity of the extant tribes, especially in view of some of the tribes, (e.g., Galo (Adi Galong) or Missing) claim separate identity and disregard common ancestry.

Based on mtDNA sequence variation, the present study investigates the antiquity, genetic affinity among the five Adi regional populations: Adi Pasi, Adi Minyong, Adi Panggi, Adi Komkar, and Adi Padam so as to validate their antiquity and common origin as depicted in their folklore tales. The study also explores the possible microevolutionary trends as a result of their fission–fusion population structure in the recent past.

2 Materials and Methods

2.1 Population: Adi Tribal Cluster

Adi and several of its sub-tribes are distributed in the lower hilly terrains of the southeastern Himalayan mountain ranges in temperate and sub-tropical regions in central Arunachal Pradesh. They are distributed in three districts adjoining the river valley of Siang, viz. West Siang, East Siang, Upper Siang, and also in the river valley regions of Upper Subansiri and Dibang Valley in central Arunachal Pradesh (Gordon 2005; Tabi 2006). Morphologically, the tribes show typical morphological traits attributed as that of East Asian. Adi and its sub-tribes speak Adi language, with several regional dialects and have been classified as North Assam branch of Tibeto-

Burman sub-linguistic family (Bradley 1997; Gordon 2005; van Driem 2002, 2004, 2006, 2007a, b, 2013, 2015). The information gathered from several field visits and from the ethno-historical records reveals that the Adi and along with several other regional tribes of the region trace their origin from a common ancestral pool called as *Tani* group who had lived in Tibet about a few centuries ago—5th–7th century AD (Tabi 2006; Lego 2005; Nath 2000; Roy 1997). During their southeastward migration along the Brahmaputra river, in due course of time, the Tani group has fractioned into several groups and drifted and settled at different river valleys in Arunachal Pradesh. The ancestors of Adi, about a few centuries ago, similar to Tani group, in pursuit of their livelihood as hunters-and-gatherers and shifting cultivators (Jhum cultivation), in due course of time, got further divided as sub-groups and drifted and occupied, primarily, along the river valleys of the several tributaries of Siam and other rivers in central Arunachal Pradesh. These regional tribes, during course of time, have diversified to an extent of getting identified as independent groups and as tribes with separate names, with their own dialects, culture—such as dress pattern, customs, marriage regulations, clans, political systems, and socioeconomic activities. As of the present times, there are about 12 regional Adi tribes who live in different mountain regions of the Siang river valley. Based on differences in their dialect, cultural traits religious customs, and historical distribution and settlement pattern along the Siang river valley, the tribe can be recognized or can be classified into two major clusters. Sub-tribes, viz. Minyong, Padam, Shimong, Milan, and Komkar form as one major cluster. The second major cluster includes sub-tribes such as Gallong, Ramo, Bokar, Pailobo, and Bori. Among the 12 tribes, Minyong and Padam are the two largest tribes with a population about a few thousands, whereas the other tribes number only a few hundreds and live in remote and inaccessible areas in upper hilly regions along the eastern extension of Himalayan mountain range (Nath 2000; Roy 1997). Apart from the above 12 tribes, the (Adi) Missing tribe is one large fraction that has got separated from the original Adi ancestors about a few centuries ago and they came down the hills and occupied the lower banks of the Brahmaputra river in upper or northern parts of Assam, though a small fraction of Missing live in central Arunachal plains near Pasighat town. These are the largest and their number will be about 5 lakhs or so. Missing have adapted some of the regional and local culture, religion—some of them follow Hindu tradition and speak Assamese language, etc.

From the folklore tales about their ethno-historical antiquity and origin indicates that different regional Adi tribes were formed due to tribal warfare and conflicts and feuds in recent past (Lego 2005; Tabi 2006). In general, the formation of Adi tribes is typical of fission–fusion population structure that has been observed in other regions of the world among similar tribes, e.g., among South American Indians (e.g., Yanomamo, Makritare, Xavante) and among Southeast Asia (e.g., Semai-Senoi (Neel 1970, 1973; Neel and Salzano 1967; Fix Alan 1975a, b, 1978; Torroni et al. 1992)). Some Adi tribes those live in higher mountain regions—such as Komkar, Pasi (upper), Bokar, Ramo, Bori—are small in size and practice hunting–gathering. Some sections of Adi tribes—such as Padam, Minyong—are large and some of them live in nearby urban areas and work in government jobs, practice settled agriculture, while some of their kin continue to live in remote villages where they practice shifting culti-

vation and hunting–gathering mode of life. As a result of separation and isolation, the Adi tribe shows wide variation in their culture, dress pattern, dialect variation, house types, religious beliefs (animistic, Hindu, Christian), customs, food habits, ornaments, etc. Interestingly, some of the tribes such as Gallon(g) or Gallo, Missing are also treat themselves as independent tribes as well. However, due to recent developmental activities in the region (communication, education market economy), there are changes in their traditional lifestyle and this is expected to result to a pan-Adi tribe on one side, and due to sociopolitical reasons there is a tendency to declare them as separate, e.g., in case of Gallo and Missing. In a way, Adi provide an interesting opportunity to investigate the microevolutionary trends and population genetics aspects of a tribe that has emerging from fission–fusion population structure. As such some of the genetic consequences of the fission–fusion population structure reported from similar such tribes are expected to be validated among the Adi tribes.

2.2 Blood Samples

Blood samples were collected, with the prior informed consent, from healthy voluntary participants belonging to five sub-regional tribes, viz. Adi Pasi (lower and upper), Adi Minyong, Adi Padam, Adi Panggin or Adi Panggi and Adi Komkar of the Adi tribe of Arunachal Pradesh, Northeast India. While Minyong and Padam samples were obtained from the villages of East Siang district (low altitude), Panggi and Komkar were sampled from remote villages of the Upper Siang district (high altitude). Pasi was sampled from both East (lower mountain range) and Upper Siang (higher mountain range) districts (Fig. 1). We have performed AMOVA to examine the inter-group differences in their mtDNA samples collected from the sub-tribes. For the ANOVA, the samples collected from two localities of the Pasi sub-tribe (Lower Pasi and Upper Pasi) were considered together as a single group except while performing AMOVA wherein the groups were treated separately as Adi Pasi-lower and Adi Pasi-upper, respectively. We have conducted several field trips during 2005–2010. During every field trip, we had the approval of the District Circle Officer (DCO) of East, West, and Upper Siang districts and also excellent rapport and support from the "*Gao Burah*" (Village Head) and all the villagers of the villages and collected blood samples, demographic, and other information from different section of the Adi sub-tribes at different locations along the mountain valley of Siang river. The molecular genetic study among the Adi tribe of Arunachal Pradesh has been approved from the committee "Indian Statistical Institute Review Committee for Protection of Research Risk to Humans" and financial support was obtained from Indian Statistical Institute, Kolkata. Selected unrelated samples (97 samples in total), from the studied five Adi sub-tribes, were analyzed for polymorphisms in the control/non-coding regions (HVS1 and HVS2) and some coding regions (7 restriction site polymorphisms and 1 insertion/deletion polymorphism) of the mitochondrial DNA (mtDNA) to characterize the genetic variation, genetic relationships, and population structure of the studied groups.

