Springer Proceedings in Mathematics & Statistics

Ratan Dasgupta *Editor*

Growth Curve and Structural Equation Modeling

Topics from the Indian Statistical Institute

Springer

Springer Proceedings in Mathematics & Statistics

Volume 132

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Ratan Dasgupta
Editor

Growth Curve and Structural Equation Modeling

Topics from the Indian Statistical Institute

Springer

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

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-17328-3 ISBN 978-3-319-17329-0 (eBook)
DOI 10.1007/978-3-319-17329-0

Library of Congress Control Number: 2015938346

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Growth Curve Models (GCM) are of immense help to study the movement of important characteristics over time in different subject matters including branches of natural science. This volume is an outcome of presentations made in a GCM workshop held at the Indian Statistical Institute, Giridih, during the period 18–19 February 2014. The book describes some of the recent trends of research in GCM on different subject areas in different disciplines, theoretical and applied.

We requested for contributions to the present volume, with more than one article if possible, from authors working in different areas associated with GCM; like in the case of compiling earlier volume 'Advances in Growth Curve Models'. We are thankful to the readers for well accepting the earlier volume. All the contributions are peer-reviewed.

A volume of this size and nature can only be successfully completed with constant support and encouragements from the contributing authors. My sincere thanks and appreciation are for all my fellow contributors. I also thank the reviewers, who in spite of their busy schedule could find time to assess the contributions at my request. Thanks are also due to *Springer* for their keen interest in the project and continuous encouragement from the very beginning.

The present endeavour about the research works on GCM that is going on by the scientists of Indian Statistical Institute in different branches of science will be considered successful if this can provide readers an insight on the broad area of research in GCM.

Kolkata, India — Ratan Dasgupta

Fig. 1 Garlanding the statue of Professor P.C. Mahalanobis, founder of the Indian Statistical Institute (ISI); before the workshop is inaugurated

Fig. 2 Some of the workshop participants in the year 2014

Fig. 3 Intercropping of Sisal and Elephant foot yam

Fig. 4 Tuber crop potatoes are being harvested from the experimental plots

Fig. 5 Rain water harvested in a water reservoir, besides protected forest like area in farmland

Fig. 6 Non-regular shaped yam produced in farm

Fig. 7 A budding flower from a yam of 10.8 kg weight, grown in the year 2014; soil of Giridih is conducive for yam plantation, see https://www.youtube.com/watch?v=w0NtZWuuVZ0. See also, https://www.youtube.com/watch?v=LboKuNcUp9E

Fig. 8 Honeycombs on Simul (Bombax) tree in Rosevilla campus of ISI Giridih. These consist of two layers of small regular shaped sturdy hexagonal structures. The small bees are sometimes friendly, see https://www.youtube.com/watch?v=tEFZqnZqcww

Fig. 9 Sisal in wild and dry environment amongst rocks, near *Ushri* falls

Fig. 10 Flooded *Ushri* in rainy season, flowing by the boundary of ISI Giridih Farm

Contents

Plant Sensitivity and Growth Curve Analysis of Elephant Foot Yam

Ratan Dasgupta

Abstract Longitudinal growths of Elephant-foot-yam are studied by taking yam from the ground, then measure underground growth and replant the structure. The growth curve has a spike and takes a sharp upturn towards the end, indicating that a small increase in plant lifetime at mature stage increases yield substantially, leading to a possibility of high production. A plant was seriously endangered accidentally, during the remaining short lifespan the growth slope of affected plant changed significantly to a higher level than its growth slope before intervention, compared to other plants. Two growth curves are compared by modelling the error component by a continuous Gaussian process viz., Ornstein–Uhlenbeck process. Growth curve corresponding to seed weight 650 g seems to indicate superior yield.

Keywords Amorphophallus paeoniifolius • Elephant foot yam • Growth curve • Longitudinal study • Cross sectional study • Archimedean principle • Ornstein–Uhlenbeck process

1 Introduction

Elephant foot yam (*Amorphophallus paeoniifolius*) may grow even in barren land and is a cash crop like a boon to farmers. In order to decide about the appropriate harvest time of yam cultivated in a land of lateritic soil full of gravels as in Giridih, Jharkhand (India), with extreme weather in summer (26–44 °C), in winter (1–20 °C), and having moderate (14 mm) to intense (70 mm) rainfall during May–October, we consider the problem of estimating the growth of underground yam deposition over time in a production season. In the above stated profile of cultivation scenario at Indian Statistical Institute (ISI) Giridih Farm, we consider plant lifetime and seed weight to be the independent variables in the present longitudinal growth study for yam via nonparametric regression. It turned out that about 6 months are required for yam to mature in the stated environment. Time component may play the dominant

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_1

role in different growth studies, e.g., see Diggle et al. (2003) for a number of examples. Growth curve model by considering skew-normal distribution for the error terms and multiplicative heteroscedasticity covariance structure is studied in Louzada et al. (2014).

In the present study we consider time and seed weight to be main variables for yam growth experiment conducted in Giridih, assuming other variables to have homogeneous effect on experimental units. The results obtained are comparable over different production seasons over years, provided environmental parameters, which regulate the physiological processes governing the matter partitioning to underground corm, are more or less similar. We adopt nonparametric analysis with general response function. Search for specific allometric relationships, and associated error due to model misspecification is thus circumvented.

In longitudinal study the same experimental unit is followed over time. Frequent checks on underground deposition by digging the yam out in order to take readings and then replanting the structure to grow further over time may disturb the experimental set-up. Uprooting even once, so as to take growth reading may adversely affect the plant, if the delicate yam root structure is not properly taken care of during replanting. Although a procedure that involves replanting is stressful for plants in general, in this particular experiment care was taken during digging out, then taking growth measurements and finally while replanting. The precaution taken is of excellent category except for stress induced on plant number 2, caused by two inexperienced field workers. The stress inflicted on this plant in the experiment is seen as a marked outlier in the data analysis. Presence of weak spike is observed in the growth curves in this longitudinal analysis.

Estimating interim weight of underground yam, without detaching the above-ground plant from the yam in a longitudinal study is another task in growth experiments. The problem may be circumvented in a cross-sectional study where data on the final yam yield versus plant life are analysed, and the growth curve is estimated by nonparametric regression. In earlier cross-sectional studies, sharp upward turn and presence of a spike towards the end of the estimated growth curve were observed; see Dasgupta (2013). Similar findings are observed in the present non-destructive longitudinal study. A spike in the growth curve leads us to infer that slight increase in the lifetime of the plant towards the end may increase yield considerably, although farmers may sometimes prefer early harvesting for monetary reasons.

In the present study, we estimate the yam growth curve in a longitudinal analysis over a production season. The experiment was undertaken at the ISI Giridih farm, with three fixed weight choices of cut-seed corm viz., 500, 650, and 800 g; and with two plantations for each chosen corm weight; so as to study within and between variations of growth curves over different seed weights.

Plant cells are known to be able to sense and respond to environmental stresses such as light, hormone, carbon dioxide, temperature, gravity, humidity, etc., see, e.g., Foyer et al. (1994) and Jia and Zhang (2008). Plant response to shock stress is of interest, see, e.g., Bose (1902), Wildon et al. (1992), and Zimmermann et al. (2009) on electrical nature of the conduction of stimuli. In the conducted experiment

at ISI Giridih farm, when a plant's delicate root structure is accidentally damaged and consequently further survival of the plant is seriously endangered, then within the remaining short lifespan the plant stored underground yam in a relatively faster rate compared to its undisturbed state. Test of significance in parametric and nonparametric set-up indicates that the slope change in yam storage in the affected plant is significantly higher than that for other healthy yam plants surviving longer in the conducted experiment. This may have potential in growth regulation.

Growth curve corresponding to seed weight 650 g seems to indicate superior yield. The curve is markedly smooth indicating steady growth that has a spike towards end. Error component in growth curves may be modelled by a continuous Gaussian process viz., Ornstein–Uhlenbeck process; the process parameters are estimated and interpreted in the present context. Growth curves with seed weight 650 g are seen to have less variation from error component. Technical details are given in Appendix.

The objectives of the present study are: (1) to have an insight of yam growth in a longitudinal study via parametric and nonparametric modelling; in view of the fact that growth patterns observed in cross-sectional studies (Dasgupta 2013) confirmed presence of a spike in yam growth curve towards end of plant lifetime, (2) to examine the plant sensitivity from inadvertently induced possible grievous hurt, in course of conducted experiment, involving uprooting and replanting in the middle production season, especially on growth; and (3) to recommend appropriate yam corm weight to farmers for field plantation in Giridih region located at 24.18°N, 86.3°E; at an average elevation of 289 m above sea level, in Jharkhand (India).

The paper is organized as follows. In Sect. 2 we study a general nonparametric model on yam growth and propose a Gaussian process, viz., Ornstein–Uhlenbeck process for error component. The section includes materials and methods along with the results obtained. Section 3 provides discussion and conclusion of the study. Some theoretical results are explained in the appendix.

2 Growth Curve: Longitudinal Study on Yam

First we present a description of materials and methods used in the experiment.

2.1 Materials and Methods

Elephant foot yam of "Bidhan Kusum" variety was planted in the experimental site of ISI, Giridih farm. Ready-to-sprout cut seed corms, much similar in shape to apple slices having a part of "main eye" on top, with weights 500, 650, and 800 g were made out of healthy yam. Two cut corms, with healthy skin on outer side, and each of abovementioned weight were planted in a row. Each pit was of 1 foot deep. Distance between plants in each row/column was 1 meter. Pieces were dipped in

thick cow dung slurry mixed with Mancozeb and dried under shade before planting. The experimental plot was treated with organic margosa cake powder before starting the experiment on 3 July 2012. Plants were administered a little bit of organic manure after 7/8 days from sprouting.

We examined the yam deposition by digging out six plants around the middle of a season on 17 October 2012. The plants were uprooted with much care (except for plant number 2 that suffered a damaged root structure); one at a time and then the volume of grown yam attached below the plants were estimated by submerging the soil-cleaned yam part in a water container and measuring the amount of displaced water. An estimate of yam weight was then available by multiplying the volume with density of yam.

To estimate density, another underground yam grown at that time was dug out and detached from the plant, its weight found, and volume of the yam was measured by Archimedean principle from the amount of displaced water when the yam was totally submerged; the yam density was found to be $\approx$4 g/cc.

All the plants were taken out of the ground on the same day for taking measurement during the experimental period. Although sown simultaneously, the date of sprouting is different for different plants. Thus plant lifetimes (counted from sprouting date) are different on the same calendar date and that resulted in different growths. Variation of growth may also be due to different seed weights.

While taking interim observations during the experiment, precaution has to be taken to minimise the time-span between a plant being uprooted and replanted, so as to minimise disturbance to the experimental set-up and damage to the root structure. Replanting is done immediately after taking the relevant measurements. The process is repeated sequentially, one plant at a time.

2.1.1 Data and Some Preliminary Look

Table 1 provides the six plant characteristics under the growth experiments: initial yam weight, weight during intermediate lifetime, final weight, etc.

Table 1 Growth data on yam

Plant no.	Seed wt (in gm)	Sprouting date	Yam weight on 17.10.12	Date till plant alive	Final weight	Plant life (day)
1	500	01.08.12	960.84	23.12.12	1,762	142
2	500	03.08.12	565.20	02.11.12	864	89
3	650	18.07.12	2,486.88	08.01.13	4,198	170
4	650	07.07.12	2,430.36	15.12.12	3,558	158
5	800	10.07.12	1,865.16	29.11.12	2,406	139
6	800	14.07.12	3,052.08	07.12.12	3,992	143

The table above provides the plant characteristics under the growth experiments: initial yam weight, weight during intermediate lifetime, final weight, etc.

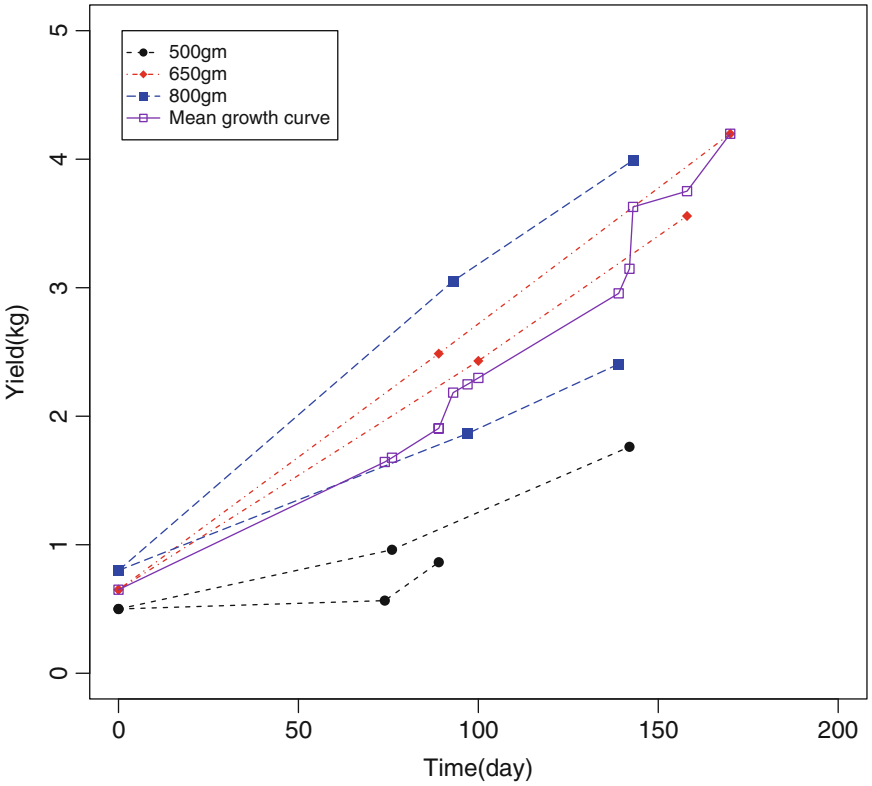


Fig. 1 Individual growth curves of yam. Growth curves corresponding to seed weight 500 g (*lower* two from start), 650 g (*middle* two from start) and 800 g (*top* two from start) are shown, with intermediate weights joined by *straight line*. Except for plant number 6 the slope differences in the two time segments are all positive, indicating presence of a weak upward spike in the growth curves, see Table 2; this difference is the highest for plant number 2. The curve in the *middle* joining 12 points represents overall growth i.e., the mean response $\mu(x)$ of Eq. (2) estimated as arithmetic mean of the weights (i.e., y values) over different plants, computed at a time point x where at least one recorded observation on y is available for some plant

In Fig. 1, individual growth curves of yam are shown. Hollow square shaped marks represent the mean of the y values over individual growth curves corresponding to 12 distinct values of plant lifetime (x values). Figure 2 shows the same with smoothing spline.

In this longitudinal analysis of underground yam growth data, presence of a weak upward spike is seen in individual growth curves. Growth slopes before and after the intervention, when yams are taken out of the ground, are estimated for each plant.

Two casual labours assigned to plant number 2 did not take proper care of the plant while uprooting, thereby damaging the root structure of the plant. The plant did not survive long afterward compared to other plants; see Table 1.

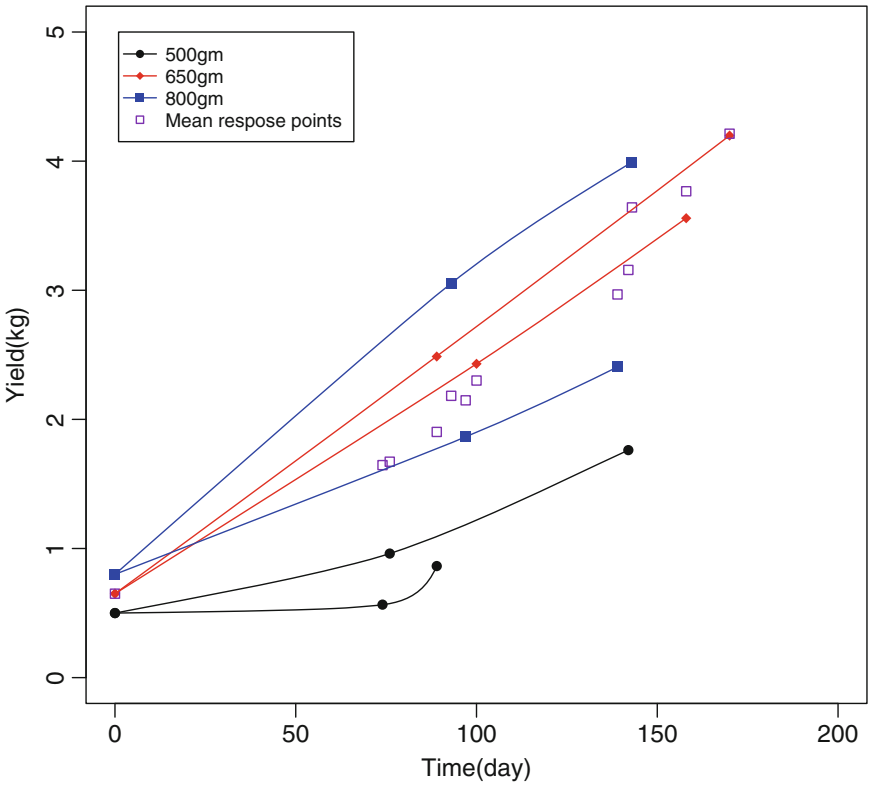


Fig. 2 Individual growth curves (spline) of yam. In this counterpart of Fig. 1, individual growth curves are obtained by spline smoothing in R software with shape parameter 0.5, instead of joining the points by *straight lines*. Hollow *square shaped* marks represent mean response $\mu(x)$ of Eq. (2) estimated as arithmetic mean of the weights (i.e., y values, now obtained by spline technique) over different plants computed at a time point x, where at least one recorded observation on y is available for some plant. The sharp upturn towards the end of the lowermost growth curve corresponding to plant number 2 is seen to be quite prominent

Growth slopes are computed in Table 2, these show homogeneity under normal plots, as seen in Figs. 3, 4, 5, and 6, except for plant number 2, after intervention of uprooting. These homogeneity justify combining individual growth data to obtain mean growth curve, as explained below.