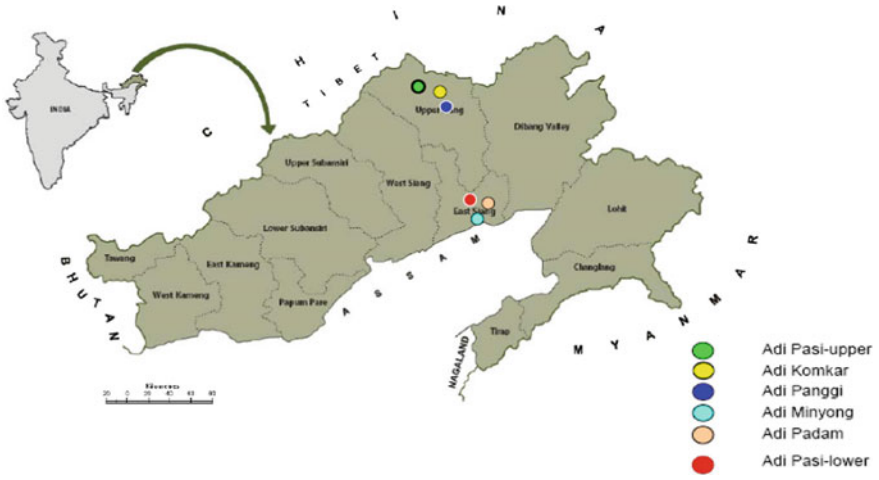

Fig. 1 Map of Arunachal Pradesh showing the geographic distribution of the studied Adi sub-tribes

2.3 DNA Isolation, Amplification, and Sequencing

From the collected blood samples of Adi sub-tribes, high-molecular-weight DNA was isolated using the standard phenol/chloroform method (Sambrook et al. 1989). Hypervariable sequence (HVS) I and II of mtDNA were amplified using the primers L-15997 5′-CACCATTAGCACCCAAAG CT-3′ and H-16391 5′-GAGGATGGTGGTCAAGGGAC-3′, and L-048 5′-CTCACGGGAGCTCTCCATGC-3′ and H-408 5′-CTGTTAAAAGTGCA TACCGCCA-3′, respectively, in 97 individuals by using an ABI Prism BigDye Terminator version 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), as resolved on an ABI 3100 DNA Sequencer (Applied Biosystems, Foster City, USA). Sequences were aligned using BioEdit software (Hall, 1999), and substitutions were reported with respect to the Cambridge Reference Sequence (CRS) (Anderson et al. 1981). Restriction fragment length polymorphism (RFLP) analysis of the mtDNA coding regions (*Hae*III np 663, *Alu*I np 5176, *Dde*I np 10,394, *Alu*I np 10,397, *Hinf*I np 12,308, *Hinc*II np 13,259, *Hae*III np 16517, and COII/tRNA^{Lys} intergenic 9-bp deletion) was performed using methods described elsewhere (Torroni et al. 1993, 1996, 2001). Based on the control and coding region substitutions obtained, haplogroups were assigned to individuals by following the mtDNA phylogeny published earlier (Kong et al. 2006; Yao et al. 2002).

2.4 Statistical Analyses

To understand the degree of genetic variation and the within-population heterogeneity of the studied groups, we estimated the required population genetic parameters, viz., gene diversity, nucleotide diversity, and the mean number of pair-wise differences (along with their standard deviations). The afore-mentioned diversity measures were computed utilizing the software package ARLEQUIN version 3.01 (Excoffier et al. 2005).

To understand the genetic relatedness between the studied Adi sub-populations, F_{ST} distances between pairs of populations and associated *P*-values based on 1000 simulations were computed using the ARLEQUIN package. Subsequently, the F_{ST} distance matrix was used to construct the neighbor-joining (NJ) tree (rectangular and radiation forms) by employing the phylogenetic software MEGA v3.1 (Kumar et al. 2004b).

A covariance analysis has been attempted for the principal component analysis (PCA) components so as to reduce the data dimensionality and obtain clustering pattern of the studied Adi sub-populations. The analysis was attempted using the Fst distance matrix by SPSS software (Version 11.0). Similar clustering in both the PCA plot and the dendrogram indicates the consistency of the results obtained, especially when the bootstrap values of the dendrogram are considerably low. Further, to investigate the phylogenetic relationship between the haplogroups observed among the studied Adi individuals, Median joining (MJ) networks were constructed using the NETWORK 4.5.0.1 program (Bandelt et al. 1999) with default settings.

To examine the genetic variation within and between the sub-populations of Adi, analysis of molecular variance (AMOVA) was performed (Excoffier et al. 1992) using the ARLEQUIN software. The significance of the AMOVA values was estimated by use of 10,000 permutations. This analysis was performed at three levels to examine the influence of geographic isolation and the ethno-historical formation of the sub-tribes that has lead to possible genetic differentiation among the different local Adi sub-tribes. For the first-level analysis, we have considered the six Adi sub-groups as constituents of one group, viz. Pasi-upper, Pasi-lower, Minyong, Panggi, Komkar, and Padam as a "single group." For the second-level analysis for ANOVA, the six Adi populations were classified "two groups." The two groups were considered based on their geophysical locations: The two groups are: [the three sub-tribes: Padam, Minyong, Pasi-lower as "Group 1" and the remaining three sub-tribes: Panggi, Komkar, Pasi-upper as "Group 2"]. The "Group 1" consisting of Padam, Minyong, and Pasi-lower are located at the lower plains of the Siang river valley and geographically separate from the populations of the "Group 2," viz. Panggi, Komkar, Pasi-upper which, on the contrary, are isolated and settled at the higher mountain ranges. For the third level, we have classified the Adi tribes based on ethno-historical information about their migration and settlement history. The three groups considered are: Group 1 included Panggi and Komkar. The two major Adi tribes Padam and Minyong were considered as Group 2 and whereas both Adi Pasi sub-tribes, viz. Upper Adi Pasi and Lower Adi Pasi were considered as the third group.

To investigate the microevolutionary trends of possible past demographic shifts in the population size, as a result of internal tribal warfare, we have considered mismatch distribution based on mtDNA sequences. Further, we have estimated Fu's "Fs" and associated P-values based on 1000 simulations, Tajima's D, and raggedness index "r" with the help of ARLEQUIN (version 3.01). As per Harpending et al. (1993), a unimodal distribution is an indication of recent demographic expansion of the population, whereas the multimodal distribution is a case of constant population size. The estimates of raggedness index "r" are lower than 0.05 and negative and Fu's "F" statistic (Fu 1997) significantly different from zero are indicatives of recent expansion in the population size.

3 Results

3.1 Extent of Mitochondrial DNA Diversity

The various diversity indices including the gene diversity, nucleotide diversity, and the mean pair-wise differences (MPD), among the studied sub-groups of Adi, are shown in Table 1. Gene (haplotype) diversity is found to be high (1.000) among all the six studied Adi sub-tribes, in congruence with the high haplotype diversity (0.984 ± 0.010) reported among Adi in a previous study (Cordaux et al. 2003). This study on mtDNA variation among Indian tribal populations reported the overall haplotype diversity to range from 0.671 to 0.995 among the studied tribal groups sampled from different geographic regions of the subcontinent. Among Indian populations, significantly higher haplotype diversity values were observed in northern, eastern, and northeastern populations (0.940–0.995) than among southern populations (0.671–0.939). Similarly, the nucleotide diversity values, among the studied sub-groups of Adi, are found to be high (0.2072–0.2989). Among the five Adi regional populations, both Adi Pasi and Adi Komkar show the least nucleotide diversity and the highest is observed among the Adi Panggi. The estimates of mean pair-wise differences (MPD) show an average of 10.13 for the Adi population and the values range between 8.523 (±4.13427) and 11.3187 (±5.46778) in Adi Komkar to Adi Minyong, respectively. The lowest values observed among the Adi Komkar could be due to its small population size, high rates of endogamy, and relative isolation since several generations of its formation after an inter-tribal war or feud during earlier times.