2.1.2 The Model

Interim growth observations thus collected over the experimental period are a few in number. Such recorded observations serve as a natural nonparametric estimate of plant growth at that time point. For other time points we may assume growth to be linear in between two immediate upper and lower time points where observations

Table 2 Longitudinal slope in yam growth curve

Plant no. i	First slope	Second slope	Slope difference $u_i = \text{col}(3) - \text{col}(2)$
1	0.0060636842	0.01213879	0.0060751037
2	0.0008810811	0.01992000	0.0190389189
3	0.0206391011	0.02112494	0.0004858371
4	0.0178036000	0.01944207	0.0016384690
5	0.0109810309	0.01287714	0.0018961119
6	0.0242159140	0.01879840	−0.0054175140

The table above provides the nonparametric estimates of growth slopes as growth increments divided by the time taken for growth in two time segments. The difference between two slopes for each plant is also computed

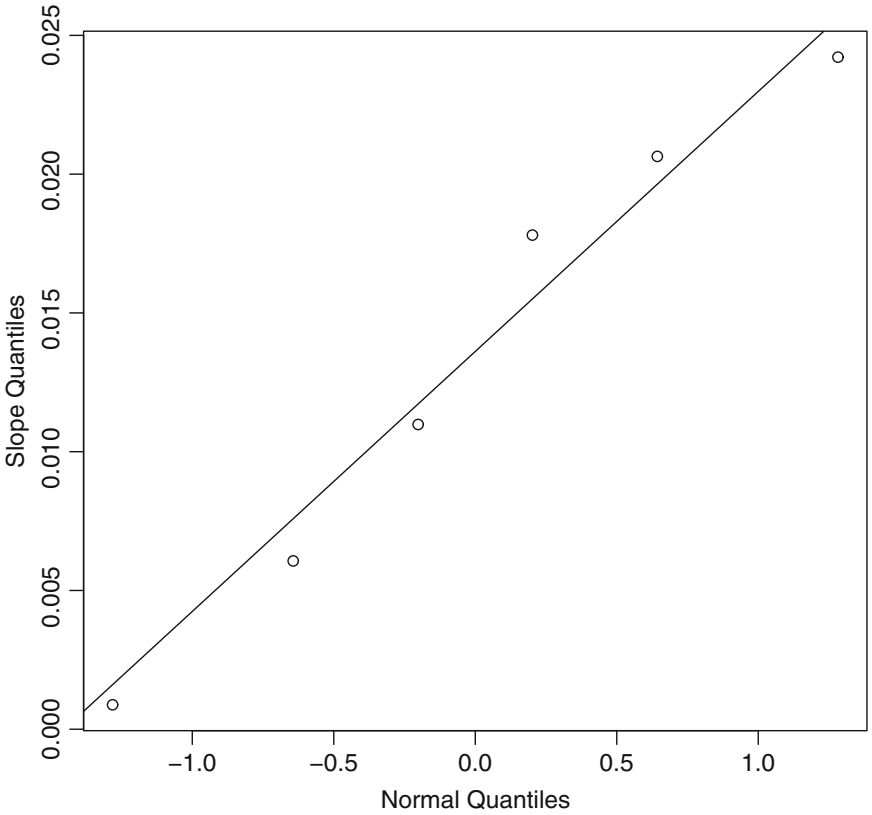


Fig. 3 Normal quantile plot for first slope. Quantile–quantile plot of first slopes of yam growth in the first time segment of plant lifetimes before intervention by uprooting the plants. The plot indicates a normal distribution as the points lie around a *straight line*

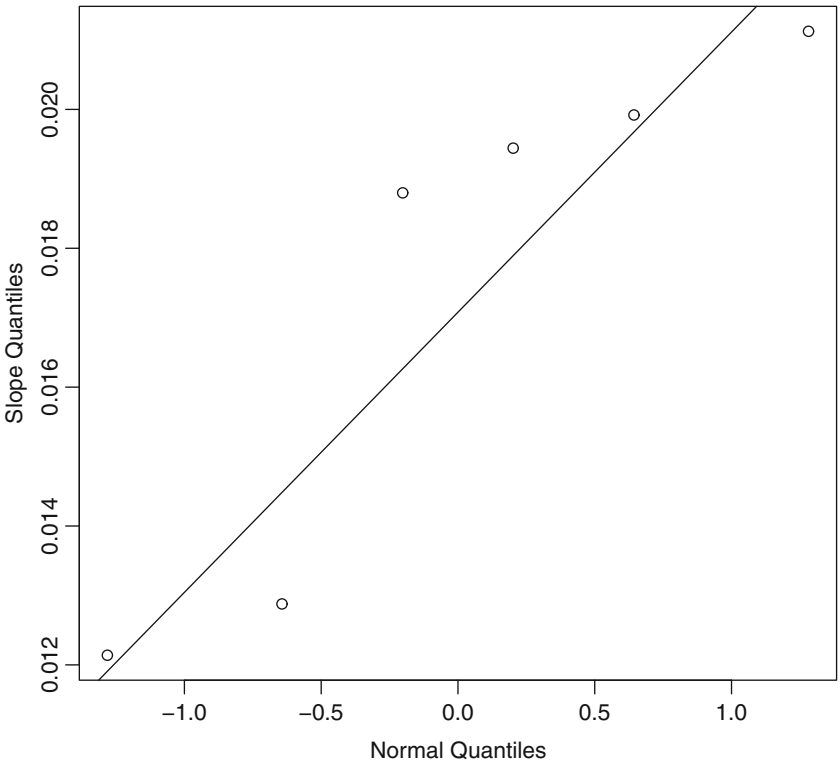


Fig. 4 Normal quantile plot for second slope. This is quantile–quantile plot of second slopes of yam growth in the second time segment of the plant lifetimes after intervention. Normality is somewhat doubtful

are recorded. Thus individual estimates of growth curves corresponding to different plants are obtained by linear interpolation in between observations, when the number of interim recorded observations for growth is small.

This accommodates a plant specific longitudinal slope β_L in growth models varying over different time regions, extending the linear growth model in the following manner.

$$E(Y_{ik}|x) = \beta_{c(i)} + \beta_{L(i1)}x_{i1} + \ldots + \beta_{L(ik)}(x_{ik} - x_{i\,k-1}) + \beta_{L(i\,k+1)}(x - x_{ik}) \quad (1)$$

where for the i-th plant at lifetime $x > x_{ik}(> 0)$, expected yam yield $E(Y_{ik}|x)$ is piecewise linear with longitudinal slopes β_L over time segments $(0, x_{i1}], \ldots, (x_{ik}, x_{i\,k+1}]$.

In the present case there are two time segments viz., before and after taking intermediate readings for $i = 1, \ldots, 6$. The first term, namely $\beta_{c(i)}$ may be considered as initial seed weight. Coefficients β_L represent expected change in Y over time per unit change in x, see Diggle et al. (2003). These plant and time specific

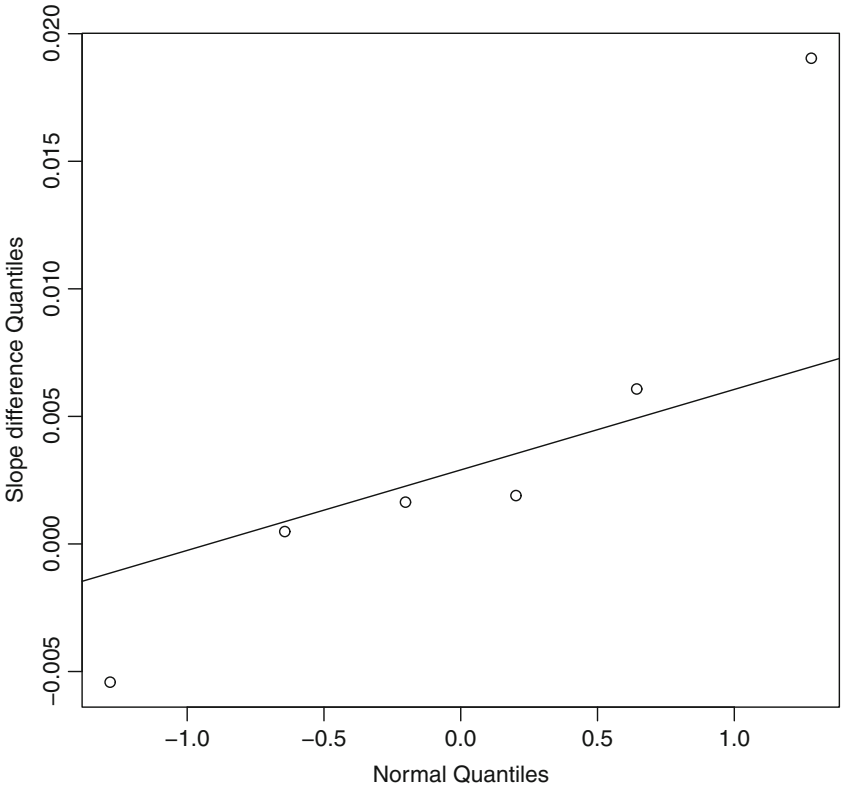


Fig. 5 Normal quantile plot for slope differences. This is quantile–quantile plot of the difference viz., second slope minus first slopes for individual plants. Topmost observation at *right corner* (corresponding to plant number 2) seems to be an outlier

longitudinal slopes $\beta_{L(ij)}$ are estimated in a nonparametric manner without assuming any parametric functional relationship amongst variables. There are two plantations per seed weight of 500, 650, and 800 g, and hence two growth curves are estimated from two plants for each of the fixed seed weight. One may thus compute variations between two growth curves for a fixed corm weight. Equation (1) is a special case of the following model.

$$E(Y_i|x) = \mu(x) + v_i(x) \tag{2}$$

where for each x, average of $v_i(x)$ over all the plants is 0.

With a simple nonparametric model postulated in Eqs. (1) and (2), we estimate the slope in growth curves for six plants (Table 2), and show that plant number 2 has markedly high growth rate after intervention, see Figs. 1 and 2. Statistical analysis reconfirms this visual finding.

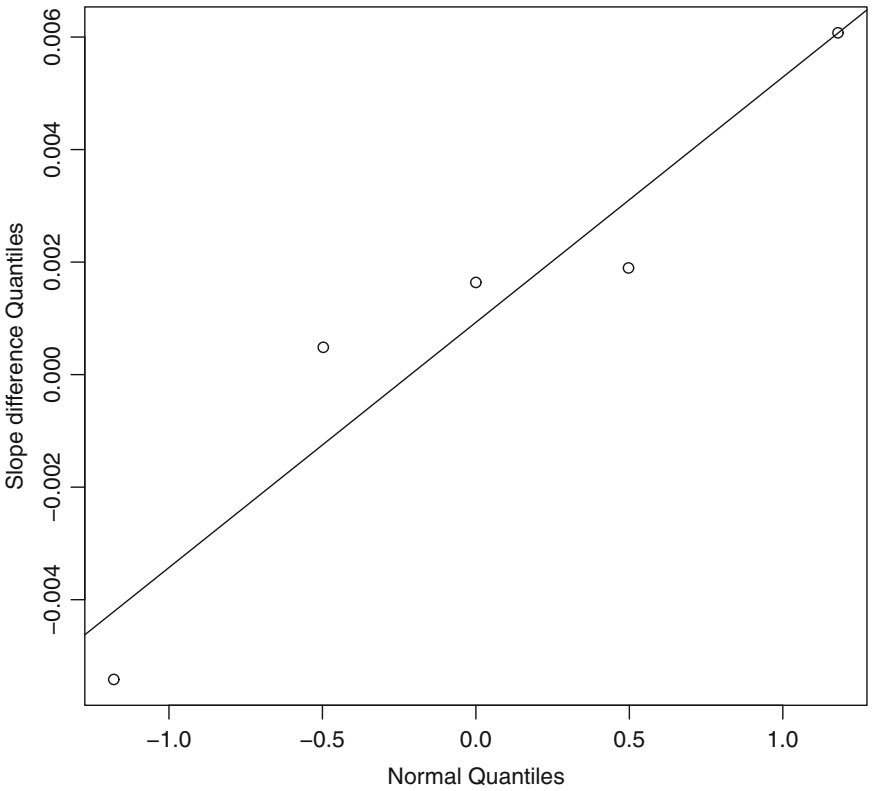


Fig. 6 Normal quantile plot for slope differences deleting extreme observation of plant 2. Deleting the extreme observation in Fig. 5, normality seems to hold for slope difference in this quantile–quantile plot

Estimate of the common growth or overall mean response $\mu(x)$ is taken as the arithmetic mean of the predicted weights (i.e., y values) over different plants computed at a time point x, where at least one recorded observation is available for some plant. One may take the average of the y quantiles for a fixed x value to obtain pooled values of (x, y) for plotting. Finally the overall response curve is estimated by nonparametric spline smoothing or lowess regression; see Cleveland (1979).

Lowess, a local polynomial regression estimator with smooth tricubic kernel and variable bandwidth based on k-th nearest neighbour, employs weighted least square criterion that assigns less weights to distant observations, to have a robust estimate of response curve insensitive to large-residual outliers, by down-weighting these over several iterations. However, lowess does not provide an explicit functional form of response variable with predictor variables.

A broad idea about the growth of yam over plant lifetime is explained via these techniques.

The average curve reflects the overall pattern of growth representing the mean component in curves. The individual curve is the mean component $\mu(x)$ plus a part $v_i(x)$ specific to i-th experimental unit. Thus an estimate of plant specific component $v_i(x)$ is the difference between individual curves from the average curve. In Eq. (1) the slope β_L may be taken as simple nonparametric estimates viz. increment of y per increment of x, as seen in individual growth curves obtained by joining the points (x, y). The procedure circumvents specification of allometric relationship, thereby avoiding error due to (allometric) model misspecification. With longitudinal data the first graph is the scatter plot of the response variable over time. See Fig. 1 for growth curves corresponding to grown plants arising out of field experiment with different seed weights.

Regression analysis estimates the conditional expectation of the dependent variable given the independent variables, i.e., the average value of the dependent variable when the independent variables are fixed. Linear increments of growth in time intervals are one of the simplest assumptions. However, other growth pattern like smoothing spline is also undertaken. Apart from nonparametric regression techniques of lowess and smoothing spline to obtain mean response or, growth curve, we may also adopt parametric Gaussian Ornstein–Uhlenbeck (O–U) process to model the error component and obtain a measure of variability of observed growth (of two plants, for each seed weight) from the corresponding mean response curve. The residuals, distance of the curve from mean response curve, will fluctuate more when variability in curves is high. Actual growth Y may then be modelled by the mean part of (1) plus a zero mean random residual component. Like many error components, these are usually assumed to follow (correlated) normal distribution, as in continuous time Ornstein–Uhlenbeck (O–U) process, see Uhlenbeck and Ornstein (1930); since successive observations over time on an experimental unit may be correlated.

Apart from some pathological examples, Ornstein–Uhlenbeck process is the only continuous normal process that is strongly Markov (i.e., it depends on past values only through the immediate past), strictly stationary (i.e., any finite dimensional distribution of it is invariant under time shift). This satisfies the following differential equation.

$$dV(x) = -\alpha V(x)dx + \gamma dB(x),\ \alpha > 0,\ \gamma > 0 \tag{3}$$

where $B(x)$ is the standard Brownian motion, γ is the diffusion parameter, α is the drift parameter; $\alpha V(x)$ is a restoring force directed towards origin proportional to the distance $V(x)$; see Karlin and Taylor (1981). In Dasgupta (2006), trimmed edge curve of thin waste metal sheets to give these a regular shape were modelled by O–U process. Distribution of an industrial characteristic burr is modelled by O–U process in Dasgupta (2011). Parameters of the process are estimated from realisation of curves therein. Let us see how the assumptions for O–U process work for error component in yam growth.

While estimating the mean response curve via lowess regression in Dasgupta (2013), position of the curves with respect to data points in different figures indicates

symmetric error component. The same is observed in the present study as well. Markov property is apparent from the fact that additional growth on a day is added with earlier growth status of the day before. Continuity of the process is assumed as growth is continuous over time. The distribution of measurement errors being invariant over time contribute towards stationary. Different normal plots in the present and earlier studies on yam indicate distribution of growth readings to be normal.

For a class of generalised Ornstein–Uhlenbeck processes, which has particular application in financial modelling, see e.g., Maller et al. (2009).

With a continuous response $\mu(x) + \nu_i(x)$, the growth curve

$$Y_i(x) = \mu(x) + \nu_i(x) + V(x) \tag{4}$$

is continuous almost surely for x. This curve may mimic the actual growth when an estimate of $\mu(x) + \nu_i(x)$ is available. Since there are two realised growth curves for each seed weight, average curve of these two may be taken as nonparametric estimate of the response $\mu(x) + \nu_i(x)$, for $i = 1, 2$ for seed weight 500 g; $i = 3, 4$ for seed weight 650 g; and $i = 5, 6$ for seed weight 800 g; with the understanding that $\nu_i(x)$ are same for i in a group, see Fig. 11 for three estimated response curves $\mu(x) + \nu_i(x)$, for three seed weights 500, 650 and 800 g. Data from Plant number 2 after intervention is not taken into account to calculate mean response curve for seed weight 500 g, as it turned out that the plant was damaged after intervention of uprooting.

Estimation of overall response curve is of primary interest in regression analysis. In the general nonparametric regression, the response function is left unspecified. Most methods of nonparametric regression implicitly assume that response is a continuous function.

Instead of linear interpolation used in the first stage to estimate intermediate points of individual growth, we also considered cubic spline for smoothing individual growths.

2.2 Results

In Fig. 1, individual growth curves corresponding to six different plants are shown, where the yam weights are joined by straight lines. Arithmetic mean of the weights (i.e., y values) over different plants is computed at a time point x where at least one recorded observation on y is available for some plants. There are 12 such distinct x points of yam lifetimes in the present case. The corresponding (x, y) points marked by hollow squares are joined by lines.

One may use spline technique to estimate the individual growth curves as shown in Fig. 2, with shape parameter 0.5, by R software. This procedure seems more accurate than linear interpolation between successive points (Fig. 1). Hollow square

shaped marks in Fig. 2 represent the mean of the y values over individual smoothed growth curves corresponding to 12 distinct values of plant lifetime (x values).

The curve corresponding to plant number 2 refers to a short lived plant. After an intermediate observation was taken its life was cut short; see lowermost curves in Figs. 1 and 2. There is a sharp upward turn in the corresponding growth curve towards the end for plant number 2. Root structure for plant number 2 was considerably damaged while digging it out, as the concerned field workers were not experienced enough. Also, outside exposure for a long period after digging out, while taking observation on 17.10.2012, may have caused the subsequent short life span for the injured plant number 2 after replanting. Within the remaining short lifespan, the plant deposited carbohydrate in a faster manner. In Table 2, for each plant we compute the growth slopes for two time segments, before and after the yams were taken from the ground for recording interim growth. For a particular plant i, growth slopes $\beta_{L(i1)}$ and $\beta_{L(i2)}$ are obtained by dividing the growth increments by the time length; see Eq. (1).

The second and third columns of Table 2 provide the estimate of the two growth slopes for the six plants. The fourth column provides the difference of the second slope from the first slope. The second yam plant with damaged root structure while taken off the ground corresponds to the highest difference, indicating a sharp change of speed in food storage. Except for plant number 6 the slope differences are all positive, indicating presence of a weak upward spike in growth curves. Higher rate of yam deposition underground, towards end of lifetime is also seen to hold in cross-sectional analysis on yam growth made earlier in Dasgupta (2013). It seems plausible that a yam plant may sense when its lifespan is going to be over, which prompts the plant to finish the remaining tasks faster when the end is approaching.

We now proceed to test whether the observed phenomenon of shift to a faster growth rate in plant number 2 under duress of damaged root structure is statistically significant compared to other healthy yam plants.

Observations related to different plants may be considered independent. From Eq. (1), estimates of slope (and slope difference) do not involve initial seed weight $\beta_{c(i)}$. Out of six independent observations, the second entry in column (4) is the largest. Thus, considering the cases to be equally likely, the event that the highest slope difference occurs for plant number 2, and the plants are of homogeneous pattern, has probability 1/6. That is, the second plant incidentally accumulated more food after root damage and the plants behaviour are indeed similar has a (low) probability 1/6.