3.2 Haplogroup Distribution among the Adi Sub-groups

The frequency distribution of macrohaplogroups M and N and their respective haplogroups and sub-haplogroups, among the six studied Adi sub-tribes, is shown in

Table 1 Diversity parameters deduced from mtDNA HV1 and HV2 sequences of Adi sub-populations

Population	Number of Sequences	Number of polymorphic sites	Gene diversity	Nucleotide Diversity	Mean pair-wise differences
Adi Pasi	19	49	1.0000 ± 0.0171	0.207185 ± 0.110673	10.152047 ± 4.852679
Adi Minyong	14	47	1.0000 ± 0.0270	0.235806 ± 0.127870	11.318681 ± 5.467784
Adi Panggi	19	34	1.0000 ± 0.0171	0.298747 ± 0.159287	10.456140 ± 4.988735
Adi Komkar	18	40	1.0000 ± 0.0185	0.207875 ± 0.112763	8.522876 ± 4.134276
Adi Padam	13	37	1.0000 ± 0.0302	0.255128 ± 0.140121	10.205128 ± 4.986200

Table 2 Frequency distribution of mtDNA haplogroups M and N among Adi

Population	Haplogroup M	Haplogroup N
Adi Pasi	D (38.88%); M* (11.11%); M-G (5.55%); M9 (22.22%)	A (16.66%); B (5.55%)
Adi Minyong	M* (15.38%); D (23.07%); M8 (7.69%); M8 (7.69%)	A (7.69%); F (30.77%)
Adi Panggi	D (20%); M9 (13.33%); M10 (6.66%)	A (33.33%); F (26.66%)
Adi Komkar	–	R (81.25%); A (18.75%)
Adi Padam	M9 (42.86%); D (21.43%); M* (7.14%)	A (14.28%); F (14.28%)

Table 2 and Fig. 2. M shows the highest frequency among Pasi (77.77%; $n = 18$) and Padam (71.43%; $n = 14$). Except Panggi ($n = 15$) and Komkar ($n = 16$), the rest exhibit a higher frequency of haplogroup M than N. While all the studied Komkar samples belong to haplogroup N, 60% of the Panggi samples belong to N.

3.3 Genetic Differentiation Among Adi Sub-groups

AMOVA analyses, based on the mtDNA HV1 and HV2 sequences information of the studied groups, were performed to understand the possible role of geography and ethno-history toward the genetic differentiation of Adi. The results of the analyses are shown in Table 3. When the studied sub-tribes of Adi were considered as a single group, the results show that 13.33% of variation is attributed to the differences among populations, whereas 86.67% of variation is accounted for differences among the individuals within populations. The corresponding F_{ST} value of 0.13328 indicates genetic differentiation, among the studied groups, to a certain extent.

The ethno-historical records suggest that the fission processes among the Adi tribes were the result of their inter-tribal conflicts and the consequent relative geographic isolation that shaped the extant Adi sub-tribes. These two major factors might possibly have played a key role in the genetic differentiation of Adi (Roy 1997; Nath 2000; Lego 2005). Accordingly, AMOVA analyses were considered so as to under-

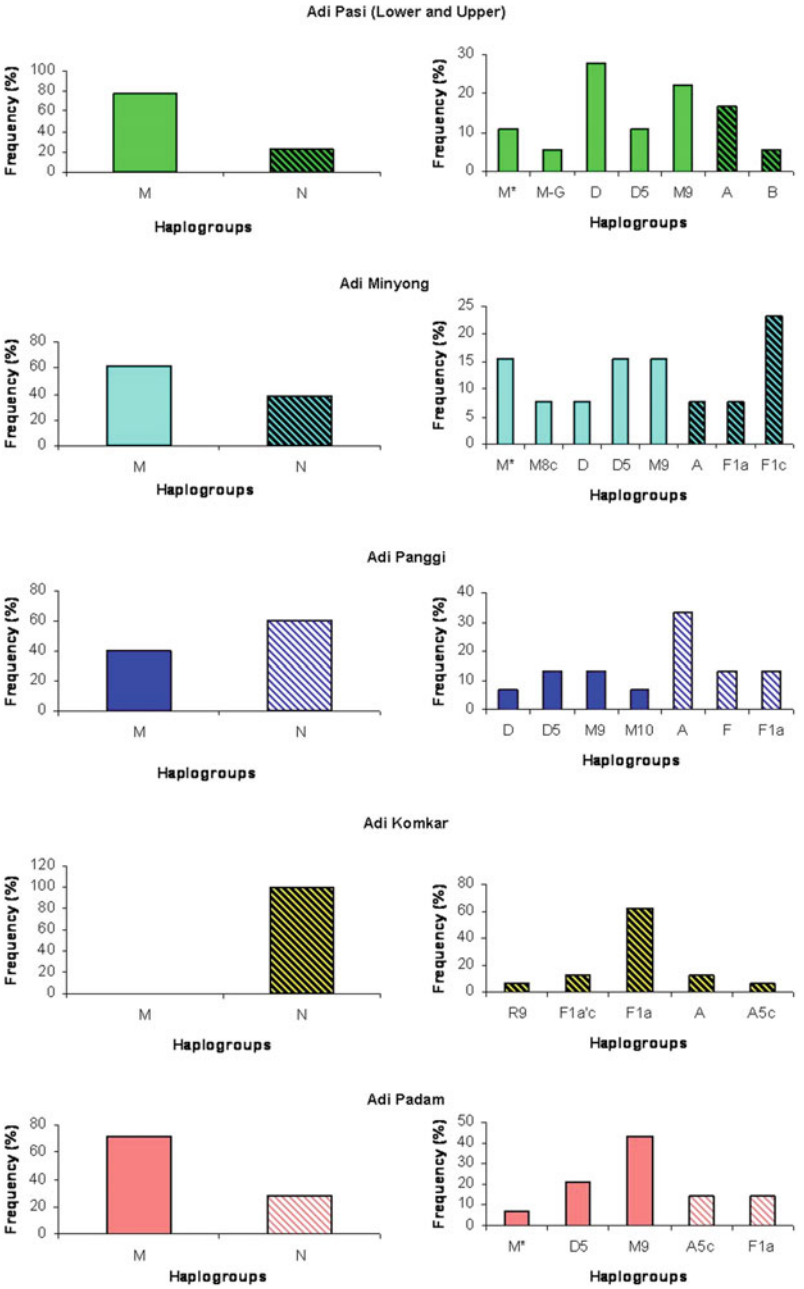

Fig. 2 Frequency distribution of mitochondrial DNA haplogroups among the studied Adi sub-tribes

Table 3 AMOVA based on mtDNA sequences of the studied Adi groups

Grouping	Adi populations in group	Source of variation	% of variation	Fixation indices
Single group	Pasi-upper, Pasi-lower, Minyong, Panggi, Komkar, Padam	Among populations	13.33	F_{ST}: 0.13328
		Within populations	86.67	
Two groups based on geophysical location	(Pasi-lower, Padam, Minyong) versus (Pasi-upper, Komkar, Panggi)	Among groups	−0.43	F_{SC}: 0.13545 F_{ST}: 0.13169 F_{CT}: −0.00435
		Among populations within groups	13.6	
		Within populations	86.83	
Three groups based on ethno-history	(Panggi–Komkar) versus (Padam–Minyong) versus (Lpasi-Upasi)	Among groups	−1.4	F_{SC}: 0.14274 F_{ST}: 0.13076 F_{CT}: −0.01398

stand the relative influences of both these factors toward the genetic differentiation of Adi (Table 3). The results indicate that the grouping of populations based on their geophysical location (F_{ST}: 0.13169) and their ethno-history (F_{ST}: 0.13076) did not reveal any significant differences among the Adi regional populations. In both the above factors considered, around 14% of the variation is accounted for the among the populations and within the groups, whereas the variation was around 86% in case of within the populations (as observed in case of the single group analysis).

To investigate the haplotype diversity among the studied sub-populations, the median joining network analysis was considered (Fig. 3). The network shows five clades, each distinct from the other with respect to the branching pattern and distribution of the haplotypes. Clade "e" is the smallest clade comprising of five individuals from Panggi, Pasi, and Komkar groups. Haplotype sharing is observed within Panggi (2 cases), Pasi (2 cases), Minyong and Komkar (1 case each). Padam does not show any haplotype sharing within its sampled individuals, reflecting wide mtDNA diversity within the group. Of the five clades (group(s) of individuals descendant from a common ancestor) observed among the Adi sub-tribes, the "Clad d" (Fig. 3) comprises mostly of the Komkar individuals (63.15%), thereby indicating close affinity, at the mtDNA level, among the individuals of the Komkar group. This is expected in view of their small size of the population and remote isolated location in upper hills. Also, the clade shows characteristic absence of Pasi and Padam groups, indicating that these individuals were derived from different ancestral individuals. The other four clades, however, show individuals from different sub-tribes, suggesting common ancestry.