Estimate of slopes and their differences are linear combination of growth observations Y_{ik}; these are normally distributed when the error components implicit in Eq. (1) are normal. Figures 3 and 4 show normal quantile plots for the first and second slopes of yam growth given in Table 2 (column no. 2 and 3, respectively). Normal distribution apparent in the first slopes seems to have undergone a little bit of change in the second slopes, those observed after replanting subsequent to taking the interim readings.

The normal quantile plot for the difference of slopes is shown in Fig. 5. The extreme point on the right top corner represents reading for plant number 2.

Deleting the highest slope difference corresponding to plant number 2 from the set of points, the modified quantile plot is shown in Fig. 6; the coefficient of determination $r^2 = 0.9066$ from linear regression. This seems to suggest that growth-slope differences in remaining five plants may be considered as normally distributed, i.e., $u_i \sim N(\mu, \sigma^2), i \in \{1, 3, 4, 5, 6\}$.

Mean and variance based on the remaining five entries in last column are $\hat{\mu} = \sum_{i\in\{1,3,4,5,6\}} u_i/5 = 0.0009356015$ and $\hat{\sigma^2} = \sum_{i\in\{1,3,4,5,6\}}(u_i - \hat{\mu})^2/5 = 1.709886 \times 10^{-5}$, respectively. With these maximum likelihood estimates (m.l.e.) of mean and variance, the estimated density function of growth-slope differences u may be written as

$$f(u) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(u-\hat{\mu})^2}{2\hat{\sigma^2}}} \tag{5}$$

The entry $u_2 = 0.0190389189$ corresponding to plant number 2 seems to be an outlier under normal set-up (5) as, $\tau = (0.0190389189 - \hat{\mu})/\hat{\sigma} = 4.377988$, with level of significance $p = 5.989001 \times 10^{-6}$, indicating that it is highly improbable that the high change in rate of food storage, as observed in plant 2 in contrast to other plants may be attributed to chance.

The conclusion of the above approximate normal test may be validated by an exact test with further calculations. Consider sum of squares due to errors, SSE $= \sum_{i\in\{1,3,4,5,6\}}(u_i - \hat{\mu})^2 = 5\hat{\sigma^2}$, where $\hat{\sigma^2}$ is m.l.e. of σ^2 under normal set-up. Then SSE/σ^2 has a Chi-square distribution with 4 degrees of freedom (d.f.), and this is independently distributed of $(u_2 - \hat{\mu}) \sim N(0, 6\sigma^2/5)$, under the hypothesis that $\{u_i, 1 \le i \le 6\}$ are independent and identically distributed $N(\mu, \sigma^2)$ random variables. Then it can be shown that the test statistic $t = 2\tau/\sqrt{6}$ has an exact distribution as that of a t-statistic with 4 d.f.; and the conclusion that growth-slope difference u_2 corresponding to plant number 2 is an outlier holds, in view of large value of observed $t = 3.5746121$. Level of significance in exact test is $p = 0.01164$.

Computed growth slopes over different plants (shown in Figs. 3, 4, 5, and 6) indicate homogeneity under normal plot, as mentioned before. This justifies combining individual growth data, except for plant number 2 (after intervention).

To obtain the mean response curve in Fig. 1, arithmetic mean of the weights (i.e., y values) over different plants was computed at a time point x where at least one recorded observation on y is available for some plants. There are 12 such distinct x points of yam lifetimes in the present case. The corresponding (x, y) points marked by hollow squares are joined by lines in Fig. 1. This provides an estimate of the yam growth curve. To examine for possible presence of a spike in the growth curve, points are smoothed by lowess and spline techniques; these are shown in Figs. 7 and 8, respectively. Possibility of a spike towards the end of the growth curve is seen to be more prominent in Fig. 8.

A similar procedure may be followed for growth curves of Fig. 2. The hollow square shaped marks in Fig. 2 representing average points are smoothed by lowess

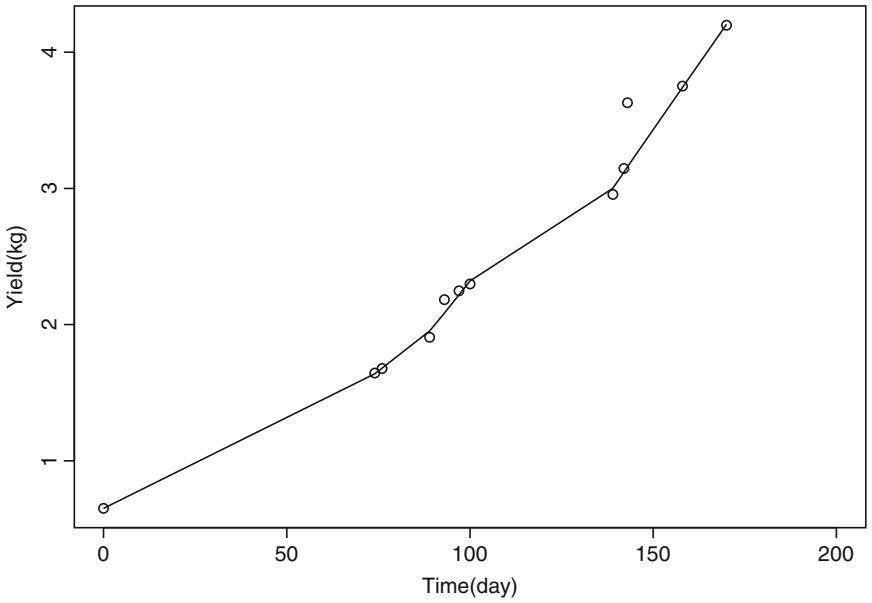


Fig. 7 Growth curve (lowess) of yam yield. Individual points of overall growth curve (see the curve joining 12 points in Fig. 1) represented as *circles* are smoothed by lowess technique with $f = .51$ in SPLUS software, and then joined by *lines* to show a smooth continuous broken line as an estimate of $\mu(x)$ in Eq. (2)

and spline techniques in Figs. 9 and 10, respectively. As before, these curves indicate possibility of having a spike towards the end of the yam growth curve.

For O–U process of Eq. (3), α is a drift parameter (i.e., pull of the process towards the mean value); and γ is the diffusion parameter, as it relates to the spread of the process. Note that $\sigma_v = \gamma/(2\alpha)^{1/2}$ may be interpreted as the standard deviation of the limiting distribution of $V(s)$, $s \to \infty$.

The growth curve lying on the top in Fig. 1 with seed weight 800 g has $t = 143$ days. For the higher curve with seed weight 650 g, $t = 170$. Mean response curves are based on both upper and lower curves for a specific seed weight. Deviations of upper curve from mean response is taken as $\hat{V}(s)$. The asymptotic distribution of the maximum likelihood estimate (m.l.e.) $\hat{\alpha}$ is normal see, e.g., Brown and Hewitt (1975). By Eqs. (8) and (9) of Appendix, the estimated process parameters from realised curves $\hat{V}$ are as follows (Figs. 12 and 13).

For seed weight 800 g; $\hat{\alpha} = 0.002144056$, $\hat{\gamma^2} = 0.001029093$, $\hat{\sigma}_v = 0.4898851$.

For seed weight 650 g; $\hat{\alpha} = 0.0006754383, \hat{\gamma^2} = 1.977903 \times 10^{-5}, \hat{\sigma}_v = 0.1210026$.

Small values of γ and large values of α signify that the random variation of growth curve from mean response is likely to be small in magnitude.

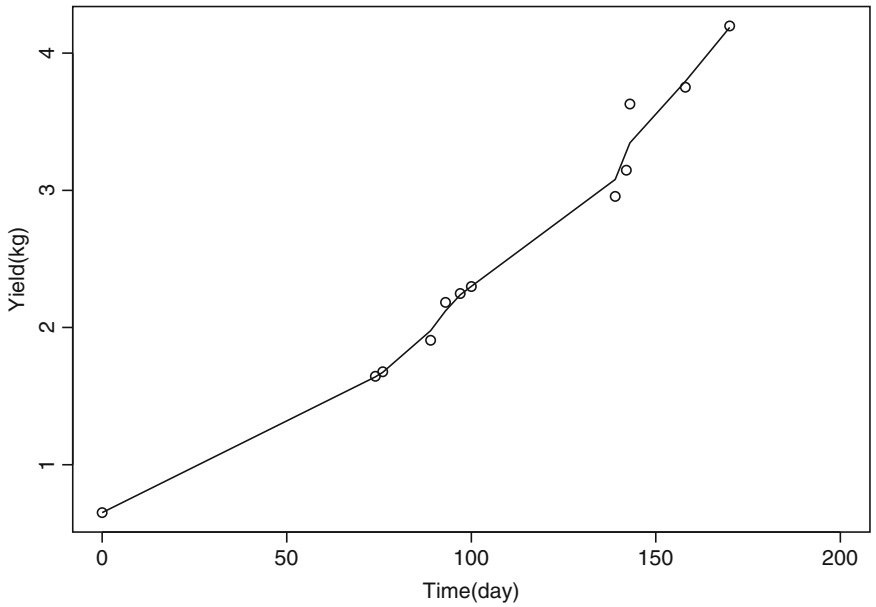


Fig. 8 Growth curve (spline) of yam yield. Individual points (12 points in Fig. 1) of overall growth curve represented as *circles* are smoothed by spline technique (using smooth.spline, spar=.00001 in SPLUS software) for a smooth continuous curve as estimated $\mu(x)$

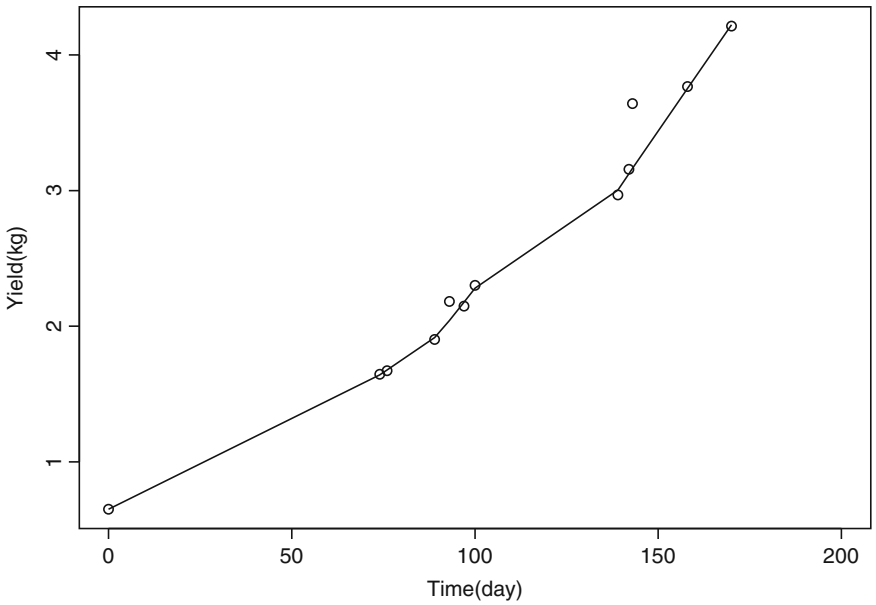


Fig. 9 Growth curve (lowess) of yam yield. In this counterpart figure of Fig. 7, hollow *square shaped* points of Fig. 2 are smoothed by lowess technique with $f = .51$ in SPLUS software, and then joined by *lines* to obtain a smooth continuous broken line as an estimate of $\mu(x)$ in Eq. (2)

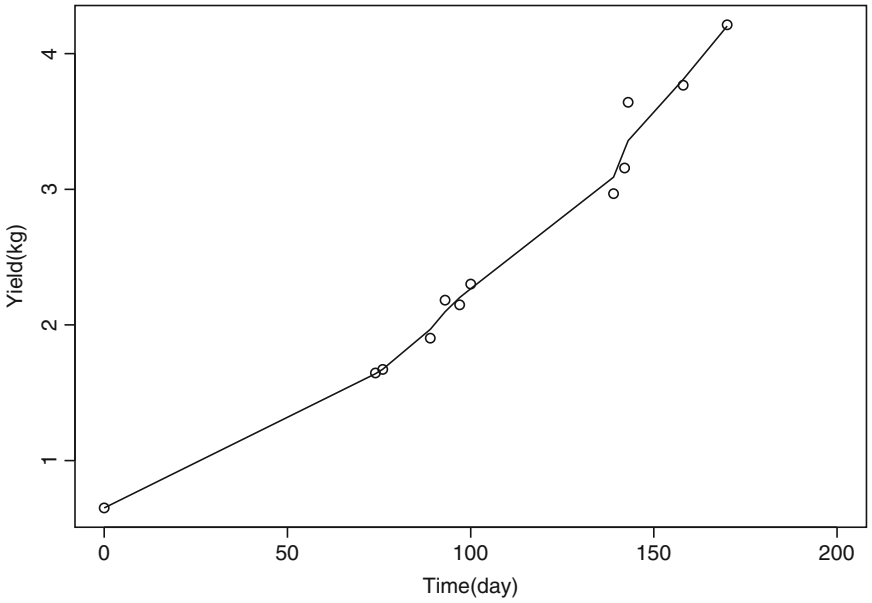


Fig. 10 Growth curve (spline) of yam yield. Hollow *square shaped points* of Fig. 2 are smoothed by spline technique (using smooth.spline, spar=.00001 in SPLUS software) to obtain a smooth continuous line as an estimate of $\mu(x)$. This is similar to Fig. 9. Presence of an upward spike towards the end is seen in these figures for overall growth

Asymptotic standard deviation (s.d.) of the process corresponding to seed weight 650 g is smaller than that for 800 g. Ratio of the estimated values of s.d. of residual process corresponding to 800 g, to that for 650 g is approximately 4. The values of α and γ are smaller for seed weight 650 g, compared to those for 800 g.

Yet another estimate $\tilde{\sigma}_v = \max_{0\le s\le t} \mid \hat{V}(s) \mid /\sqrt{2\log t}$ of the asymptotic standard deviation is available from the maximum fluctuation of observed process $\hat{V}(s)$ on time segment $[0, t]$, around mean response curve, see Appendix; Eqs. (12) and (13). This spread index of residual process is nonparametric in nature without model assumptions, as these are based on maximum fluctuation. See Figs. 12 and 13.

These are actually absolute values of the residuals. We only need maximum fluctuation of residuals for comparison. Central line estimated is equidistant from two growth curves for each seed weight, thus there is a mirror reflection of Figs. 12 and 13 below the line $y = 0$ for residuals computed from second curve for a fixed seed weight. These two curves (with fixed seed weight) have maximum fluctuation of same magnitude. O–U process is mean reverting, that is the process is pulled back to the (zero) mean. Since intermediate growths are linearly interpolated, the residual pattern is so. This pattern will be different if nonlinear spline regression is used. Here we are interested only in maximum fluctuations of process to compare two processes.

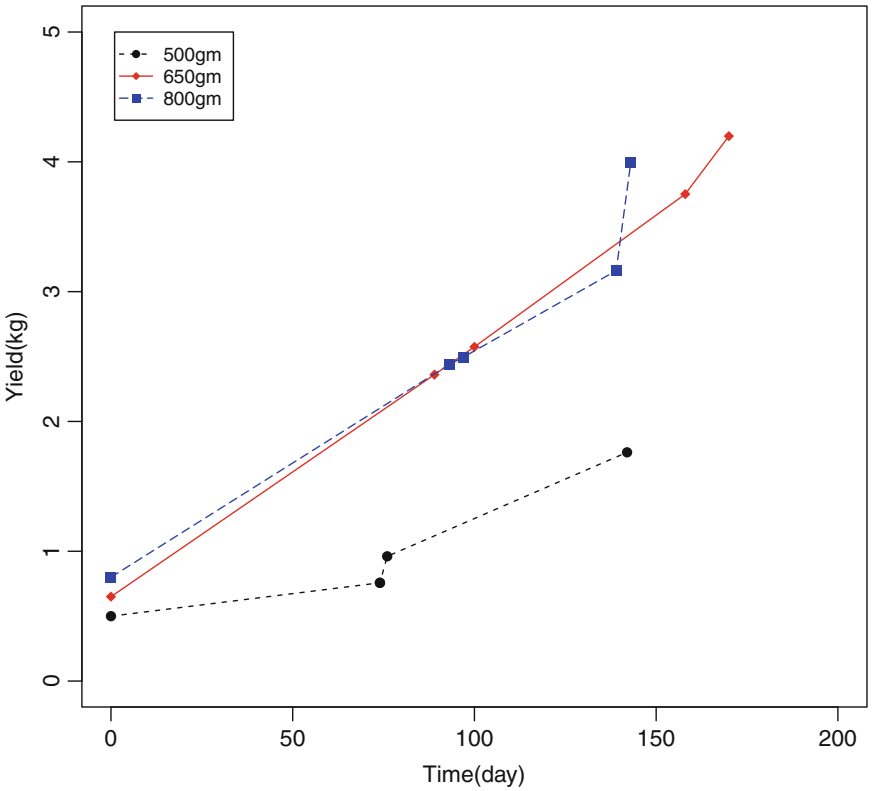


Fig. 11 Three mean response curves. Three estimated response curves $\mu(x) + v_i(x)$ for three seed weights 500, 650 and 800 g are shown. Curve corresponding to seed weight 500 g lies *below* the other two. Curve corresponding to seed weight 650 g seems to be steady and seems superior to curve corresponding to 800 g after about 100 days. Data from Plant number 2 after intervention is not taken into account to calculate mean response curve for seed weight 500 g, as it turned out that root structure of the plant was damaged after intervention of uprooting. Although in this figure the points in a curve are not smoothed by lowess and spline techniques, the curve corresponding to seed weight 650 g is markedly smooth indicating a steady growth that has a spike towards end

For seed weight 800 g, $\tilde{\sigma}_v = 0.239772$; and for seed weight 650 g, $\tilde{\sigma}_v = 0.06029778$. The ratio of $\tilde{\sigma}_v$ for residual process corresponding to seed weight 800 g, with that for 650 g is almost same viz., 4; as before.

The growth curve for seed weight 650 g seems to have less fluctuation compared to that for higher seed weight of 800 g. Estimated overall growth curve corresponding to seed weight 650 g indicates a superior yield than seed weight 800 g.

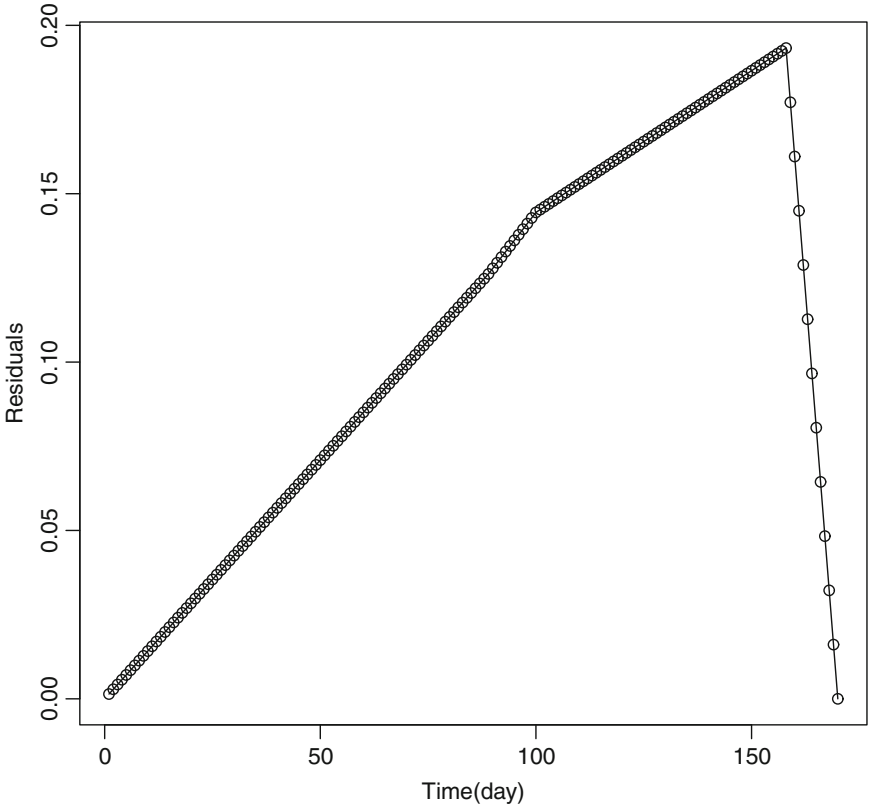


Fig. 12 Residuals in O–U model (seed wt 650 g). The deviations of the individual growth curve from mean response curve are shown for seed weight 650 g. Maximum fluctuation is of order 0.2

3 Discussion and Conclusions

Presence of a spike in yam growth curve was noted in cross-sectional studies (Dasgupta 2013). In this longitudinal study of Elephant foot yam growth, we observe that the growth curve has a spike and the curve takes a sharp upturn towards the end, indicating that a small increase in plant lifetime at mature stage increases yield substantially, leading to a possibility of high production. Farmers usually do not wait till the end for yam plant to die in its own course due to monetary reasons, as an early harvest fetches good price in local markets. A little more wait may produce good yield. A plant was seriously endangered accidentally in course of conducted experiment. During the remaining short lifespan the growth slope of affected plant changed significantly to a higher level than its growth slope before intervention (of uprooting and replanting), compared to other plants. In view of superior features observed in the yam growth curve corresponding to the seed weight 650 g, this

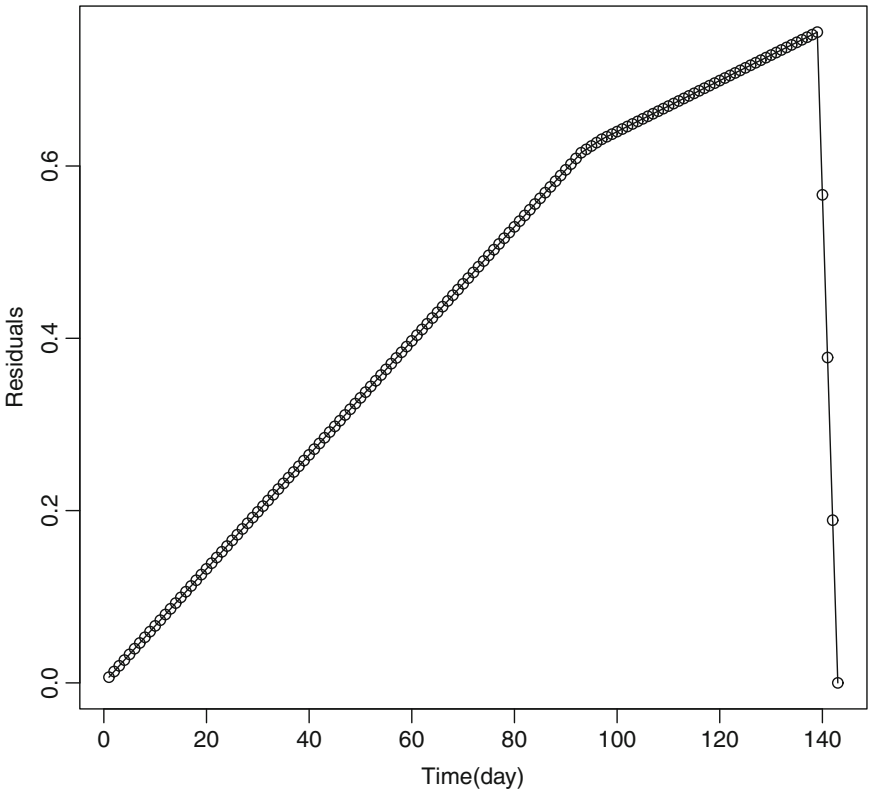


Fig. 13 Residuals in O–U model (seed wt 800 g). The deviations of the individual growth curve from mean response curve are shown for seed weight 800 g. Maximum fluctuation is of order 0.8, approximately four times than that for seed weight 650 g

seed weight may be recommended for plantation by Elephant foot yam farmers in Giridih, Jharkhand.

A follow-up study is conducted on longitudinal growth of yam with 60 plants, as reported in Dasgupta (2015). This reconfirmed a number of results stated in the present work including normality of growth slopes and slope differences. Seed weight 650 g turns out to be recommended choice again in the follow-up study. Even in small sample size, data may contain a lot of information, as the present study explains.

Acknowledgements Thanks are due to the referees for constructive comments that improved the presentation.

Appendix: Parameter Estimation in O–U Process

In this section we provide some technical details regarding parameter estimation of O–U process used in the main part.

Parameters of O–U processes in the proposed model (3)–(4) may be estimated from the observed growth curves, and interpreted in the present context. By transforming the process $V(s)$ to the corresponding Brownian motion $B(s)$ one may write, according to a result given in Lemma 4.2, page 212 of Basawa and Rao (1980), the following:

$$\lim_{n\to\infty} \frac{1}{t}\sum_{j=1}^{2^n}[V(jt2^{-n}) - V((j-1)t2^{-n})]^2 = \gamma^2 \text{ almost surely (a.s.)} \tag{6}$$

One may consider grids of finer length and then select that grid size for which the estimate of γ^2 as evident from (6) stabilizes, see also Dasgupta (2006). In the present case we may consider grid spacing as consecutive days. For each seed-weight the deviations of realised growth curve (Fig. 1) from the corresponding mean response curve (Fig. 11), computed at consecutive days, may be considered as observed values of residual $V = Y_i - (\mu + \nu_i)$ of model (4). Here t is plant lifetime.

Following the example 5.4, page 187–188 of Basawa and Rao (1980), the m.l.e. of α is the following:

$$\hat{\alpha} = -\int_0^t V(s)dV(s) / \int_0^t V^2(s)ds = \frac{1}{2}[\int_0^t V^2(s)ds]^{-1}[\gamma^2 t + V^2(0) - V^2(t)] \tag{7}$$

Estimate of γ^2 is given by

$$\hat{\gamma^2} = \frac{1}{t}\sum_{j=1}^{t}[V(j) - V(j-1)]^2 \tag{8}$$

Replacing the integral by finite sum, one may also write from (7) and (8)

$$\hat{\alpha} \approx \frac{1}{2}[\sum_{j=1}^{t}\{V(j)\}^2]^{-1}[\hat{\gamma^2} t + V^2(0) - V^2(t)] \tag{9}$$

An estimate of the drift parameter i.e., pull of the process towards the mean value may be obtained from (9). By law of iterated logarithm (LIL) of standard Brownian motion, e.g., see Chung (1948);

$$\overline{\lim}_{t\to\infty}(2t\log\log t)^{-1/2}B(t) = 1 \quad \text{a.s.} \tag{10}$$

and

$$\overline{\lim}_{t\to\infty}(2t\log\log\ t)^{-1/2}\sup_{0\le s\le t}\mid B(s)\mid = 1 \quad \text{a.s.} \tag{11}$$

Using the relationship $V(s) = e^{-\alpha s} B[\gamma^2(e^{2\alpha s} - 1)/2\alpha]$, see, e.g., Karlin and Taylor (1981); one may thus write from (10) and (11)

$$\overline{\lim}_{t\to\infty}[\frac{\gamma^2}{\alpha}(1 + o(1)) \log t]^{-1/2} V(t) = 1 \quad \text{a.s.} \tag{12}$$

and

$$\overline{\lim}_{t\to\infty}[\frac{\gamma^2}{\alpha}(1 + o(1)) \log t]^{-1/2} \sup_{0\le s\le t} \mid V(s) \mid = 1, \text{ a.s.} \tag{13}$$

Hence the fluctuation of the O–U process as seen from (12) and (13) is dependent on the parameter $\sqrt{2}\sigma_v = \gamma/\alpha^{1/2}$. Equating the observed value of the maximum fluctuation of the realised curve $\hat{V}(s)$ with $\sqrt{2 \log t}\ \sigma_v$, one may have an estimate $\tilde{\sigma}_v$ of σ_v.

In general, maximum fluctuations of two processes may serve as basic estimates of spread, without any model assumption.

References

Basawa IV, Rao BLSP (1980) Statistical inference for stochastic processes. Academic, London

Bose JC (1902) Response in the living and non-living. Longmans, Green & Co, London

Brown BM, Hewitt JI (1975) Asymptotic likelihood theory for diffusion processes. J Appl Probab 12:228–238

Chung KL (1948) On the maximum partial sum of sequences of independent random variables. Trans Am Math Soc 64:205–233

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74 (368):829–836

Dasgupta R (2006) Modeling of material wastage by Ornstein-Uhlenbeck process. Calcutta Stat Assoc Bull 58:15–35

Dasgupta R (2011) On the distribution of burr with applications. Sankhya B 73:1–19

Dasgupta R (2013) Yam growth experiment and above-ground biomass as possible predictor, Chap 1. In: Dasgupta R (ed) Advances in growth curve models: topics from the Indian Statistical Institute. Springer (USA) proceedings in mathematics & statistics, vol 46. Springer, New York, pp 1–33. doi:10.1007/978-1-4614-6862-2_1

Dasgupta R (2015) Longitudinal growth of elephant foot yam and some characterization theorems, Chap 14. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Diggle PJ, Heagerty P, Liang KY, Zeger SL (2003) Analysis of longitudinal data. Oxford University Press, New York

Foyer CH, Lelandais M, Kunert KJ (1994) Photooxidative stress in plants. Physiol Plant 92:696–717

Jia W, Zhang J (2008) Stomatal movements and long-distance signaling in plants. Plant Signal Behav 3(10):772–777

Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic, London

Louzada F, Ferreira PH, Diniz CAR (2014) Skew-normal distribution for growth curve models in presence of a heteroscedasticity structure. J Appl Stat 41(8):1785–1798

Maller RA, Müller G, Szimayer A (2009) Ornstein-Uhlenbeck processes and extensions. Handbook of financial time series. Springer, Berlin, pp 421–437

Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian motion. Phys Rev 36:823–841

Wildon DC, Thain JF, Minchin PEH, Gubb IR, Reilly AJ, Skipper YD, Doherty HM, O'Donnell PJ, Bowles DJ (1992) Electrical signalling and systemic proteinase inhibitor induction in the wounded plant. Nature 360:62–65. doi:10.1038/360062a0

Zimmermann MR, Maischak H, Mithöfer A, Boland W, Felle HH (2009) System potentials, a novel electrical long-distance apoplastic signal in plants, induced by wounding. Plant Physiol 149(3):1593–1600

Some Remarks on Pseudo Panel Data

Ratan Dasgupta, Jayanta K. Ghosh, Sugato Chakravarty, and Jyotishka Datta

Abstract We discuss the possibility of constructing pseudo panel data from cross-sectional data, sampled at different points in time, by aligning individuals sharing some common characteristics into groups called "cohorts". Based on a real-life example on income distribution in the USA, we construct and validate a pseudo panel data and compare this with real panel data. The agreement is encouraging.

Keywords Repeated cross section • Synthetic panel data • Cohort • Error-in-variable

1 Introduction

In many problems, where we wish to study a cross section of the population over time the natural approach is to use a panel data, also called longitudinal data by statisticians. To ensure comparability, true panel data should be based on responses to similar questions posed in a consistent manner and data collected from the same individuals repeatedly over time. Several researchers including Moffitt (1993) and Verbeek (2008) have pointed out the relative advantages of using a true panel data over repeated cross-sectional data. Russell and Fraas (2005) have noted that the formation of a true panel data is not problematic if the individuals belong to small set

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

J.K. Ghosh (✉)
Department of Statistics, Purdue University, West Lafayette, IN, USA

Statistics and Mathematics unit, Indian Statistical Institute, Kolkata, West Bengal, India
e-mail: jayantag1@gmail.com

S. Chakravarty
College of Consumer and Family Science, Purdue University, West Lafayette, IN, USA

J. Datta
Department of Statistical Science, Duke University and SAMSI, Durham, NC, USA

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_2

of entities, such as member countries of UN Security Council, and the data is based on unambiguous questions, e.g. population or other stable demographic variables of the member countries. However, issues like attrition and non-response pose a threat to the comparability even in the highest-quality panel data sets. This effect is more serious for studies involving a large number of entities defined as individual people or individual households spanning a long time interval. A natural workaround is to construct pseudo panel data from repeated cross-sectional surveys whenever they are available. Important examples of large repeated cross-sectional database are the Current Population Survey in the USA, and the Family Expenditure Survey in the UK. In this short note, we provide a brief review of the methods of constructing and validating a pseudo panel data and illustrate the methods on a publicly available data set on US Survey of Consumer Finance (SCF).

As we shall discuss later, pseudo panel data often provides a way to build models to study longitudinal effects of important variables, and has been widely applied in many fields including microeconomic research, and many important areas of social science where a genuine panel data is not available. One such example is lack of true (or "genuine") panel data on poverty related issues in India. Typically Indian data for poverty has been cross-sectional data sampled at different points in time, also called "waves". It has been felt that such data is easier to collect and less prone to error than panel data, which suffers from depletion as time proceeds and is expensive and time consuming to sample. Cross-sectional data on poverty in India in the 1960s and 1970s appear in detail in Bhattacharya et al. (1991). However, we have not seen any book length treatment of longitudinal data on similar topics. One way to solve this problem is to construct a pseudo panel data by following a number of cohorts over time, and observe how different aspects of their life, for example consumption of food, availability of clothes and shelter change.

In general, classification with respect to a suitably chosen cohort may reveal interesting growth pattern of individuals. While computing the overall growth pattern of height (in cm.) out of four yam stems sprouting on different dates from a single seed corm, in Dasgupta (2013), see Figs. 6.9, 6.11–6.14 therein; much different sprouting dates were observed for 4 stems, and corresponding dates were considered as time origin for each relevant stem. That is age (in day) of stems in a plant was taken as cohort for computing the response curve of that plant growth, thus extracting information from multiple stems.

Panel data on income considered in this note for explanation has a shorter span than the gap between successive cross-sectional data, stretching our imagination on data trend along the perceived direction (Figs. 2 and 3). With available data in hand, we construct pseudo panel data and compare this with panel data; some of the agreements as explained by figures are encouraging. The data example considered is not broad enough for demonstrating how pseudo panel data can be useful in estimating changes in variables. We plan to come back to address this issue in future, and the present report is preliminary.

Deaton (1985) observes that there are no panel data in the United Kingdom on consumer expenditure or on household labour supply, but there are several large household surveys carried out periodically. Deaton (1985) and Verbeek and

Nijman (1992) suggest an artificial panel data can be constructed by sampling the cross-sectional data suitably. This has been called a repeated cross section (RCS). Literature on panel data includes papers by Browning et al. (1985), Deaton (1985), Verbeek and Nijman (1992), Verbeek (2008) and Moffitt (1993).

One way of constructing pseudo panel data is as follows. First, synthetic panels are constructed by grouping individuals sharing some common characteristics (it may be the year of birth) into groups called "cohorts", and the averages within each cohort are taken as observations in a synthetic panel. This makes it possible to follow the same cohorts of individuals over a larger period of time. It is probable that this sort of construction would produce valid panel data, especially if the data size at each point of time is large. How large is often a subjective decision, as Propper et al. (2001) point out the trade-off between error in estimating the cohort mean (larger number of cohorts but fewer individuals per cohort) and lack of information (a few cohorts with many individuals in each cohort). In this note, we will follow the approach taken by Deaton (1985), and other literature on this topic by defining the cohorts by intervals of the year of birth and the race of the individual.

An important issue in the construction of pseudo panel data from repeated cross-sectional data is a principled comparison between pseudo panel data and true panel data. This is often rendered difficult by lack of co-existing true panel data and repeated cross-sectional data for the same variables on the same population for a common, reasonably large period of time. Assuming that such data sets are available for comparison, one may plot both time series, namely the real panel data and the pseudo panel data at the chosen point of time, and see how close the two data sets are with respect to certain averages and measures of deviation from the average. We will show a simple example of such a comparison for our SCF data set where we have both the repeated cross-sectional data and true panel data, albeit for a short period of time.

A further point worth studying is the effect of variations in the size of pseudo panel and the true panel over time, and whether the pseudo panel is too close to the true panel data.

If the pseudo panel is too close to the true panel data available; one needs to be alert. Like in Mendel's experiment, too low value of chi-square becomes significant on the left tail of test statistic distribution. Very small value of chi-square is an indication of possible over fitting and may raise uncomfortable question on validity of experiment and/or underlying assumptions of data analysis. However, if underlying error component is small, pseudo panel data may reproduce the unseen panel data. Consider a general model for ith individual $i = 1, 2, \ldots, N$, with time dependent observation $c_{yi}(t) = c_{\mu}(t) + \varepsilon_i$; where c is a cohort value and error $\varepsilon_i = \varepsilon_i(t)$ is negligible compared to the main component $\mu(t) = c_{\mu}(t)$. In such a situation whether it is real panel or pseudo panel data, the random ε part being small; average over randomly selected individuals will produce similar results over time, even with a small sample size n compared to N, where c is an arbitrarily fixed cohort. Further, use of improved nonparametric regression techniques insensitive to outliers e.g., LOWESS or spline regression with suitable choice of smoothing parameters, may correctly produce the underlying response curve; screening out the

error components $\varepsilon_i(t)$ even when fluctuation of these are of substantial magnitude, see e.g. Cleveland (1979, 1981). Response curve may then be deduced from either from pseudo panel or true panel data.

2 Pseudo Panel Data

As discussed in the Introduction, for several countries like India, the UK and the USA, there is a lack of true panel data where specific individuals or families are followed over time, but there exists a series of independent waves of cross-sectional data collected over time. One such example is the U.S. SCFs data set, which has both panel data and repeated cross-sectional data at different time-points, e.g. surveys during 1989–2007 are available as repeated cross-sections and the surveys during 2007–2009 are available as a panel data set.