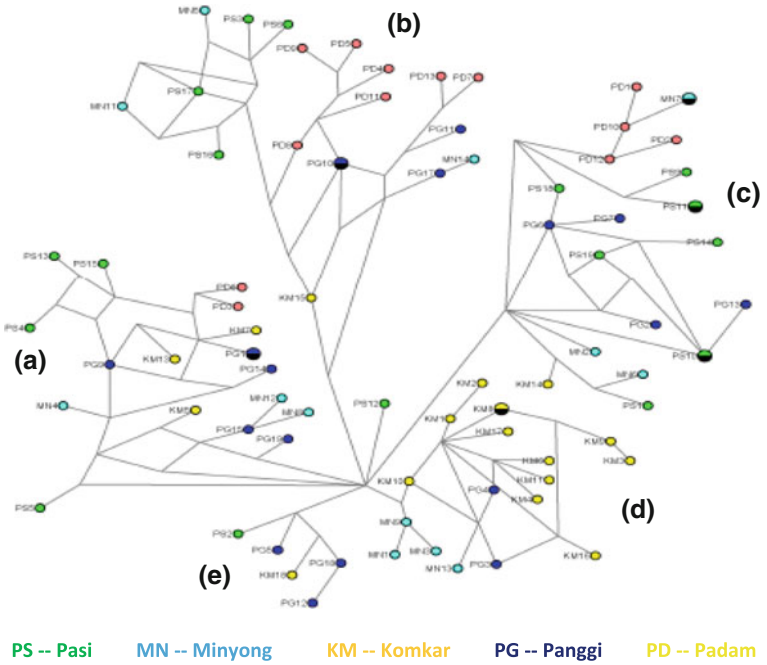

Fig. 3 Median-joining network of the mtDNA haplotypes of Adi sub-tribes

3.4 Genetic Affinity Among Adi Sub-groups

The folklore tales depict about their putative origin from an ancestral population called Tani group from Southern Tibet, and during their migration and settlement they have formed several regional populations and hence they all share a common genetic relatedness. The genetic affinity and relatedness based on mtDNA sequences have been shown in Fig. 4. The NJ phylogeny dendrograms based on Fst distance by rooted and uprooted methods show a close cluster of three populations: Panggi, Pasi, and Minyong, whereas Adi Padam and Adi Komkar deviate from the cluster and show longer branch length. The distant clustering of Adi Padam and Adi Komkar is the reflection of their earlier departure and formation of sub-regional populations. Demographically, both Adi Minyong and Adi Padam are relatively larger than the rest three populations. The PCA plot shown in Fig. 5 shows similar clustering of the three populations, viz. Minyong, Panggi, and Pasi whereas both Komkar and Padam are distantly placed.

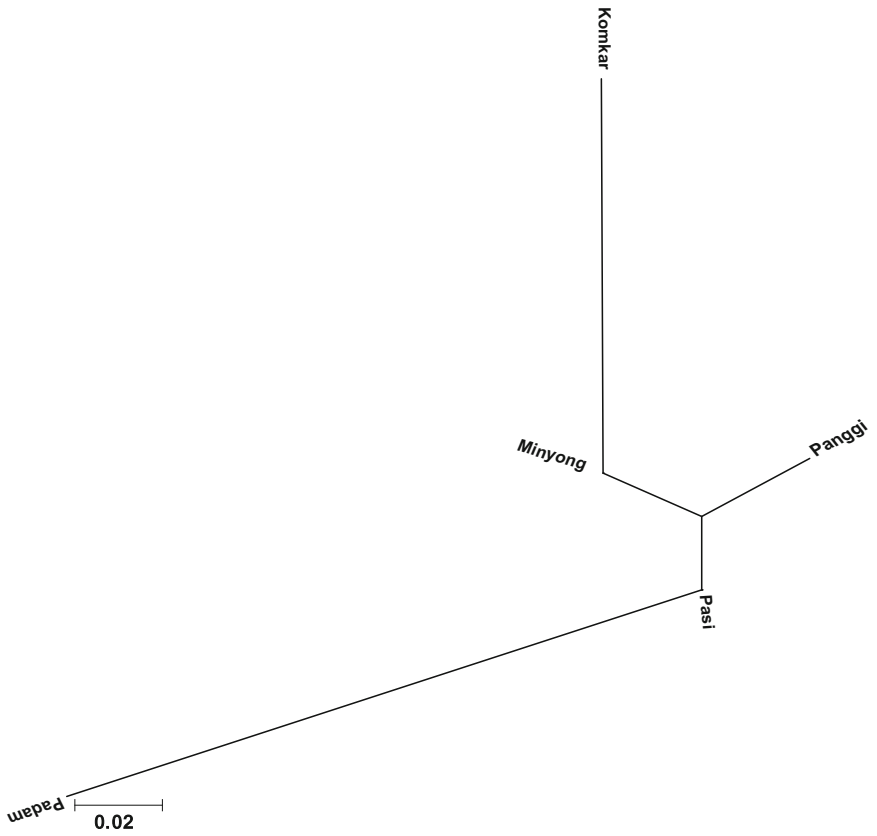

Fig. 4 NJ phylogeny, based on F_{ST} distance matrix, depicting the genetic affinity among the studied Adi sub-groups

3.5 Mismatch Distributions of the Adi Sub-groups

The fusion–fission population structure, a reflection of their inter-tribal warfare and feuds, in recent past, among the Adi tribal populations implies possibility of demographic upheaval, and this gets reflected from a comparison of mismatch distribution (MMD). The MMD obtained from the pair-wise nucleotide differences from each of the five Adi populations and described elsewhere (Krithika and Vasulu 2013). The pattern of MMD, especially the size and shape observed, differs among the Adi populations. Of the five, Adi Minyong, Adi Komkar, and Adi Pasi nearly conform to the theoretical distribution of unimodal distribution, though all the three differ in their mean, mode, and kurtosis and differ also in their overall shape as well. Both Adi Panggi and Adi Padam significantly deviate from the theoretical distribution: Adi Panggi shows a trend of bimodal distribution, though the second mode is less significant, whereas Adi Padam shows multimodal distribution, and shows increasing trend

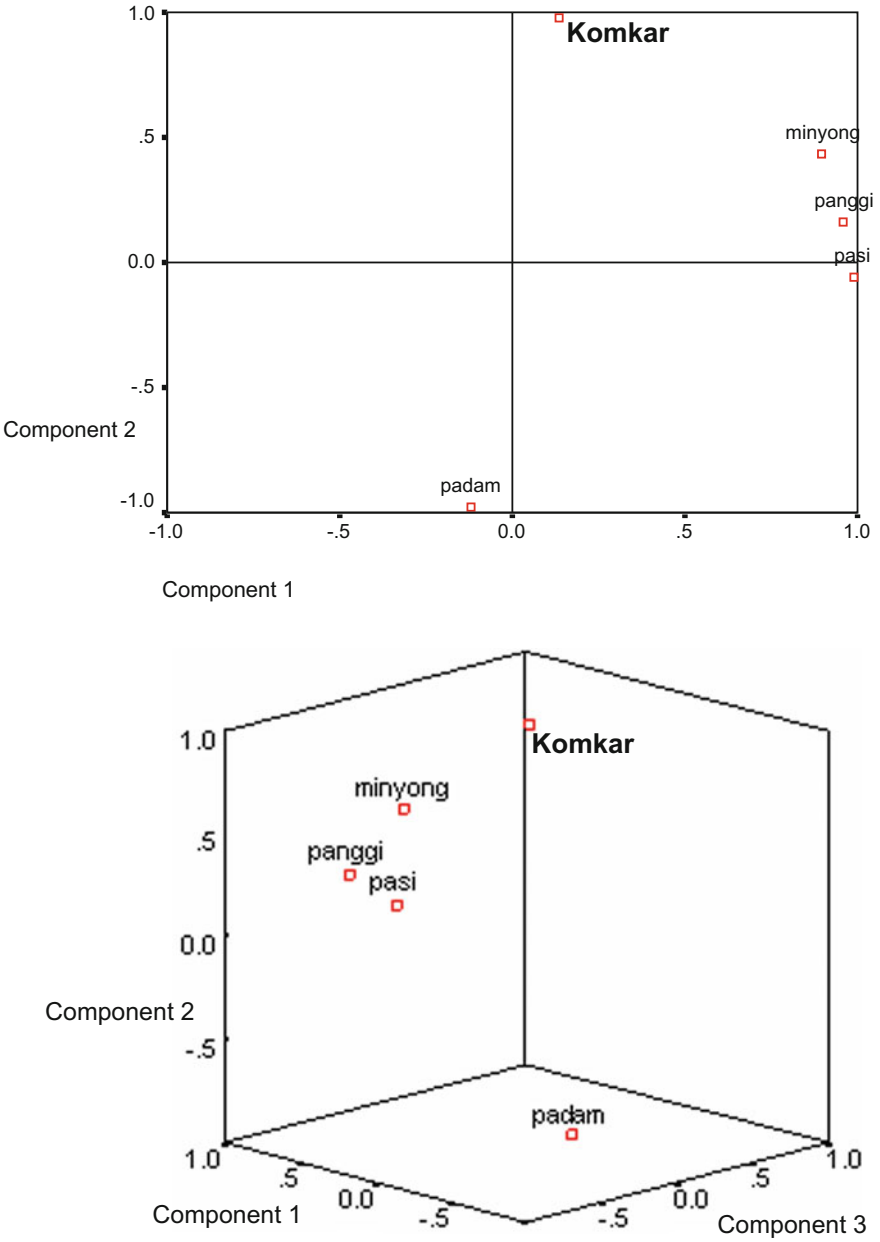

Fig. 5 PCA plots depicting the genetic affinity among the studied Adi sub-groups

of three major crests or modes. Of all the five, Adi Panggi shows wide distribution with a range 1–27 nucleotide differences (NTD), followed by Adi Minyong (1–22) and the rest three, viz. Adi Pasi, Adi Panggi, and Adi Komkar, show near equal range from 1 to 19 (NTD). In general, the three populations, viz. Adi Pasi, Adi Minyong, and Adi Komkar, to an extent, follow the demographic expansion model and as such the statistical estimates of mean, mode, and variance, etc., conform to the theoretical distribution pattern. However, both Adi Panggin and Adi Padam do not follow the expected distribution. The mean of the distributions vary from around 10 (except in case of Komkar (8.5)), and the variance of the distributions vary 12–13, (except Panggi (35.838)).

3.6 Estimates of Demographic Parameters of Adi Sub-groups

Table 4 shows the estimates of different demographic parameters, especially, Fu's Fs, Ragged Index "r", and Tajima's D computed based on mismatch distributions of mtDNA HV1 and HV2 sequences among the five Adi populations. All the five populations show negative estimates in case of Fu's Fs and the values range from −5.6437 and 5.974 in case of Adi Padam and Adi Pasi to −11.535 and −11.218 in case of Adi Komkar and Adi Pasi populations. Consistent and near equal negative and significant values ($p < 0.0$) in the Adi sub-populations suggest that they had experienced a demographic expansion in near past. The higher values suggest the tempo of demographic expansion is more intense among Adi Pasi, Adi Panggi, and Adi Komkar populations than among Adi Minyong and Adi Padam. The high values observed in the three populations were also demographically smaller populations. The estimates of Ruggedness index "r" are all positive, though non-significant. They follow a similar trend that has been observed in case of Fu's F values, viz. they vary from 0.0127 (Adi Komkar), 0.0132 (Adi Pasi), 0.0176 (Adi Panggi) to 0.0272 in Adi Minyong and 0.0362 in Adi Padam. As the values are all below 0.05 and suggest demographic expansion, similar to what was inferred from Fu's f estimates. In case of Tajima's D estimates, except for Adi Panggin, the other four populations show negative values that range from −0.638 in Adi Padam to −0.0132 in Adi Pasi. The positive value observed in case of Adi Panggin suggests demographic upheavals and bottleneck expansion in their recent past history of settlement, and this is also expected to reflect intermediate frequency in some of the polymorphic loci as well. These three demographic parameter estimates which give insight into the trends of changes in their population size changes in their past and possible genetic consequences are in agreement with the mtDNA mismatch distribution pattern of the Adi populations (Krithika and Vasulu 2013).

Table 4 Demographic indices deduced from mtDNA HV1 and HV2 sequences of Adi sub-populations

Population	Number of sequences	Fu's F_S (P-value)	Raggedness index "r" (P-value)	Tajima's D (P-value)
Adi Pasi	19	−11.21824 (0.000)	0.0132 (0.576)	−1.1216 (0.092)
Adi Minyong	14	−5.97461 (0.011)	0.0272 (0.490)	−1.02473 (0.116)
Adi Panggi	19	−10.98509 (0.000)	0.0176 (0.454)	0.2991 (0.602)
Adi Komkar	18	−11.53593 (0.000)	0.0127 (0.855)	−1.08973 (0.096)
Adi Padam	13	−5.64377 (0.005)	0.0362 (0.394)	−0.63823 (0.242)

4 Discussion

Adi tribe of Northeast India comprises of several local or regional sub-tribes (about 13) settled in relative geophysical isolation for a number of generations since their settlement abutting the Siang river of sub-Himalayan mountain extension in Arunachal Pradesh (Roy 1997). These local or regional sub-tribes show differences among them, apart from their population size and distribution they differ widely with respect to their dialect of Adi language and usage of terms of common nouns, kinship terms. Differences are also observed in their religious beliefs, cultural traits, viz. housing types, dress pattern (which specific pattern that identifies the sub-tribe, especially seen in case of "*Gale*", *a lower garment* worn by ladies); food habits, hunting tradition, ornaments (Roy 1997; Nath 2000; Blackburn 2003). The groups also possess a unique folk culture tradition describing their putative origin and different migration histories (Roy 1997; Nath 2000).

The ethno-historical accounts of Adi tribe suggest that the different sub-tribes have been formed at different times, some of them as a splinter group from the larger group as a result of inter-tribal conflicts over resource utilization (Roy 1997; Nath 2000; Lego 2005). As a consequence, each tribe is characteristic of a unique population structure, which can be noticed from their demographic aspects, like size, marriage patterns, and admixture with other tribal groups that vary among the sub-tribes (Lego 2005; Tabi 2006). These differences in their population structure variables that had cropped up for over generations after their isolation and survival will also expected to result in genetic differences between the sub-tribes. Overall, the sub-groups exhibit sociocultural as well as linguistic diversity coupled with wide variation in subsistence pattern (ranging from hunting-gathering to settled agriculture). Therefore, it will be interesting to investigate how far Adi sub-tribes differ genetically among themselves? Now, the question that arises is to what extent the different sub-tribes of Adi are genetically different? And does the unique population structure of each of the Adi

sub-tribes is reflected through the results obtained from the molecular genetic studies? The results obtained reflect considerable genetic diversity among the Adi sub-tribes at the mitochondrial DNA level, which can be observed in the mitochondrial DNA haplogroup distribution, mean number of pair-wise differences and the average gene diversity values.

The distribution of mtDNA haplogroups among the sub-tribes of Adi indicates their wide genetic diversity. Adi sub-tribes show the presence of both M and N sub-haplogroups; while some (e.g., Pasi and Padam) exhibit higher frequency of haplogroup M (~>70%), some others (e.g., Panggi and Komkar) reveal higher incidence of haplogroup N. This wide distribution of the mtDNA haplogroups among the Adi sub-tribes signifies their genetic diversity at the mitochondrial DNA level. The sub-haplogroups of N, which characterize the northeast tribes of India, are East Asian specific and are virtually absent elsewhere in the Indian subcontinent. Cordaux et al. (2003) also reported the presence of East Asian haplogroups A and F, based on HV1 region of mtDNA, among the studied populations of Northeast India (Adi, Apatani, Nishi, Naga, and Tipperah). A recent study by Kumar et al. (2008) also reported the presence of both M and non-M haplogroups among the Tibeto-Burman groups of Northeast India [7: Dirang Monpa, Gallong, Lachungpa, Lepcha, Shertukpen, Toto, Wanchoo].