While repeated cross-sections are considered inferior to genuine panel data for their lack of individual histories, they are often free from problems such as attrition and non-response, and are typically larger in number of individuals and the time-span. To draw inferences on RCS datasets, synthetic panels are constructed by grouping individuals sharing some common characteristics (for example, year of birth) into cohorts, and the averages within each cohort are taken as observations in a synthetic panel. Deaton (1985) suggests an error-in-variables estimator for obtaining consistent estimators as the cohort-averages are error-ridden measurements of the true cohort population parameters. However, in many cases, such as ours, as described later, the number of observations per cohort is large, which leads one to ignore the errors-in-variables problem and work with the synthetic panel as a genuine panel to obtain reliable estimators. Early work such as those by Deaton (1985) and Browning et al. (1985) have proposed the use of such estimators. For a detailed discussion on the conditions under which the standard estimators ignore the "errors-in-variable" problem, see Verbeek and Nijman (1992). Moffitt (1993) extends the application of synthetic panel approach of Deaton to non-linear and dynamic models. For an inclusive description of all the models, see Verbeek (2008).

Construction of Pseudo Panel Data We discuss the construction of pseudo panel data. We assume that the data set is a series of repeated cross-sections. Deaton (1985) suggests the use of cohorts to obtain consistent estimators, if repeated cross-sections are available. Towards this, following the work of Deaton (1985), Browning et al. (1985) and Verbeek (2008), we define C cohorts, which are groups of individual sharing some common characteristics, for our example it is the year of birth and the race of an individual. The cohorts are defined such that each individual is a member of exactly one cohort, which remains the same for all periods. This preempts the use of time-varying variables, like earning, to be used as a cohort defining criterion. The seminal study of Browning et al. (1985) uses 5-year age bands subdivided as to whether the head-of-household is a manual or a non-manual worker. Banks et al. (1994) use 5-year age bands. In our study, involving the SCF data there are repeated cross-sections available every 3 years from 1995 to 2007,

which prompts us to use 3-year age bands as one of the variables to define our cohorts. The lowest cohort takes into account all of the individuals aged 21 or less as of 1995, and the highest cohort takes into account all of the individuals aged 60 or more as of 1995.

We also considered the public datasets available on the Federal Reserve website from 1995 to 2007 in 3-year intervals (http://www.federalreserve.gov/econresdata/scf/scfindex.htm). The website also has panel data sets for the years 2007 and 2009. For both the repeated cross-sections and the true panel data, missing values were imputed five times using a multiple imputation technique, thus there are 22,085 observations in the 2007 repeated cross-sectional datasets for 4,417 families and 19,285 observations in the 2007 panel for 3,857 families included in the survey. We have randomly chosen one out of every five implicates for our analysis. For constructing the cohorts, we chose 3-year age bands and Race class (1 = white non-Hispanic, 2 = non-white or Hispanic), and define the cohorts as age-groups divided by race class. As we discussed in the introduction, one important aspect of analysing the pseudo panel data is a comparison of the pseudo panel data with the true panel data if they are available for the same period of time. However, this is difficult in practice due to unavailability of such data sets. In our case, the cross-sectional data and the true panel data for the SCF data base overlap only in the year 2007. A simple way of validating the constructed pseudo panel data is to compare the cohort-wise distribution of a few variables from the pseudo panel data and the true panel data over the period of overlap.

Figure 1 shows the distribution of mean income, mean asset and mean networth over the cohorts for the pseudo panel data constructed from repeated cross-sectional data sets from 1995 to 2007. The similarity between the top row and bottom row of Fig. 1 indicates the cohorts for either data set have similar characteristics. There is also remarkable similarity across different variables, particularly between mean asset and mean networth.

To compare across different time-points and also to see how the cohort-wise distribution changes over time, we have plotted the mean income for the pseudo panel data for 1995–2007 and the true panel data for 2007 and 2009 versus the age cohorts in Figs. 2 and 3. The distributions look similar providing further support to the method of construction adopted in this note.

One may identify the positions of black hollow square data points in Fig. 3 with similar position of green hollow square data points in Fig. 2; these correspond to the year 2007, the common year of true panel and pseudo panel data, those we say are close.

During the Great Recession, the median U.S. household income (in 2011 dollars) dropped from $54,489 in 2007 to $52,195 in 2009, a loss of 4.2 %.

This is indicated as lowering the peak of income distribution in 2009 compared to 2007 in Fig. 3.

The above-mentioned recession is discussed in National Bureau of Economic Research (NBER), the largest economics research organization in the United States;

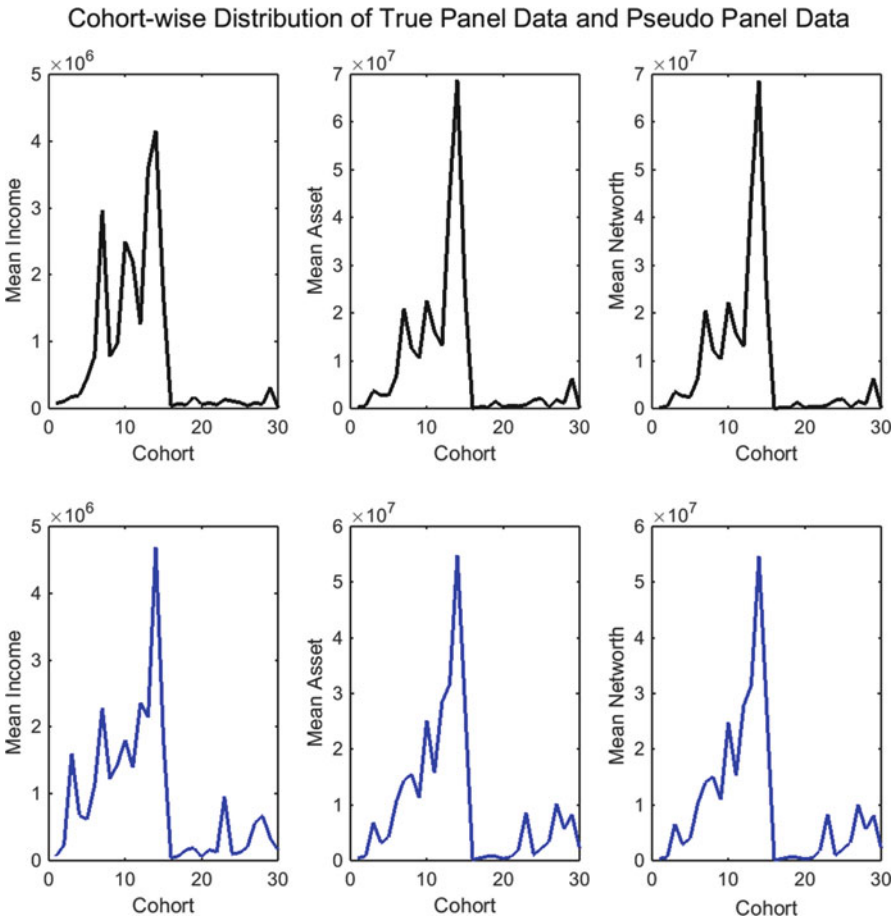


Fig. 1 Distribution of mean income, mean asset and mean networth by Cohorts for both the pseudo and true panel data for the year 2007

business cycle dates are determined by the NBER (http://www.nber.org/cycles/cyclesmain.html). According to NBER, the Great Recession started in December 2007 and ended in June 2009.

The disastrous economic event in the years 1930s is termed as "Great depression" in the literature. The later event of 2007–2009 is termed as "Great recession".

3 Linear Fixed Effects Model

In this section, we briefly state linear fixed effects model, which provides the effect of the predictors that vary over time (analysis done in Sect. 4). The key assumption underlying this model is that the unobserved variables that have an effect on both

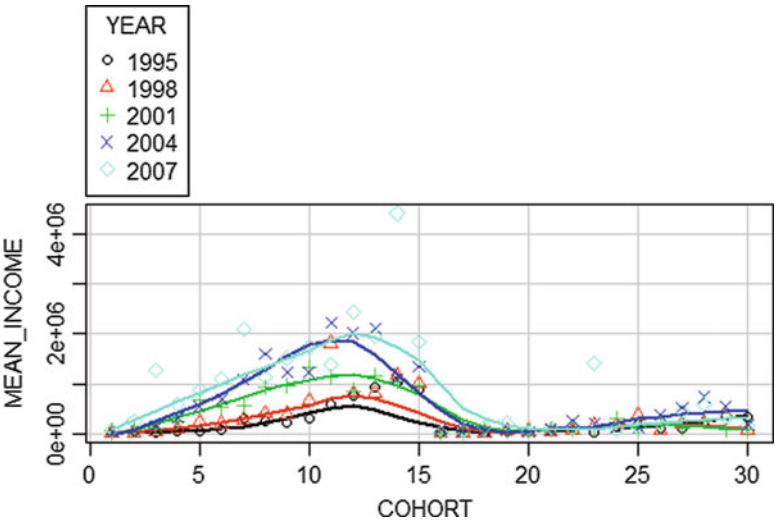


Fig. 2 Distribution of mean income by Cohorts for pseudo panel data over the years 1995–2007. First 15 cohorts are 3-year age-groups for race class = 1 (white, non-Hispanic), the last 15 cohorts are 3-year age-groups for race class = 2 (Hispanic or non-white)

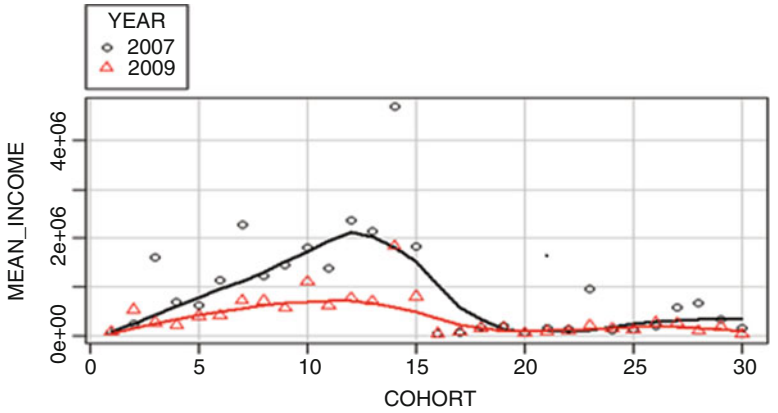


Fig. 3 Distribution of mean income by Cohorts for true panel data over the years 2007 and 2009. First 15 cohorts are 3-year age-groups for race class = 1 (white, non-Hispanic), the last 15 cohorts are 3-year age-groups for race class = 2 (Hispanic or non-white)

the predictor and the response variable are time-invariant in nature. The fixed effects model also assumes that these time-invariant characteristics are unique to the entity (in our case, the cohort) and hence they are not correlated among themselves.

$$y_{it} = x_{it}'\beta + \theta_i + \epsilon_{it}, \quad i = 1, 2, \ldots, N(t), \quad t = 1, 2, \ldots, T \tag{1}$$

where x_{it} denotes a $k \times 1$ vector of explanatory variables and β is the parameter (vector) of interest, where i indexes individuals and t indexes time. We assume, for simplicity, that

$$E\left(x'_{it}\epsilon_{js}\right) = 0 \;\; \forall s,t = 1,2,\ldots,\; T \quad and \quad \forall \;\; i,\; j$$

Aggregation of all observations to cohort level results in the following model

$$\overline{y_{ct}} = \overline{x'_{ct}}\beta + \overline{\theta_{ct}} + \overline{\epsilon_{ct}},\; c = 1,\; 2,\ldots,C,\; t = 1,2,\ldots,\; T \tag{2}$$

where $\overline{y_{ct}}$ and $\overline{x'_{ct}}$ are the averages of all observed y's and **x**'s in cohort c at time t. The resulting data set is a synthetic (or pseudo) panel data set with repeated observations on C cohorts over T time-periods. The main problem with this approach is that $\overline{\theta_{ct}}$ depends on t and is likely to be correlated with $\mathbf{x}_{\mathrm{ct}}$'s. Therefore, the use of $\overline{\theta_{ct}}$ as fixed will lead to identification problems, unless the temporal variance of $\overline{\theta_{ct}}$ over t can be neglected ($\overline{\theta_{ct}} = \overline{\theta_c}$) which is the case when the number of observations per cohort is large, as is true in our case. Consider the within-estimator on the pseudo panel,

$$\beta_W = \left(\sum\nolimits_{c=1}^{C}\sum\nolimits_{t=1}^{T}\left(\overline{x_{ct}} - \overline{x_c}\right)\left(\overline{x_{ct}} - \overline{x_c}\right)'\right)^{-1}\sum\nolimits_{c=1}^{C}\sum\nolimits_{t=1}^{T}\left(\overline{x_{ct}} - \overline{x_c}\right)\left(\overline{y_{ct}} - \overline{y_c}\right)' \tag{3}$$

We assume that the number of cohort C is constant and the number of individuals tends to infinity. Thus, the number of individuals per cohort tends to infinity.

4 Application to the SCF Data

For the results shown in this report, we consider a linear model with asset as the dependent variable and income and net worth as the regressors. As mentioned earlier, the number of observations per cohort is large for each time period. Below, we describe the model we fitted to both the datasets using net worth as the dependent variable and income as the regressor. We considered a fixed effects model as in (1), i.e.

$$y_{it} = x'_{it}\beta + \theta_i + \epsilon_{it},\;\; i = 1,\; 2,\ldots,\; N(t),\quad t = 1,2,\ldots,\; T \tag{4}$$

where $\mathrm{y_{it}}$ is the asset and $\mathrm{x_{it}}$ is the net worth, and θ_i's are the fixed effects. We use the standard within estimator for fixed effects model given in (3). The estimates and the standard errors for the RCS data and the true panel data are given below. Standard errors are small compared to estimates of parameters (Table 1).

Table 1 The estimate of β for the fixed effects model using both true panel data and pseudo panel data

Variable	Estimate	Standard error
Mean networth (true panel)	1.009518	0.00278
Mean networth (RCS)	1.0075	0.00157

5 Discussion

We give a brief overview of the method of constructing and validating pseudo panel data from the widely available repeated cross-sectional surveys. The pseudo panel data is easy to construct and it enjoys a few substantial advantages over both true panel data and repeated cross-sectional data by avoiding errors due to non-response or attrition. We illustrate, with the help of the U.S. SCF data, the construction of pseudo panel data and discuss the validation of these by comparing the cohort-wise distribution and comparing the parameter estimates obtained from same model fit to both the datasets. Panel data considered has a shorter span than the gap between successive cross-sectional data, making the task of practical demonstration a bit difficult. We explained the pseudo panel data construction and comparison with present data, some of the agreements as explained by figures are encouraging. There are a few possible directions for future research, for example, one might construct the cohorts in a non-parametric way to identify the natural clusters in the data, and then finding suitable time-invariant variables to define the clusters. This could lead to an optimal choice of the cohort width that ensures both prevention of loss of information and sufficient estimation accuracy for the cohort means.

Acknowledgements We thank the referee for providing constructive comments that helped in improving the presentation of this short note. We also thank Mr. Piyas Chakraborty for his help regarding the preparation of this note.

References

Banks J, Blundell R, Preston I (1994) Life-cycle expenditure allocations and the consumption costs of children. Eur Econ Rev 38(7):1391–1410

Bhattacharya N, Coondoo D, Maiti P, Mukherjee R (1991) Poverty, inequality and prices in rural India. Sage Publication, New Delhi

Browning M, Deaton A, Irish M (1985) A profitable approach to labor supply and commodity demands over the life-cycle. Econometrica 53:503–543

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74(368):829–836

Cleveland WS (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35(1):54

Dasgupta R (2013) Optimal-time harvest of elephant foot yam and related theoretical issues, Chap 6. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 101–129

Deaton A (1985) Panel data from time series of cross-sections. J Econ 30(1–2):109–126

Moffitt R (1993) Identification and estimation of dynamic models with a time series of repeated cross-sections. J Econ 59(1–2):99–123
Propper C, Rees H, Green K (2001) The demand for private medical insurance in the UK: a cohort analysis. Econ J 111:180–200
Russell JE, Fraas JW (2005) An application of panel regression to pseudo panel data. Mult Lin Regression Viewpoints 31(1):1–15
Verbeek M (2008) Pseudo-panels and repeated cross-sections. In: Matyas L, Sevestre P (eds) The econometrics of panel data. Springer, Berlin, pp 369–383
Verbeek M, Nijman T (1992) Can cohort data be treated as genuine panel data? Empir Econ 17(1):9–23

Rates of Convergence in CLT for Two Sample U-Statistics in Non iid Case and Multiphasic Growth Curve

Ratan Dasgupta

Abstract We obtain nonuniform rates of convergence in central limit theorem for two sample U-statistics in non iid case when moment generating function of the kernel ϕ necessarily exists, but the kernel may not be bounded. The rates are sharp when the kernel is bounded, like in the case of Wilcoxon two sample U statistics. Precision of these results motivates to explore data analysis of plant growth in the set-up of U-statistics. Growth patterns of Sisal plants, having high economic return for extracted leaf fibres, are tested for two different growth environment by two sample Wilcoxon U statistic. In the Indian Statistical Institute (ISI) Giridih farm these plants are grown in two different types of land viz., a high land with rock layer below topsoil having scarcity of irrigation, and the other with sandy soil structure near a hilly rivulet occasionally flooded in rainy seasons for a few days. The latter environment turns out to be more conducive for growth. We study plant growth viz., growth in number of leaves and plant height from field experiments. These variables are further studied for a subgroup of randomly sampled plants. Length and mid width of sisal leaves are studied for overall growth. Proliferation rates and second derivatives are also calculated. Almost sure confidence bands for sisal growth curves are computed in the set-up of U-statistics. These reveal multiphasic growth patterns. The study is of interest in assessing economic potential of sisal plantation in Jharkhand.

Keywords Agave sisalana • Longitudinal study • Nonparametric regression Bulbil • Nonuniform L_p version of the Berry–Esseen theorem • Cross-sectional data • Smoothing spline

MS subject classification: Primary: 62G08, secondary: 62P10

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_3

1 Introduction

We compute rates of convergence in central limit theorem for two sample U-statistics in non iid case under a moment condition that ensures existence of moment generating function for the kernel ϕ, but the kernel is not necessarily bounded. Nonuniform Berry Esseen bound gets sharper as the assumption varies from existence of m.g.f. to boundedness of the kernel ϕ. For decomposition of U statistic in one sample and two sample cases and convergence rates in CLT along with allied results see e.g., Hoeffding (1948), Ghosh and Dasgupta (1982), Dasgupta (1984, 2008, 2013) and the references given therein.

In India Sisal *(Agave sisalana)* is mainly grown in arid and semi-arid regions of Andhra Pradesh, Jharkhand, Orissa, Karnataka, Maharashtra, and West Bengal. By two sample U statistic we compare growth pattern of Sisal planted in two types of land. One is a high land with rock layer below topsoil and with less irrigation, and the other has sandy soil structure occasionally flooded by a hilly rivulet *Usri* in rainy season. Comparison by two sample Wilcoxon U statistic indicates that the riverside land is significantly better than the upper land for sisal growth.

Sisal fibres have high economic value, for these are stronger than jute fibre. Strength for sisal fibre is studied in Inacio et al. (2010). From each sisal leaf 4–4.5 % of hard parallel fibres are extracted by machine decortications in which the leaf is crushed between the rollers and then mechanically scraped. A healthy sisal plant produces about 200–250 leaves during its 10–12 years of life span, after which it produces long flowering axis called "pole". A pole produces the bulbils which can be 400–800 or more in numbers, and are used as seedlings for sisal cultivation. See, e.g., Lock (1969) and Gentry (1982).

We examine the longitudinal growth of sisal plant height and number of leaves for several years in Indian Statistical Institute (ISI) Giridih farm. Multiphasic growth pattern over years is seen in the data analysed.

In Sect. 2 we briefly recapitulate Hoeffding decomposition of two sample U statistic in non iid case. We examine equivalence of two assumptions, one in terms of moments of kernel ϕ, the other is in terms of moment generating function of some function of ϕ. In Sect. 3 we compute the nonuniform rates of convergence in CLT and use these to obtain moment type convergences and nonuniform L_P version of Berry–Esseen theorem. Precision of these results motivates us to explore data analysis of sisal growth in the set-up of U-statistics. Section 4 deals with cross sectional comparison of two growth environment for sisal plants by two sample Wilcoxon statistic. We further study longitudinal growth curve of sisal in terms of plant height, leaf length and width and obtain almost sure confidence bands for growth curves. Derivative, proliferation rate and second derivative curves of sisal growth are obtained. Confidence band for growth curve for variance of sisal over time is studied in a set-up of U-statistics. Seasonal variations over years are reflected in multiphasic growth patterns in estimated curves.

2 Decomposition of Two Sample U Statistic in Non Iid Case

Let $U_{n,m}$ be a two sample U-statistic based on the independent but not necessarily identically distributed random variables $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$ with kernel ϕ and degree (r,s) i.e.,

$$U = (n_{C_r} m_{C_s})^{-1} \sum_{\substack{1 \le i_1 < \cdots < i_r \le n \\ 1 \le j_1 < \cdots < j_s \le m}} \phi(X_{i_1}, \cdots, X_{i_r}; Y_{j_1}, \cdots, Y_{j_s}) \tag{1}$$

where the kernel ϕ is symmetric in X_i's and Y_j's. Without loss of generality we may assume

$$E\phi(X_{i_1}, \cdots, X_{i_r}; Y_{j_1}, \cdots, Y_{j_s}) = 0, \quad \forall\, i_1 \ne \cdots \ne i_r, \quad j_1 \ne \cdots \ne j_s. \tag{2}$$

For Hoeffding decomposition of U statistics, we adopt the notations of Dasgupta (2008, 2013). With $r = s = 2$, it was shown therein that the main part V_1 is sum of independent random variables and the remainder with components V_2, V_3, V_4 may be considered negligible. The main part V_1 is of the form

$$V_1 = \frac{2}{n} \sum_{i_1=1}^{n} \overline{\psi}^{(1)}(X_{i_1}) + \frac{2}{m} \sum_{i_3=1}^{m} \overline{\psi}^{(1)}(Y_{i_3}),$$

and

$$\begin{aligned} U &= V_1 + V_2 + V_3 + V_4 \\ &= V_1 + R_{n,m} \ \text{ where } \ R_{n,m} = V_2 + V_3 + V_4. \end{aligned} \tag{3}$$

Note that V_1 is a weighted sum of independent random variables for which the standard theory applies. V_1 involves conditional expectation of ϕ fixing one coordinate. In the general case the coefficients in the sum for V_1 are r/n and s/m in the place of $2/n$ and $2/m$, respectively.

A moment bound for the remainder $R_{n,m}$ is stated in Proposition 1 of Dasgupta (2013). The result in brief is presented below.

Proposition A. *For a two sample U statistic defined in (1) with kennel ϕ with degree (r, s) let (2) hold and for an integer $q \ge 1$, let*

$$\delta_q = \sup_{\substack{m \ge 2 \\ n \ge 2}} [\tbinom{n}{2}\tbinom{m}{2}]^{-1} \sum_{\substack{1 \le i_1 < i_2 \le n \\ 1 \le i_3 < i_4 \le m}} E|\phi(X_{i_1}, X_{i_2}, Y_{i_3}, Y_{i_4})|^{2q} < \infty. \tag{4}$$

Then, for a constant $L(> 1)$ independent of m, n and q

$$ER_{n,m}^{2q} \leq n^{-2q}\ L^q (vq)!\ \delta_q \tag{5}$$

under the assumption $m = O_e(n)$, *and* $v = r + s$ *is the number of arguments in* ϕ. *Now consider the bound*

$$\delta_q \leq L^q e^{vq \log q} \tag{6}$$

$\forall q > 1$, *where* $L > 0,\ v \in (0, 1)$. *The above condition is implied by*

$$\sup_{n\geq 1, m\geq 1}(n_{c_2} m_{c_2})^{-1} \sum_{\substack{1 \leq i_1 < i_2 \leq n \\ 1 \leq j_1 < j_2 \leq m}} E \exp(s|\phi|^{1/v}) < \infty, \tag{7}$$

where $0 < s < s_o = ve^{-1}L^{-1/v}$ *and* $\phi = \phi(X_{i_1}, X_{i_2}, Y_{j_1}, Y_{j_2})$.

This follows along the lines of Dasgupta (2006), see Proposition 2.1 and Remark 2.2 therein. Existence of m.g.f. for ϕ corresponds to the case $v = 1$.

Under the above condition, one has $ER_{n,m}^{2q} \leq n^{-2q}\ L^q e^{(v+v)q \log q}$.

3 Rates of Convergence for U Statistics

Two sample U statistic is of the form of a nonlinear statistic considered in Dasgupta (2006),

$$T_n = s_n^{-1}\ S_n + R_n, \tag{8}$$

where $S_n = \Sigma_{i=1}^n X_{ni}, s_n^2 = \Sigma_{i=1}^n EX_{ni}^2, \inf_{n\geq 1}\ n^{-1}\ s_n^2 > 0$.

$X_{n1}, X_{n2}, \ldots, X_{nn}$ are independent random variables in a triangular array with zero expectation and R_n is a negligible remainder. For some $\beta \geq 0$, let R_n be small in the sense that

$$E\ |\ R_n\ |^q \leq c(q) n^{-q/2} (\log\ n)^{\beta q}, q > 1, \tag{9}$$

where $c(q) \leq L_1^q e^{(\gamma+\delta)\ q\ \log\ q}$, for some $\delta \geq 0$ and $L_1 > 0$.

Now write, as in Dasgupta (2013)

$$U_n^* = U_{n,m}^* = [\text{var}(V_1)]^{-1/2} U_{n,m} = [\text{var}(V_1)]^{-1/2} V_1 + R_{n,m}^* \tag{10}$$

where $V_1 = \frac{2}{n}\sum_{i=1}^n \overline{\psi}^{(1)}(X_i) + \frac{2}{m}\sum_{i=1}^m \overline{\psi}^{(1)}(Y_i)$; $R_{n,m}^* = [\text{var}(V_1)]^{-1/2} R_{n,m}$,

$\sigma^{*2} = \text{var}(V_1) = 4(\sum_{i=1}^n E[\overline{\psi}^{(1)}(X_i)]^2/n^2 + \sum_{i=1}^m E[\overline{\psi}^{(1)}(Y_i)]^2/m^2) = O_e(\frac{1}{n+m})$,

$\sigma_n^2 = \sigma_{n,m}^2 = (n+m)^2\sigma^{*2} = (n+m)^2 \text{var}(V_1) = O_e(n+m) = O_e(n)$, provided

$$\inf_{n\geq 1} n^{-1}\sum_{i=1}^{n} E[\overline{\psi}^{(1)}(X_i)]^2 > 0, \quad \inf_{m\geq 1} m^{-1}\sum_{i=1}^{m} E[\overline{\psi}^{(1)}(Y_i)]^2 > 0. \tag{11}$$

Let $L > 1$ be a generic constant. The first term in the r.h.s. of (10) is then a standardised sum of independent random variables and the second term is remainder with

$$\begin{aligned} E(R^*_{nm})^{2q} &\leq n^{-q}\ L^q(vq)!\ \delta_q,\ v = r+s \text{ is the number of arguments in } \phi \\ &\leq n^{-q}\ L^q e^{(v+\nu)q\log q},\ \forall q > 1, \text{under (6)} \end{aligned} \tag{12}$$

Then proceeding like Theorem 4.1 in Dasgupta (2006), one may obtain

Theorem 3.1. *Under the assumption $m = O_e(n)$ and (6)/(7), i.e., averaged qth absolute moment of kernel ϕ is of order $L^q \exp(\nu q \log q)$, $L > 0, 0 < \nu < 1$, $\forall q > 1$; or equivalently, averaged value of $E\exp(s \mid \phi \mid^{1/\nu})$ is finite for some $s > 0$, there exist constants $b(> 0)$, and $k \in (0, 1/2)$ such that the following holds for the standardised two sample U statistics U^*_n defined in (10),*

$$\mid P(U^*_n \leq t) - \Phi(t) \mid \leq b\, n^{-1/2}(\log\ n)^{v+\nu} \exp\ (-k \mid t \mid^{2\wedge 1/(v+\nu)}), \tag{13}$$

where $v = r + s$, the number of arguments in the kernel ϕ and $-\infty < t < \infty$.

The uniform bound of convergence associated with (12) is nearly optimal, the nonuniform part depending on t is exponentially decaying, and is of interest. In the case of Wilcoxon statistic $v = 2$ and $\nu = 0$.

The following results on moment type convergence and L_q version of nonuniform Berry–Esseen theorems are immediate from Theorem 3.1. See also Theorem 2.5 and Corollary 2.1 of Dasgupta (1992).

Theorem 3.2. *Let the assumptions of Theorem 3.1 be satisfied. Let $g : (-\infty, \infty) \to [0, \infty)$ be a even function, $g(0) = 0$ and $Eg(T) < \infty$, where T is a normal deviate. Suppose, $g'(x) = O[\exp(k \mid x \mid^{2\wedge 1/(v+\nu)})(1+ \mid x \mid)^{-q}]$, $q > 1$. Then,*

$$\mid Eg(U^*_n) - E\ g(T) \mid = O(n^{-1/2}).$$

Corollary 3.1. *Under the assumptions of Theorem 3.1,*

$$\| \exp(k \mid t \mid^{2\wedge 1/(v+\nu)})(1+ \mid x \mid)^{-q/p}(\mid P(U^*_n \leq t) - \Phi(t))\|_p = O(n^{-1/2}),$$

for any $q > 1$.

4 Data Analysis

Sisal is a drought resistant plant requiring little care with high economic value for leaf fibres. To compare two growth environments, data on length (x_{ij}) and mid leaf width (y_{ij}), $i = 1, 2; j = 1, \cdots, n_i$, for all healthy leaves were collected from two fully grown mature plants, one grown in a high dry land and the other grown in a riverside plot with sandy soil structure. These were planted years back, around the same time ($\pm 2/3$ days). Data (in cm.) on $n_1 = 46$ and $n_2 = 48$ leaves from two plants are presented in Tables 1 and 2.

Wilcoxon two sample U statistics with bounded kernels:

$I(x_1 > x_2)$, $I(y_1 > y_2)$ and $I(x_1 y_1 > x_2 y_2)$ may be considered for testing equality of two population in terms of leaf length, width and a measure of leaf area, respectively, for the sisal data of Tables 1 and 2.

For the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other, the Wilcoxon test is an efficient nonparametric test with null asymptotic distribution of standardised U as $N(0, 1)$.

The standardised value of the statistic $U^* = (U - \frac{n_1 n_2}{2})/\{\frac{n_1 n_2 (n_1+n_2+1)}{12}\}^{1/2}$ with kernel involving x, y and xy are $6.739179, 0.4084351$ and 4.008715, respectively, to be compared with a normal deviate. First and last values are highly significant, indicating that riverside plot is more conducive for sisal growth.

Sisal is a sturdy plant that can adapt to harsh environment. A plant from highland was uprooted two years back that made its roots established on riverbank, outside the boundary wall of farm. This unprotected plant prone to damage has number of leaves, $n_1 = 74$. This unguarded plant nearer to hilly rivulet is compared with one of the best Sisal plants from upland with number of leaves, $n_2 = 106$. Data (in cm.) of these two plants are provided in Tables 3 and 4.

As before we compare leaf length x, mid leaf width y, and xy; a measure of leaf area by the above mentioned Wilcoxon two sample U statistics; the standardised value of the statistic U^* in this case are $3.26775, 0.596, 2.46826$, respectively, to be compared with a normal deviate. Significance of first and last values provide ample evidence of conducive growth by riverside land for Sisal plants.

We next consider longitudinal study of plants in high land. Sisal plants cultivated in ISI Giridih farm show a growth pattern $y = y(t)$ that reveals a seasonal variation over time t. Growth curves of number of leaves averaged over 180 plants on different time points, shown in Figs. 1 and 2 using the nonparametric regression techniques of lowess (using SPlus with $f = .11$) and smoothing spline (using SPlus smooth.spline, with spar= 0.00001), over a period of four years exhibit slow rate of growth in summer and winter, both being extremely harsh in that region.

The same is reflected in the growth curve of height of 179 plants in Fig. 3 (lowess smoothing in SPlus with $f = .11$), and Fig. 4 (spline smoothing in SPlus smooth.spline, with spar= 0.00001); exhibiting features of step function.

Table 1 Leaf height and width (in cm.) of 46 Sisal leaves in a plant (riverside)

Sl. no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Leaf height	82.5	96.4	87.0	79.0	74.0	106.0	103.0	88.0	102.0	97.0	109.0	94.0	88.0	97.0	104.2
Leaf width	4.5	6.5	7.0	4.0	5.5	8.0	7.0	4.5	7.3	7.0	7.6	6.0	5.0	7.0	7.0

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
84.5	98.0	98.0	100.0	110.0	111.0	99.0	92.0	96.0	104.0	112.0	103.0	98.0	96.0	94.0
3.8	6.5	6.0	6.0	7.5	8.0	7.5	4.5	7.0	7.0	8.0	6.5	7.5	6.5	6.0

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
95.5	106.0	98.0	112.0	104.0	102.0	105.5	94.0	96.5	104.0	91.0	96.0	111.0	111.5	106.5	102.0
7.0	8.0	6.0	7.5	4.5	6.5	6.5	7.2	5.5	7.5	5.5	6.5	8.4	8.0	6.5	6.2

Table 2 Leaf height and width (in cm.) of 48 Sisal leaves in a plant (high land)

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Leaf height	80.0	78.0	87.0	71.0	85.0	49.0	90.0	91.0	81.0	75.0	73.0	73.5	86.0	79.0	91.0	83.5
Leaf width	6.0	5.5	6.0	5.5	6.5	7.0	7.0	7.0	6.5	5.5	3.5	5.0	6.0	6.5	7.0	6.5

17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
90.0	89.0	90.5	88.0	90.0	72.0	79.5	75.0	85.5	74.5	80.0	90.5	90.0	88.5	86.0	81.0
7.5	7.5	6.5	6.5	8.0	5.0	6.0	6.0	6.0	5.5	6.0	7.0	8.5	7.0	6.0	6.0

33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
71.0	75.0	79.0	46.0	84.5	92.5	94.5	93.0	91.0	87.0	83.0	93.0	92.0	91.0	91.0	90.5
5.0	6.0	6.0	5.5	6.5	6.5	6.5	7.5	6.0	7.0	7.0	6.5	7.0	6.5	6.5	6.0

Table 3 Leaf height and width (in cm.) of 74 Sisal leaves in a plant (riverside)

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Leaf height	108.0	111.5	112.5	115.5	115.5	116.0	114.5	115.0	117.0	107.0	104.0	96.0	91.0	93.5	106.0
Leaf width	7.5	8.5	6.5	8.0	8.5	6.5	7.5	7.5	7.7	7.0	6.0	7.5	4.5	7.0	8.0

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
97.0	80.5	84.0	114.0	105.0	115.0	120.0	119.0	111.0	98.0	96.0	119.0	120.0	98.0	105.0
8.5	6.5	7.0	8.0	8.5	7.5	8.0	9.0	8.0	7.5	7.0	8.5	7.5	7.5	9.0

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
124.0	120.0	102.0	112.5	108.0	121.5	120.0	107.0	99.0	106.0	119.0	120.0	117.0	109.0	120.5
10.0	8.5	8.0	8.1	8.5	10.0	10.5	9.0	8.5	8.5	10.0	8.2	9.5	9.0	9.0

46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
101.0	94.0	105.0	111.0	122.0	99.0	119.0	120.0	119.0	108.0	103.0	115.0	116.0	115.5	96.0
8.0	8.0	8.5	8.6	10.0	8.0	9.0	8.5	8.0	8.5	8.0	7.5	8.0	8.5	8.0

61	62	63	64	65	66	67	68	69	70	71	72	73	74
92.0	94.0	71.0	80.0	106.0	115.0	114.5	116.0	116.5	115.5	116.0	116.5	111.0	116.0
6.0	5.5	6.5	7.5	8.0	9.0	8.0	9.5	9.5	8.0	8.0	8.0	7.5	8.0

Table 4 Leaf height and width (in cm.) of 106 Sisal leaves in a plant (high land)

Sl. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Leaf height	105.0	99.0	93.0	98.0	90.0	95.0	103.0	108.0	114.0	119.0	118.0	97.0	94.0	92.0	89.0
Leaf width	7.0	7.5	7.5	7.5	5.0	7.5	7.5	8.5	7.5	9.0	9.0	7.5	7.5	7.5	6.5

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
96.0	79.0	107.0	99.0	106.0	118.5	118.0	117.5	120.0	109.0	95.0	88.0	77.0	85.0	77.0
7.5	7.0	8.0	8.0	8.5	10.0	10.5	10.5	9.5	8.5	7.5	7.5	6.0	7.0	6.5

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
86.0	95.0	89.0	97.0	102.0	93.0	95.0	111.0	116.0	116.0	117.0	110.0	106.5	98.0	93.0
6.5	6.0	6.5	7.5	8.0	7.0	7.5	7.0	10.0	8.5	9.5	7.5	8.0	7.5	6.5

46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
79.0	94.0	94.0	102.0	94.0	113.0	107.0	112.0	115.0	114.0	108.0	113.5	106.0	97.0	92.5
6.0	6.0	7.0	8.0	7.0	8.0	8.0	9.0	10.0	9.0	7.5	10.5	8.0	8.0	6.5

61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
91.0	92.0	89.0	91.0	104.0	108.0	112.0	115.0	110.0	108.0	97.0	94.0	92.0	102.5	94.0
7.0	7.0	7.5	7.0	7.5	7.5	9.5	10.0	7.0	8.0	7.5	6.5	7.0	8.0	7.5

76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
93.0	109.0	114.0	113.0	114.0	118.0	108.0	94.0	95.0	91.0	94.0	103.0	114.5	112.0	96.0
6.5	7.5	8.5	8.5	10.0	9.5	8.0	6.5	7.0	7.0	7.0	7.5	9.5	7.0	8.0

91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106
112.5	118.5	117.0	120.0	116.5	102.0	94.0	97.0	93.0	105.0	110.0	117.0	119.0	121.0	119.0	120.5
7.0	10.5	9.5	10.0	9.0	8.0	7.0	7.5	6.5	7.5	8.0	9.5	8.5	9.5	10.0	9.0