High gene diversity (1.000) and nucleotide diversity values (0.2072 to 0.2989) among all the studied populations also imply that the Adi regional populations exhibit higher range of values with respect to mtDNA hypervariable regions. The observed high value of gene diversity that is consistent in all the five regional populations are in congruence with that reported among the Northeast Indian populations, e.g., 0.940 among Naga to 0.995 among Tipperah (Cordaux et al. 2003). In case of nucleotide diversity, barring Panggi, Padam followed by Minyong exhibits the highest nucleotide diversity (0.255128 ± 0.140121 and 0.235806 ± 0.127870, respectively) which is expected taking into consideration their marriage practices that includes marriages with other sub-tribes and with other tribes as well (field observation). The least nucleotide diversity values among the Pasi (0.207185 ± 0.110673) and Komkar (0.207875 ± 0.112763) groups might probably be explained due to their small size and remote location which in turn prevents external gene flow into the population. In this regard, it is to be noted that among 19 samples analyzed in Adi Pasi, only four samples belonged to Pasi-lower group while the rest were Pasi-upper. So the least nucleotide diversity among Pasi is expected in view of the population size and relative geographic isolation located on higher mountain ranges of the Pasi-upper group. Unexpectedly, Panggi, being an isolated population, shows the highest nucleotide diversity. Being an isolated population and with small demographic size of maximum of about two thousand, such high nuclear diversity apparently is not expected. However, details of their population structure give some interesting explanation, especially admixture with other groups could be the reason for the higher values of the diversity index. The high nucleotide diversity could plausibly be due to the fact that the Panggi group also consists of some non-Panggi surnames (Maji et al. 2007; Maji and Vasulu 2008).

The different mean values of mtDNA pair-wise differences (MPD) observed among the five Adi populations is in agreement with the trend observed with the nuclear diversity values. The observed MPD values among the Adi are higher than some of the studied northeastern populations in neighboring regions reported by Cordaux et al. (2003) that ranges from 5.73 ± 2.79 among Apatani (Arunachal Pradesh) to 7.17 ± 3.51 among Tipperah (Tripura). The results also reflect their differential population structure as a result of their fission–fusion and warfare in the recent past as has been depicted in their folklore tradition. In particular, the highest values (11.318 ± 5.46) observed among the Adi Minyong community can be attributed to their large population size and their location not far from the urban area, which possibly had facilitated external gene flow. In contrast, the smallest value (8.5228 ± 4.134) among Adi Komkar is in agreement with other results and is expected in view of their remote location, small population size. Possibly the Fst values of 1.333 obtained from AMOVA analysis are in agreement with the trends of diversity values as well. The AMOVA estimates also show near-zero negative values in case of Fct estimates in case of two and three groups. The negative values obtained can be expected in case the populations are small in size and practice high endogamy rate or inbreeding. Adi Komkar and Adi Panggin are two small populations, live in isolation in upper higher mountain ranges and practice high endogamy as has been revealed from their demographic and isonymy analysis (Maji et al. 2007, 2008); therefore, the near-negative values Fct observed is reflection of their small population size and high endogamy rates.

Overall, the results imply the role of several microevolutionary forces operating differently among the Adi regional populations in accordance with the varying population structure of the Adi—viz. marriage pattern, migration, admixture and due to recent demographic upheaval of as revealed by their folklore tradition.

In the phylogeny and the PCA plot, the two smaller groups, viz. Panggi and Pasi, indeed show greater affinity among themselves and also with one of the larger groups—the Minyong of Minyong–Padam cluster. Indeed, this supports the ethnohistorical accounts as narrated in their folklore tradition, especially that the two small groups of Adi Pasi and Panggi were the splinter groups formed in the past inter-tribal warfare during their formative period of settlement history. They were supposed to have been derived from Minyong of the Padam-Minyong group. It is of interesting that Adi Komkar is distantly located in the higher mountain valley miles and miles away from the location of the larger group Adi Padam on the opposite bank and at lower mountain heights. This possibly reflects their severe inimical and animosity relationship between the groups during their past historical inter-tribal warfares, and what separated them needs to be interrogated. This indeed supports the hypothesis based on the ethno-historical information of Adi sub-tribes: "*Adi sub-populations separated out from a single common larger group are expected to exhibit greater genetic affinity among themselves and with the single group from whom they were derived*".

The ethnographic and folklore information on Adi suggest that the sub-tribes are the factions as a result of inter-tribal warfares, wherein the formation of new sub-tribes from the splinter kin groups took place in the recent past (Roy 1997; Nath

2000; Lego 2005). This was the scenario especially with the formation of a separate settlement of Pasi-lower group from the parental Pasi group a few generations ago. Similarly, Panggi and Komkar are the splinter groups formed in the recent past as a result of conflicts over resource sharing (Roy 1997). These events have had influence in their genetic profiles, which can be inferred from their mismatch distributions.

The mismatch distribution based on mtDNA hypervariable sequences among the Adi regional populations show, in general, the unimode pattern with discrepancies indicating possible recent demographic expansion, in confirmation with their folklore tales depiction of inter-tribal warfare and redistribution and settlement of different factions in the recent past, a few generations ago. The effect of fission–fusion population structure had a differential effect among the Adi populations, in a way that some have had experienced severe demographic upheavals than others. Such effects can be detected from the mtDNA mismatch distributions, viz. especially the multimodal distribution observed among the Adi Padam is one such example. Adi Padam is relatively the largest of the all the Adi regional populations with a population size of about some thousands and well established. The multimodal with negatively skewed distribution indicating change in the demographic size can be explained due to migrations of large families in recent generations (after Indian Independence) from their original abode, located at higher mountain ranges at a place called "*Damuru*" to urban settlements, in the plains for jobs and other opportunities. Another relatively large population is Adi Minyong, which shows unimodal distribution with narrow peak suggesting that the population has stabilized in recent times, after the factional feuds, and continues to grow with little migration. Adi Panggi and Adi Komkar are two small populations and live in relative isolation mostly depended on hunting—gathering subsistence economy; however, since two decades some of them are getting settled near by developing urban areas. Adi Panggi and Adi Komkar show a tendency of multimodal and bimodal distributions with negative values observed in some of the estimates, viz. Fu's F and lower value of raggedness index "r" are in agreement with their small population size, and the same trend is also seen among other such populations (Cordaux et al. 2004).

The details of surname frequency among the Adi Panggi give some explanation for the positively skewed distribution with a tendency to form another mode. A similar trend can be inferred from the demographic indices of positive Tajima's D and Fu's F (Table 4). Since the population size is only about a few hundred and most of the marriages happen within the community along the clans and among the surnames. The results of the isonomy analysis reveal that the populations consist of 37 surnames shared among 154 husbands and 147 wives. About 70% of husbands and Panggi wives belong to seven surnames. Interestingly since it is patrilocal, the wives represent more diversity than their husbands. For example, in recent years there were wives from other Adi populations through marriages. There are at least 15 (40%) non-Panggi surnames admitted through marriages with 17 non-Panggi wives (Maji and Vasulu 2008). Possibly this explains the positive skewness and tendency of another mode in the distribution.

5 Conclusions

The results of the study on mitochondrial DNA investigated among the five Adi regional populations, to an extent, suggest differences in mtDNA haplotype frequencies, the relative genetic affinity and differential trends in the mismatch distribution and estimates of three demographic parameters. The results are in conformity and in agreement that is expected from the fission–fusion population structure and the possible genetic consequences (e.g., bottleneck effect of genetic drift especially among Adi Panggin) that these populations have experienced in the recent past.

References

Arunachal Pradesh Human Development Report 2005. (2006). Itanagar, Arunachal Pradesh: Department of Planning, Government of Arunachal Pradesh.

Arunkumar, G. P., Tatiana, V. T., Duty, J., et al. (2015). Genome-wide signature of male-mediated shaping the Indian gene pool. *Journal of Human Genetics* 1–7. https://doi.org/10.1038/jhg.2015.51.

Arunkumar, G. P., Soria-Hernanz, D. F., John Kavitha, V., Santhakumari Arun, V., Syama, A., & Samy Ashokan, K., et al. Lineages correlates with agricultural expansions predating the caste system. *PLoS ONE 7*(11), e50269. https://doi.org/10.1371/journal.pone.0050269.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290,* 457–465.

Bandelt, H., Forster, P., & Rohl, A. (1999). Median joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution, 16,* 37–48.

Basu, A., Sarkar-Roy, N., & Majumder, P. P. (2016). Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences of the USA, 113*(6), 1594–1599.

Bhasin, M. K., & Walter, H. (2001). *Genetics of castes and tribes of India*. Delhi: Kamla-Raj Enterprises.

Blackburn, S. (2003/2004) Memories of migration: Notes on legends and beads in Arunachal Pradesh, India. *European Bulletin of Himalayan Research, 25/26*, 15–60.

Blinkhorn, J., & Petraglia, M. D. (2017). Environment and cultural change in the indian subcontinent. *Current Anthropology, 58*(17).

Bradley D. (1997). Tibeto-Burman languages of the Himalayas. In Papers in Southeast Asia linguistics. No.14, Pacific Linguistis, A-86, pp. 1–72.

Chaubey, et al. (2015). The genome wide analysis of the Bhils. The second largest trial populations of India. *Man in India, 95* (4), 279–289.

Cordaux, R., Saha, N., Bently, G. R., Aunger, R., Sirajuddin, S. M., & Stoneking, M. (2003). Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. *European Journal of Human Genetics*, *11*(3),253–264.

Cordaux, R., Weiss, G., Saha, N., & Stoneking, M. (2004). The northeast Indian passageway: A barrier or corridor for human migration? *Molecular Biology and Evolution, 21,* 1525–1533.

Dani, A. H. (1960). *Prehistory and protohistory of Eastern India*. Calcutta: Firma KL Mukhopadhyay.

Das, B. M., Deka, R., & Das, R. (1980). Haemoglobin E in six populations of Assam. *Journal of the Indian Anthropological Society, 15,* 153–156.

Deka, R., Reddy, A. P., Mukherjee, B. N., Das, B. M., Banerjee, S., et al. (1988). Haemoglobin E distribution in ten endogamous population groups of Assam, India. *Human Heredity, 38,* 261–266.

van Driem, G. (2002). Tibeto-Burman phylogeny and prehistory: Languages, material culture and genes. In P. Bellwood & C. Renfrew (Eds.), *Examining the farming/language dispersal hypothesis* (pp. 233–249). Cambridge: McDonals Institute for Archealogical Research.

van Driem, G. (2004). 'Hodgson's Tibeto-Burman and Tibeto-Burman today'. In D. M. Waterhouse (Ed.), *The origins of himalayan studies: Brian houghton hodgson in Nepal and Darjeeling 1820–1858* (pp. 227–248). London: Routledge Curzon.

van Driem, G. (2006). The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese. *Bulletin of Chinese Linguistics, 1*(2), 211–270.

van Driem, G. (2007a). Non-human genetics, agricultural origins and historical linguistics in Asia. In M. D. Petraglia & B. Allchin (Eds.), The evolution of human populations in South Asia: Interdisciplinary studies in biological anthropology, linguistics and genetics (pp. 393–443). Dordrecht: Springer.

van Driem, G. (2007b). The diversity of the Tibeto-Burman language family and the linguistic ancestry of Chinese. *Bulletin of Chinese Linguistics, 1*(2), 211–270.

van Driem, G. (2013). East Asian ethnolinguistic phylogeography. *Bulletin of Chinese Linguistics, 7*(1), 135–188.

van Driem, G. (2015). 'Tibeto-Burman'. In W. S.-Y. Wang & C. Sun (Eds.), *Oxford handbook of Chinese linguistics* (pp. 135–148). Oxford: Oxford University Press.

Elwin, V. (1959). *A philosophy for NEFA*. Shillong: North-East Frontier Agency.

Fix Alan, G. (1975a). Genetic micro differentiation in the Semai Senoi of Malaysia. *American Journal of Physical Anthropology*, *43*,47–55.

Fix Alan, G. (1975b). Fission-Fusion and linear effect: Aspects of the population structure of the Semai Senoi of Malaysia. *American Journal of Physical Anthropology*, *43*,295–302.

Fix Alan, G. (1978). The role of kin-structured migration in genetic micro differentiation. *Annals of Human Genetics*, *41*, 329–339.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics, 147,* 915–925.

Fuller, D. Q. (2006). Agricultural origins and frontiers in South Asia: A working synthesis. *J World Prehist, 20,* 1–86.

Fuller, D. Q. (2007). Non-human genetics, agricultural origins and historical linguistics in South Asia. In M. D. Petraglia & B. Allchin (Eds.), The evolution and history of human populations in South Asia (pp. 393–443). Dordrecht, The Netherlands: Springer.

Gordon, R. G. (Ed.). (2005). *Ethnologue: Languages of the World* (15th ed.). Dallas, Tex.: SIL International.

Hackinger, S., Kraaijenbrink, T., Xue, Y., Mezzavilla, M., Asan, van G. D., Jobling, M. A., Kniff, de P., Tyler-Smith, C., & Ajub, Q. (2016). Wide distribution and altitude correlation of an archaic high-altitude-adaptive *EPAS1* haploype in the Himalayas. *Human Genetics* https://doi.org/10.1007/s00439-016-1641-2.

Harney, E., Niraj, R., Nick, P., Kumarasamy, T., & David, R. (2017). The skeletons of Rookund Lake: Genomic insighits into the mysterious identity of ancient Himalayan Travellers. Human Evolution 2017 Conference http://eurogenes.blogspot.com.

Harpending, H. C., Sherry, S. T., Rogers, A. R., & Stoneking, M. (1993). The genetic structure of ancient human populations. *Current Anthropology, 34,* 483–496.

Kashyap, V. K., Chattopadhyay, P., Dutta, R., & Vasulu, T. S. (2004). Genetic structure and affinity among eight ethnic populations of eastern India: Based on 22 polymorphic DNA loci. *American Journal of Human Biology, 16,* 311–327.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., et al. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *American Journal of Human Genetics, 72,* 313–332.

Kong, Q. P., Bandelt, H. J., Sun, C., Yao, Y. G., Salas, A., Achilli, A., et al. (2006). Updating the East Asian mtDNA phylogeny: A prerequisite for the identification of pathogenic mutations. *Human Molecular Genetics, 15*(13), 2076–2086.

Kraaijenbrink, T., van der Gaag, K. J., Zuniga, S. B., Xue, Y., Carvakho-Silva, D. R., Tyler-Smith, C., et al. (2014). A linguistically informed autosomal STR Survey of human population residing in the greater Himalayan region. *PLoS ONE, 9*(3), e91534. https://doi.org/10.1371/journal.pone.0095134.

Krithika, S., & Vasulu, T. S. (2013). Effect of past demographic events on the mtDNA diversity among the Adi tribe of Arunachal Pradesh. In R. Dasgupta (Ed.), *Advances in growth curve models* (pp. 199–2014). Springer proceedings in Mathematics and Statistics 46. https://doi.org/10.1007/978-1/4614-6862-2_11. New York.

Krithika, S., Maji, S., & Vasulu, T. S. (2009). A microstellite study to disentangle the ambiguity of linguistic, geographic, ethnic and genetic influences on tribes of India to get a better clarity of the antiquity and peopling of South Asia. *American Journal of Physical Anthropology, 139,* 533–546. https://doi.org/10.1002/ajpa.21018.

Krithika, S., Suvendu, M., & Vasulu, T. S. (2013). Molecular bioilogical perspectives of tribes of India. *Indian Journal of Anthropological Society*, Special volume of conceptualising Tribes in India, *62*(2), 775–804.

Krithika, S., Trivedi, R., Kashyap, V. K., Bharati, P., & Vasulu, T. S. (2006a). Antiquity, geographic contiguity and genetic affinity among Tibeto-Burman populations of India: A microsatellite study. *Annals of Human Biology, 33*(1), 26–42.