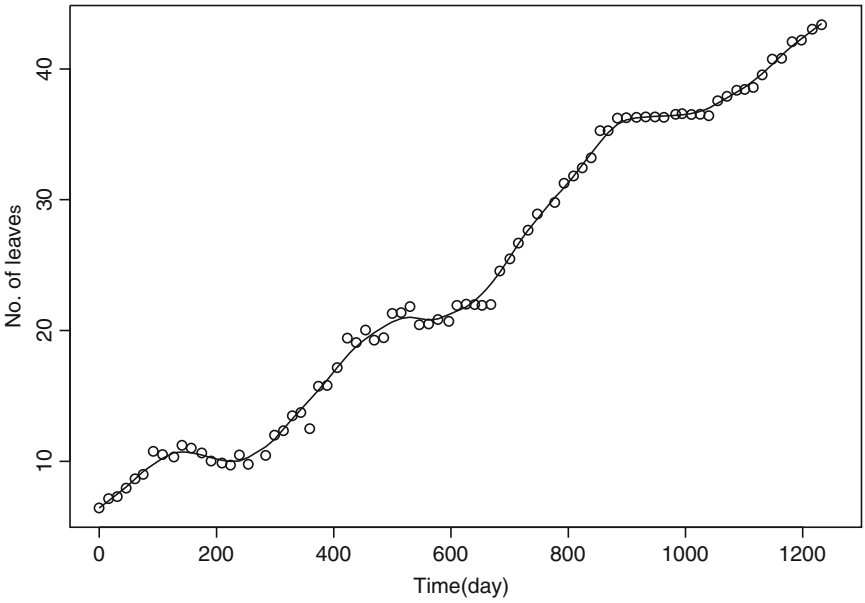


Fig. 1 Growth curve of number of leaves averaged over 180 plants on 160 different time points is shown in Fig. 1 using the nonparametric regression techniques of lowess (using SPlus with $f = 0.11$) over a period of about four years. This exhibits slow rate of growth in summer and winter, both being extremely harsh in that region. The curve has step-function like features

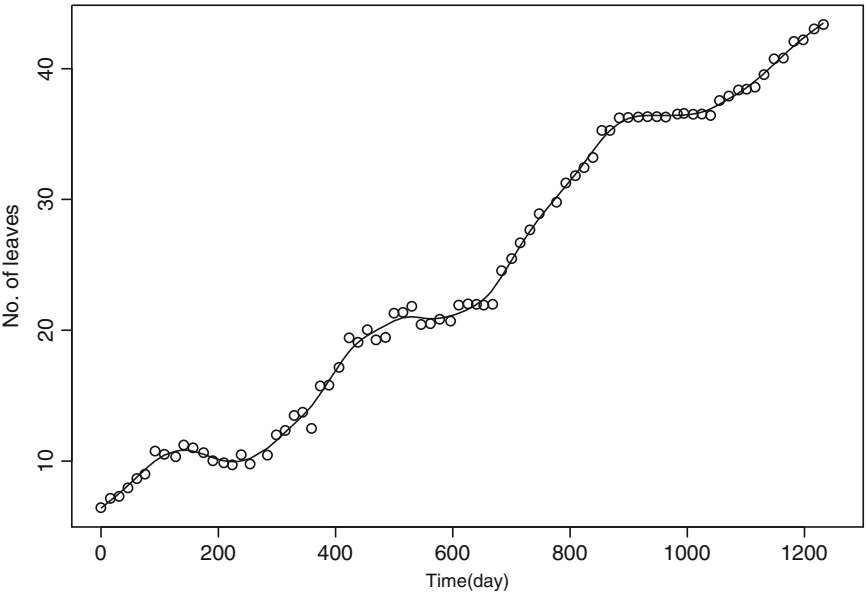


Fig. 2 In this counterpart of Fig. 1, nonparametric spine regression using SPlus smooth.spline software with spar $= 0.00001$ provides a smoother growth curve of sisal leaf number. Features of step function like growth present are similar to the previous figure

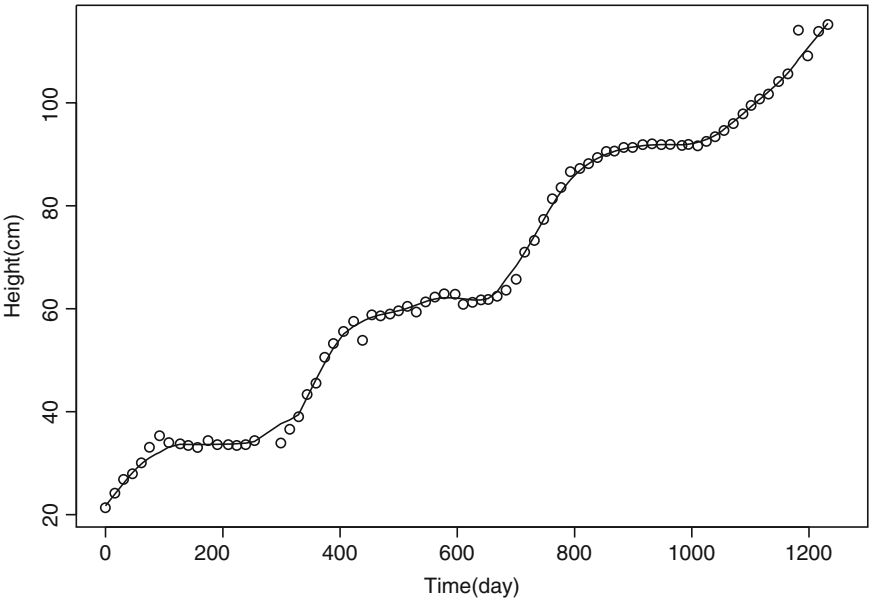


Fig. 3 Similar step function like feature is reflected in the growth curve of height of 179 plants in Fig. 3 obtained by lowess smoothing in SPlus with $f = 0.11$

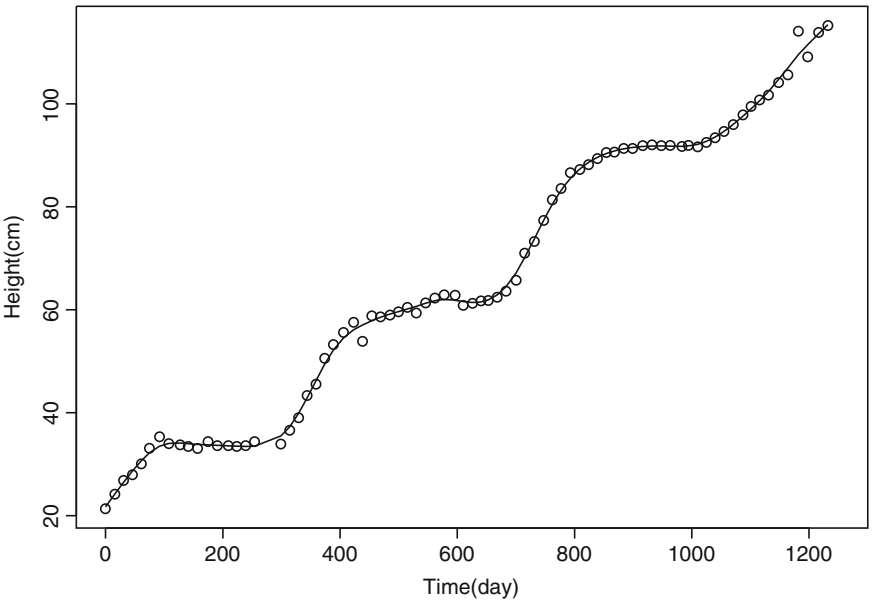


Fig. 4 The growth curve of height of 179 plants shown in Fig. 4 is obtained by spline smoothing in SPlus smooth.spline, with spar $= 0.00001$. The curve is relatively smooth compared to the curve in Fig. 3. Growth of sisal is steep in rainy season as seen in the curves. Giridih has two spells of rain in a year, causing fluctuations in growth curves

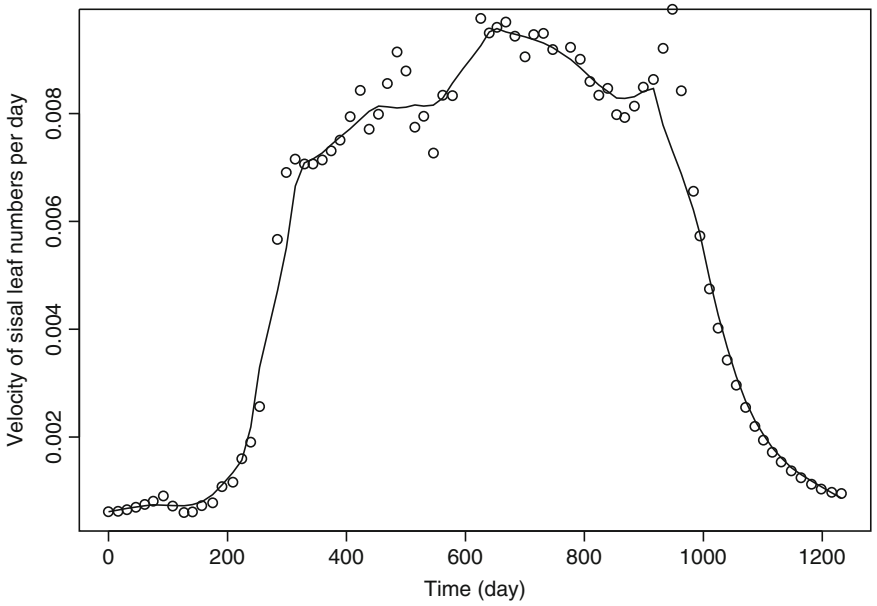


Fig. 5 Figure 5 shows the derivative of the growth curve for number of leaves with input from Fig. 1; following a technique proposed in Dasgupta (2013) by lowess in SPlus with $f = 1/7$

Growth of sisal is steep in rainy season as seen in the curve. Giridih usually has two spells of rain; the additional one is during September–October of a short duration.

Figure 5 shows the derivative of the growth curve for number of leaves with input from Fig. 1; following a technique proposed in Dasgupta (2013) by lowess smoothing in SPlus with $f = 1/7$. Figure 6 shows the same by smooth.spline, with spar= 0.00001.

Figure 7 shows the derivative of the growth curve for sisal plant height with input from Fig. 3; by lowess smoothing in SPlus with $f = 1/7$. A similar variation in growth velocity is nicely explained in Fig. 8 obtained by smooth.spline, with spar= 0.00001.

Proliferation rate $\frac{d}{dt} \log y = \frac{1}{y} \frac{dy}{dt}$ does not depend on the unit of measurements for y, the proliferation rate is calculated for number of leaves and height of sisal plants by smooth.spline, with spar= 0.00001 in Figs. 9 and 10, respectively. There is a sharp upturn in the beginning in these figures, the rates oscillates and slowly decrease in a step function like manner. Like growth and velocity, the proliferation rates of sisal are also affected by seasonal variation.

Similar oscillation in cell count and proliferation rate of wild type cells is observed while incorporation of Bromodeoxyuridine, an analogue of thymidine, in infected megakaryocytes; see Fig. 4b of Horsley et al. (2008).

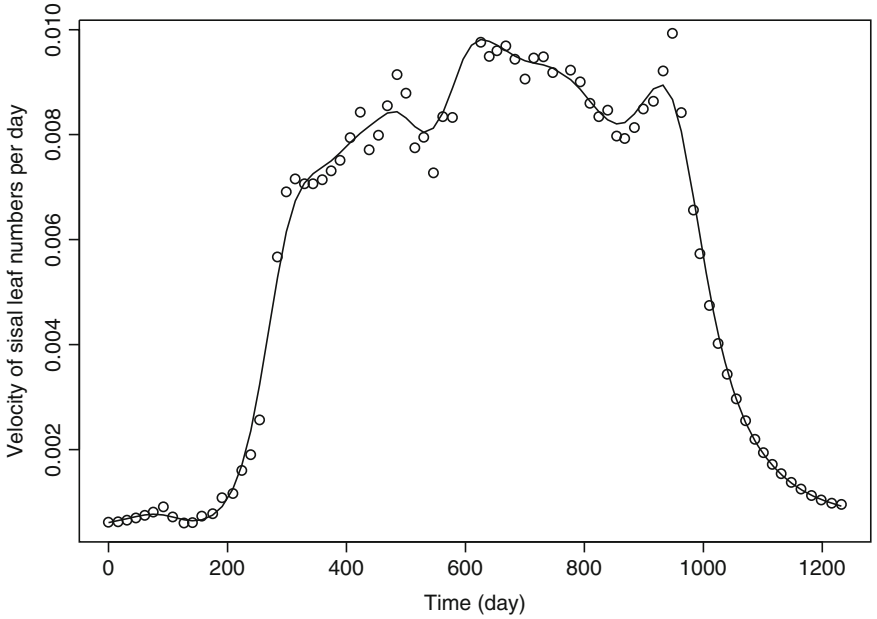


Fig. 6 Figure 6 is smooth spline counterpart of Fig. 5. This smoother curve is obtained by smooth.spline in SPlus, with spar = 0.00001

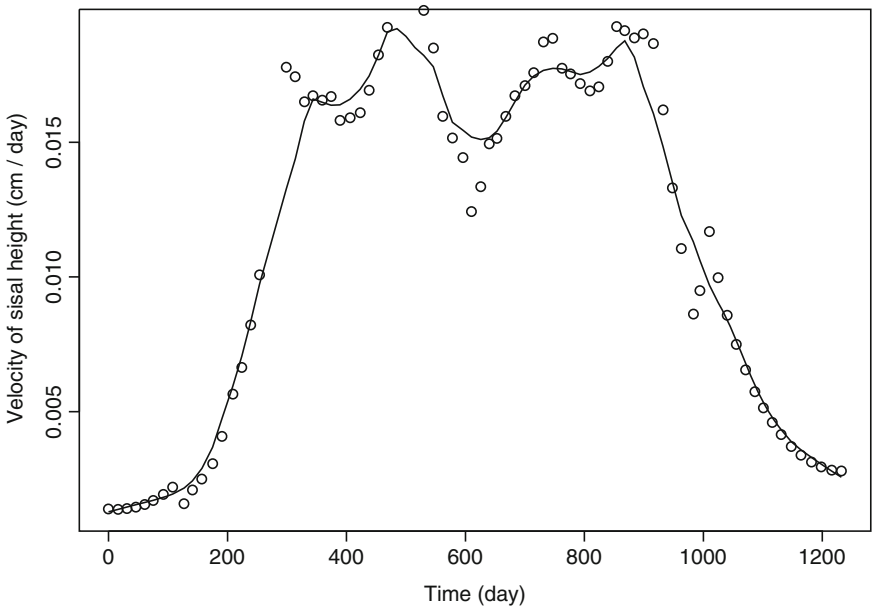


Fig. 7 Velocity of sisal height with trimmed mean, wt. exp(−.01 x); lowess

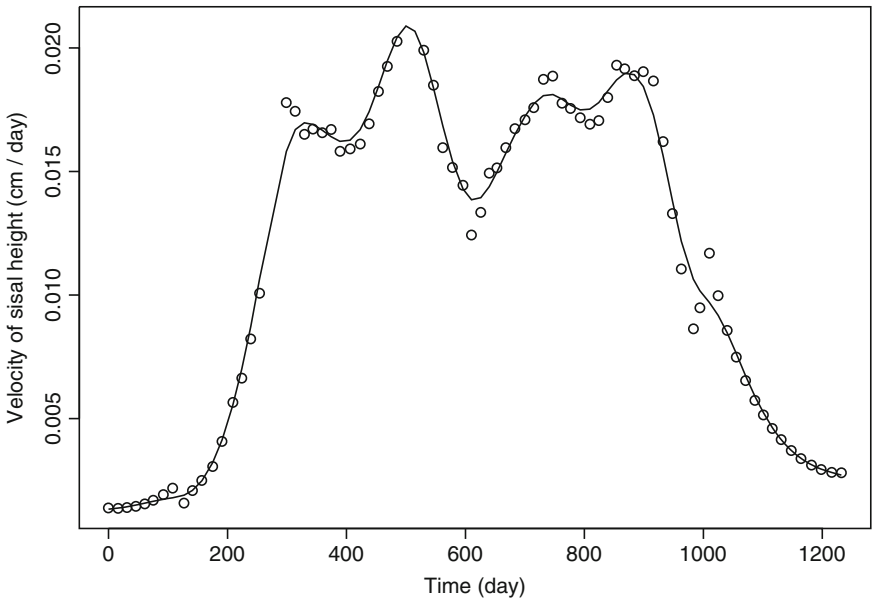


Fig. 8 Velocity of sisal height with trimmed mean, wt. exp(−.01 x); spline

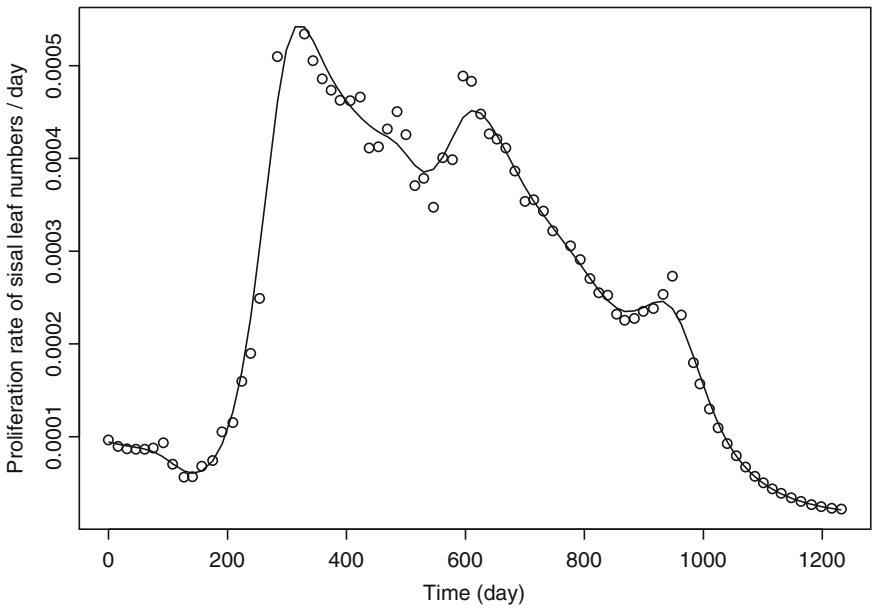


Fig. 9 Proliferation rate of leaf no. with trimmed mean, wt. exp(−.01 x); spline

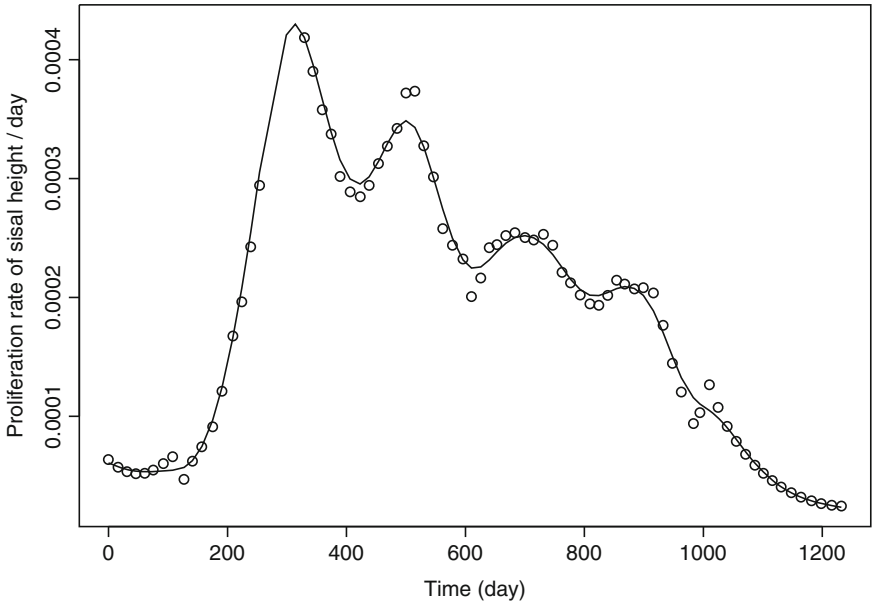


Fig. 10 Proliferation rate of sisal height with trimmed mean, wt. exp(−.01 x); spline

To obtain second derivative of growth curve, we proceed by lowess smoothing with $f = .11$ and compute estimate of the growth curve from averaged points of basic data, and next compute the first derivative by lowess smoothing with $f = .11$. Finally lowess (with $f = .11$) and spline technique (smooth.spline, with spar= 0.00001) are used for smoothing the second derivative point estimates obtained by the technique of Dasgupta (2013). The resultant curves are shown in Figs. 11 (with lowess) and 12 (with spline) for growth of sisal leave number. For sisal plant height, the curve of second derivative is calculated in a similar manner and is shown in Figs. 13 and 14 obtained by lowess and spline method, respectively. Spline technique at final stage produces relatively smooth curves, although at initial stages we preferred to use lowess technique (with $f = .11$) to preserve the main features of data, avoiding much data smoothing.