Krithika, S., Maji, S., & Vasulu, T. S. (2006b). Genetic heterogeneity among three Adi tribes of Arunachal Pradesh, India. *Human Biology, 78*(2), 221–227.

Krithika, S., Maji, S., & Vasulu, T. S. (2007). Intertribal and temporal allele-frequency variation at the ABO locus among Tibeto-Burman-speaking Adi subtribes of Arunachal Pradesh, India. *Human Biology, 79*(3), 355–362.

Kumar, V., Basu, D., & Mohan Reddy, B. (2004a). Genetic heterogeneity in northeastern India: Reflection of Tribe-Caste continuum in the genetic structure. *Amrican Journal of Human Biology*, *16*, 334–345.

Kumar, S., Timura, K., & Nei, M. (2004b). MEGA3: Integrated software for molecular genetics analysis and sequence alignment. *Briefing in Bioinforamtics*, *5*(2), 150–163.

Kumar, S., Padmanabham, P. B., Ravuri, R. R., Uttaravalli, K., Koneru, P., Mukherjee, P. A., et al. (2008). The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. *BMC Evolutionary Biology, 8*(1), 230.

Kumar, V., Reddy, A. N., Babu, J. P., Rao, T. N., Langstieh, B. T., et al. (2007). Y-chromosome evidence suggests a common heritage of Austro-Asiatic populations. *Evolutionary Biology, 7*, 47.

Lego, N. (2005). *History of the Adis of Arunachal Pradesh*. Itanagar, Arunachal Pradesh: Peregrine Graphics.

Maji, S., Krithika, S., & Vasulu, T. S. (2007). Genetic kinship among an isolated Adi tribe of Arunachal Pradesh: isonymy in the Adi Panggi. *Human Biology, 79*(3), 321–337.

Maji, S., & Vasulu, T. S. (2008). Genetic structure of an isolated sub-tribe of the Adi people of Arunachal Pradesh state in Northeast India: Isonymy analysis and selective neutrality of surname distribution in Adi Panggi. *Journal of Genetic Genealogy, 4*(1), 1–11.

Majumdar, D. N. (1980). Northeast India: A profile. In T. C. Sharma & D. N. Majumdar (Eds.), *Eastern Himalayas: A study on anthropology and tribalism*. Cosmo: New Delhi.

Majumder, P. P. (2010). The human genetic history of South Asia. *Current Biology, 20,* R184–R187.

Mayukh, M., Anders, B., Yali, X., et al. (2017). Y-chromosomal sequences of diverse Indian populations and the ancestry of the Andamanese. *Human Genetics*. https://doi.org/10.1007/s00439-017-1800-0.

Nath, J. (2000). *Cultural heritage of tribal societies* (Vol. 1) (The Adis). New Delhi: Omsons Publications.

Neel, J. V. (1970). Lessons from a primitive people. *Science*, *170*, 815–822.

Neel, J. V. (1973). "Private" genetic variation and the frequency of mutation among South American Indians. *Proceedings of National Academy of Sciences (USA)*, *70*, 3311–3315.

Neel, J. V., & Salzano, F. M. (1967). Further studies on the Xavante Indians. X: Some hypotheses-generalizations resulting from these studies. *American Journal of Human Genetics*, *19*(4), 554–574.

Niraj, R., Gyneshwer, C., Rakesh, T., et al., (2012). The phylogeography of Y-chromosome Haplogroup H1a1a-M82 reveals the likely Indian origin of the European Romani Populations. *PLoS One* 11.e48477.

Rapson, E. J. (1955). People and languages. In E. J. Rapson (Ed.), *Cambridge history of India* (Vol. 1, pp. 33–57). S. Chand: Delhi.

Reich, D., et al. (2009). Reconstructing Indian population history. *Nature, 461,* 489–494.

Roy, S. (1997). *Aspects of Padam Minyong culture*. Itanagar, Arunachal Pradesh: The Director of Research.

Roychoudhury, A. K. (1981). The genetic composition of the people in Eastern India. *Journal of the Indian Anthropological Society, 16,* 153–170.

Roychoudhury, A. K. (1992). Genetic relationships of the populations in eastern India. *Annals of Human Biology, 19,* 489–501.

Ruhlen, M. (1991). A guide to the world's languages. In *Classification* (Vol. 1) Stanford, California: Stanford University Press.

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., et al. (2006). Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *American Journal of Human Genetics, 78,* 202–221.

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). Molecular cloning: A laboratory manual. In N. Ford, C. Nolan, & M. C. Ferguson (Eds.) Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Shah, A. M., et al. (2011). Indian Siddis of African descendants with Indian admixture. *American Journal of Human Genetics, 89,* 154–161.

Singh, K. S. (1998). *People of India: India's communities* (Vol. 1). New Delhi: Oxford University Press.

Tabi, T. (2006). *The Adis*. Pasighat, Arunachal Pradesh: Siang Literary Forum.

Tamang, R., & Thangaraj, K. (2012). Genomic view on the peopling of India. *Investigative Genetics, 3,* 20.

Thangaraj, K., Naidu, B. P., Crivellao, T., Tamang, R., Upadhyay, S., et al. (2010). The influence of natural barriers in shaping the genetics structure of Maharashtra Populations. *PLoS ONE, 5*(12), e15283. https://doi.org/10.1371/journal.pone.0015283.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Schozzari, R., Obinu, D., Savontaus, M. L., & Wallace, D. C. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics*, *144*, 1835–1850.

Torroni, A., Petrozzi, M., Urbano, L. D., Sellitto, D., Zaviani, M., Carrara, F., Carducci, C., Leuzzi, V., Careli, V., Barboni, P., De Negri, A., & Scozzari, R. (1997). Haplotype and phylogenetic analyses suggest that one European-specific mtDNA bacground plays a role in the expression of Leber hereditary optic neuropathy by increasing the penentance of the primary mutations 11778 and 14484. *American Journal of Human Genetics*, *60*, 1107–1121.

Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., Calderon, F. L., Simionali, B., Valle, G., Richards, M., Macaulay, V., & Scozzari, R. (2001). Do the four clandes of the mtDNA haplgroup L2 evolve at different rates? *American Journal of Human Genetics*, *69*, 1348–1356.

Torroni, A., Schurr, T. G., Yang, C-C., Szathmary, E. J. E., Williams, R. C., Schanfield, M. S., & Troup, G.A., et al. (1992). Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, *130*, 153–162.

Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., Smith, D. G., Vullo, C. M., & Wallace, D. C. (1993). Asian affinities and continental radiation of the four founding native American mtDNAs. *American Journal of Human Genetics*, *53*, 563–390.

Vishnupriya, K., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., et al. (2017). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science, 5,* 171504. https://doi.org/10.1098/rsos.171504.

Walter, H., Mukherjee, B. N., Gilbert, K., Lindenberg, P., Dannewitz, A., et al. (1986). Investigations on the variability of haptoglobin, transferrin and Gc polymorphisms in Assam, India. *Human Heredity, 36,* 388–396.

Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T., & Zhang, Y. P. (2002). Phylogeographic differentiation of the mitochondrial DNA in Han Chinese. *American Journal of Human Genetics, 70,* 635–651.

Snapshots from ISI, Giridih and of Some Research Initiatives Undertaken for this Volume

Picture 1 Garlanding the statue of Prof. P. C. Mahalanobis

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6

Picture 2 Workers and students of ISI, Giridih, along with the participants of the conference

Picture 3 Moving toward the conference hall

Picture 4 In the lecture hall after inauguration attended by media personnel

Picture 5 Elephant foot yam growth experiments conducted in the farm of ISI, Giridih

Picture 6 Usri Falls, Giridih, in winter

Picture 7 A tea garden in Tripura, within the tribal belt

Picture 8 Tribal persons are interviewed in Tripura

Picture 9 Toward the Sunderban farm by launch over the saline water river *Bidyadhari*

Picture 10 Coconut trees planted in the farm for growth experiments, near the river bank