We collected data on leaf length and leaf width of sisal for a period of 16 months. Five plants in each row were selected at random to collect data. Then five leaves were selected from each plant so marked. As such in each row, there were 25 selected leaves for data collection in a particular date. Average of length and width of leaves were plotted at each time point. To examine the overall growth of the plants, data were collected once in a month within the period (1/4/2011–1/8/2012).

A study on the length and width of sisal leaves at different time period shows that these characteristics increase at a high rate initially and then a downward tendency of basic points is observed for a while when the leaves are mature, and the curve gradually increases to reach a stability. The minor ups and downs of the basic points

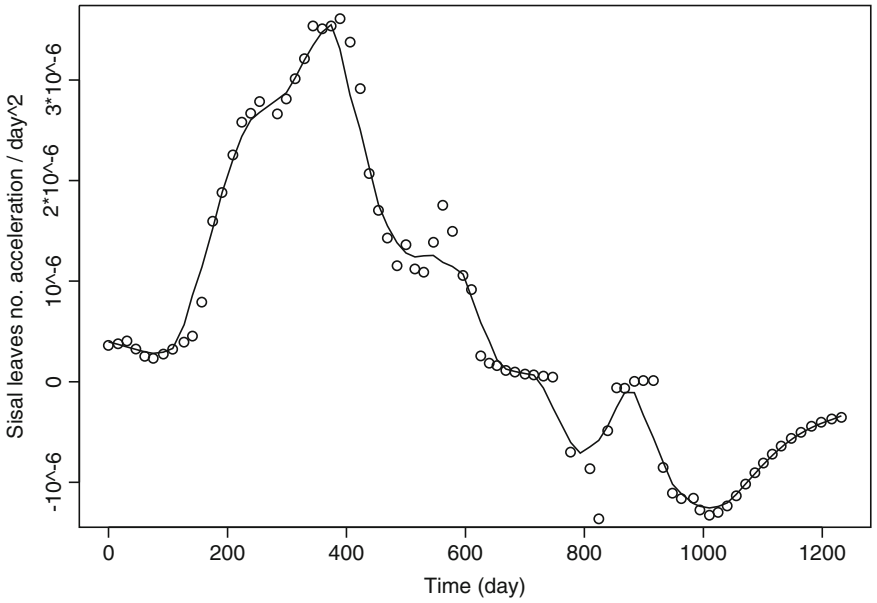


Fig. 11 Sisal leaf no. second derivative with trimmed mean, wt. exp(−.01 x); lowess

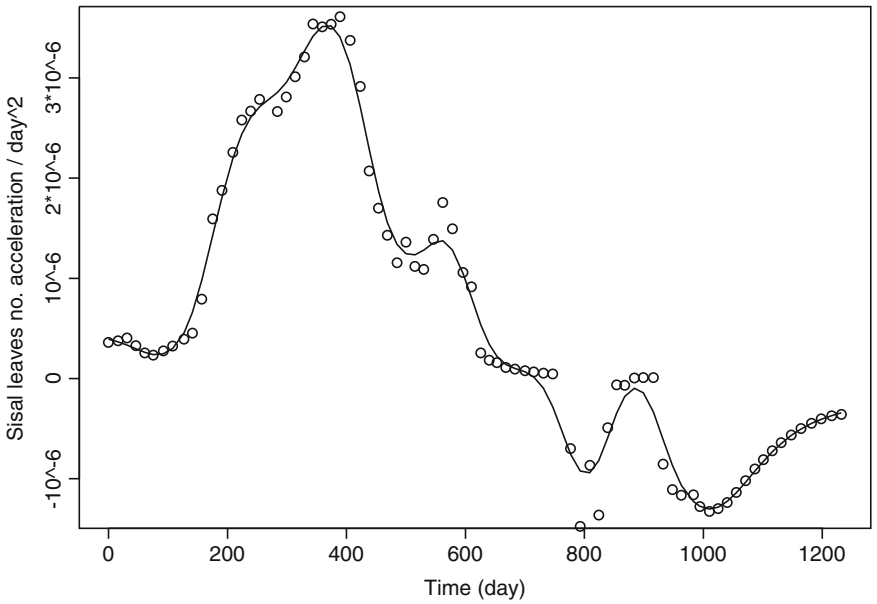


Fig. 12 Sisal leaf no. second derivative with trimmed mean, wt. exp(−.01 x); spline

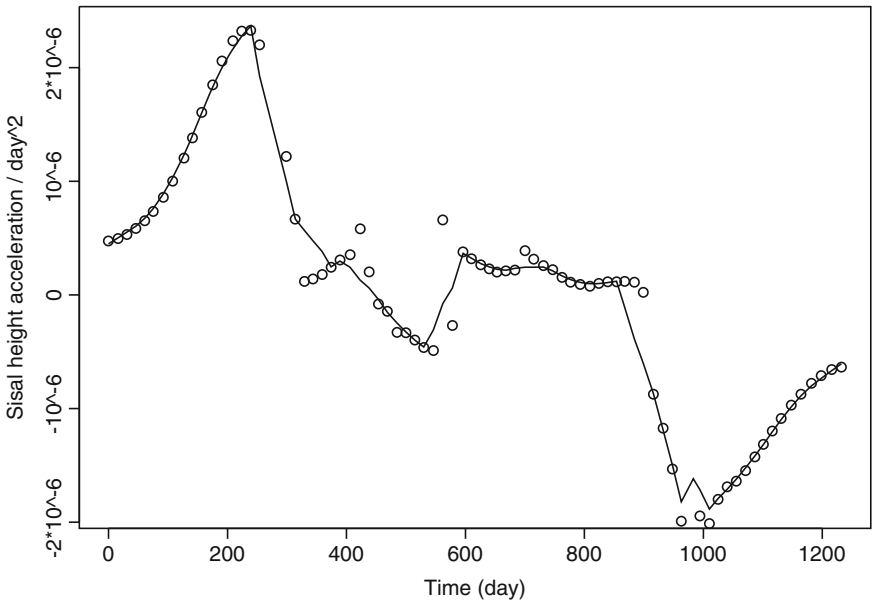


Fig. 13 Sisal leaf no. second derivative with trimmed mean, wt. exp(−.01 x); lowess

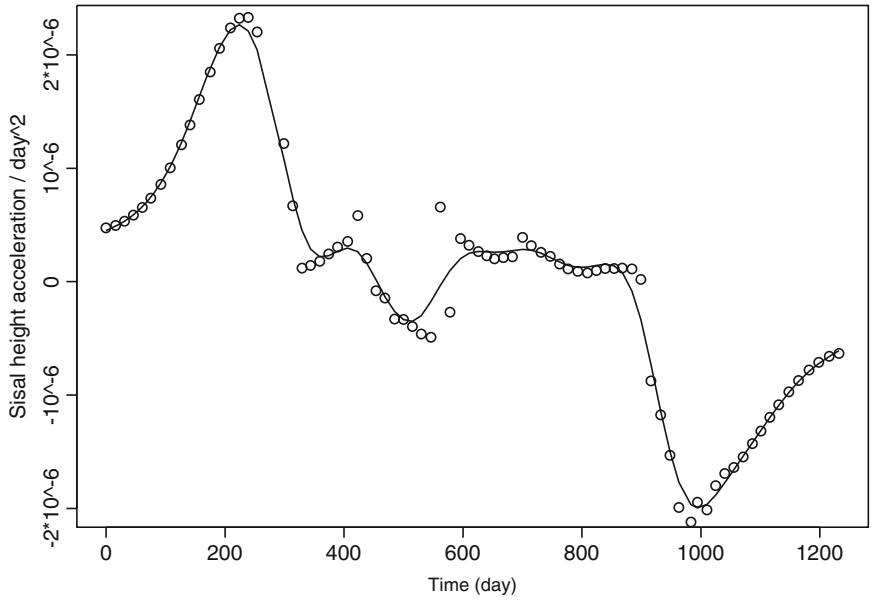


Fig. 14 Sisal height second derivative with trimmed mean, wt. exp(−.01 x); spline

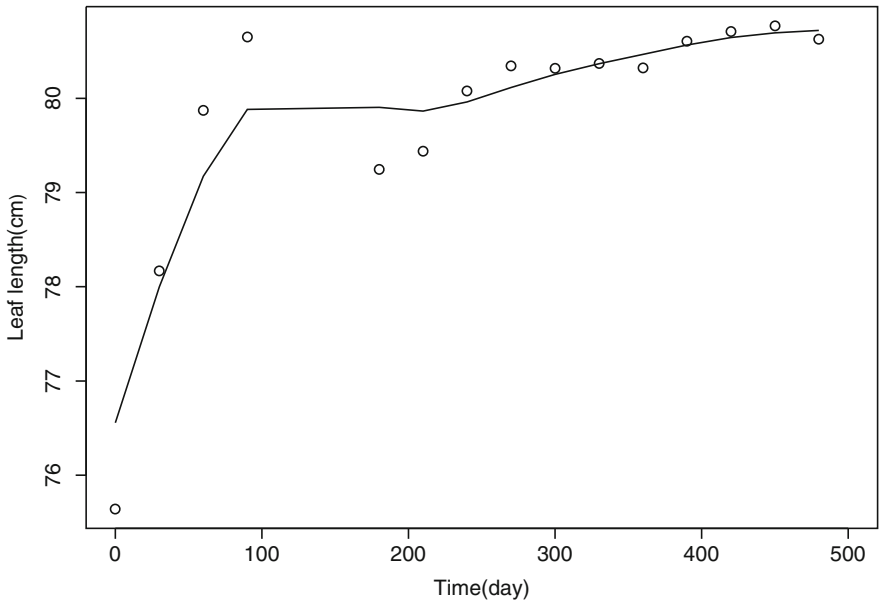


Fig. 15 Growth curve (spline) of sisal leaf length

may be due to seasonal variations, as the growth of leaf decreases during winter season and increases during summer.

Figures 15 and 16 are drawn in SPlus smooth.spline, with spar= .00081, these describe growth curves of leaf length and leaf width, respectively. The length of the sisal leaves is increasing sharply at the initial stage after emerging. Then we notice slight downward trend in the basic points, possibly due to sampling fluctuation. However an overall increasing tendency in growth curve is noticeable in this case also, as was seen for the width of the leaves.

Ten representative plants were selected with systematic sampling starting from plant number one, and the height and numbers of leaves of these plants are recorded over time in Fig. 17 and Fig. 18, respectively. The mean curves are shown in red colour. Almost sure convergence of an estimator to the parameter from a particular direction (upper or lower) is termed as "one-sided estimation". Such conservative type of convergence is of interest when loss due to overestimation and underestimation of the parameter may not be the same. For discussion on one sided estimation and resultant almost sure confidence band see Sect. 2 and Sect. 5 of Dasgupta (2015a) and the references given therein. See also Dasgupta (2015b). Based on these ten sampled plants, in Fig. 19 and Fig. 20, we compute the central line of growth curve for plant height and leaf number, respectively, by lowess with $f = 1/5$; and the upper and lower confidence band by lowess with $f = 1/7$ in respective figures. Almost sure bands specify the position of the growth curve of plant height/number of leaves with probability 1, for large sample size.

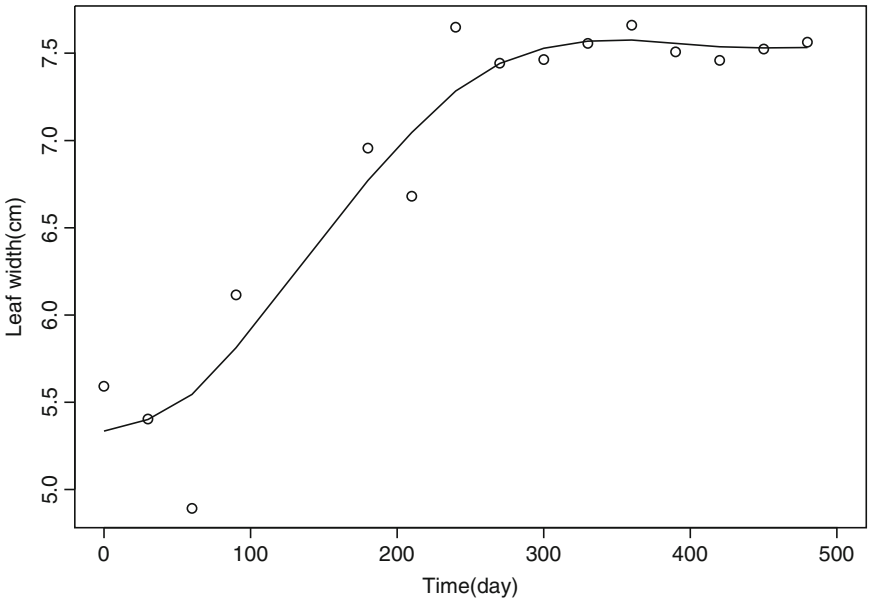


Fig. 16 Growth curve (spline) of sisal leaf width

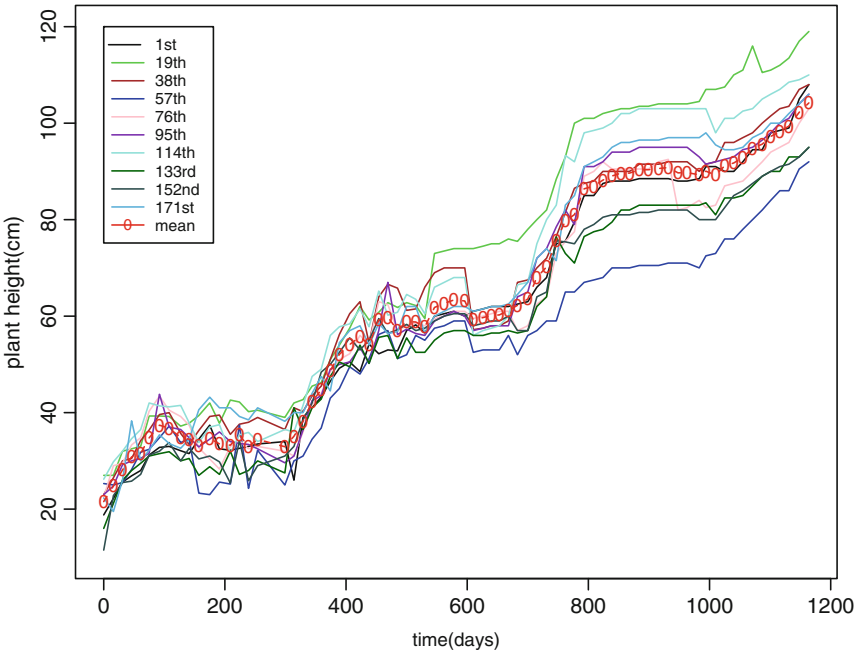


Fig. 17 Sisal plant height growth

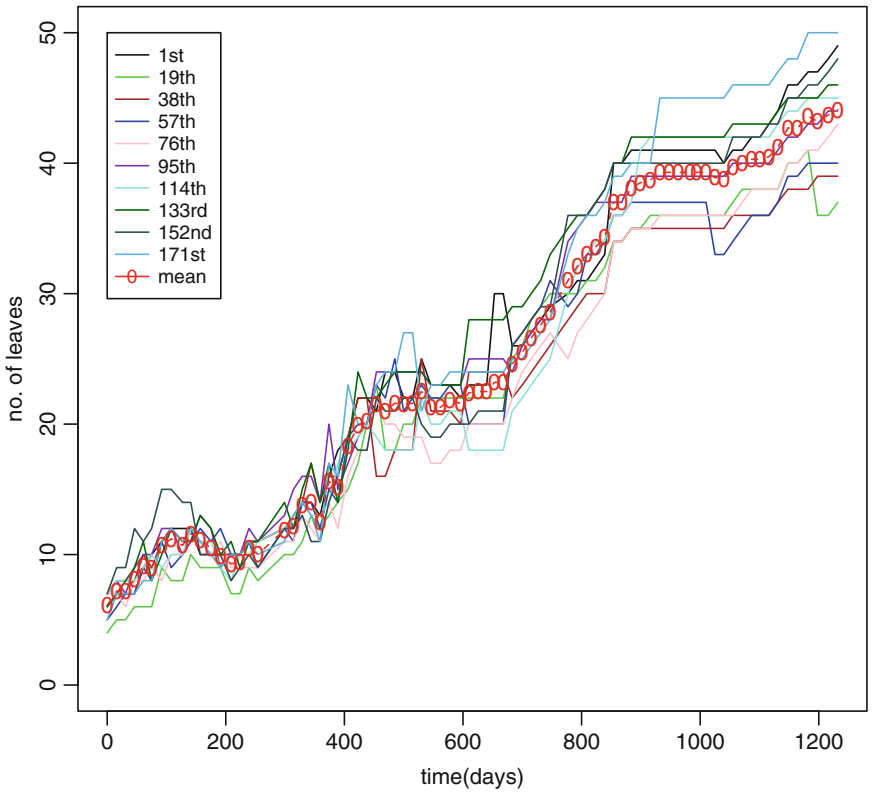


Fig. 18 No. of leaves growth for Sisal plant

The a.s. bands are of stronger conclusion compared to conventional confidence bands of probabilistic coverage. As almost sure band for the variance curve is computed in Fig. 21, with the U statistic kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$; the perturbation part being simplified to $\frac{1}{2n^\alpha}[|(x_{(1)} - x_{(2)})^2/2 - (x_{(2)} - x_{(3)})^2/2| + |(x_{(3)} - x_{(1)})^2/2 - (x_{(1)} - x_{(2)})^2/2| + |(x_{(2)} - x_{(3)})^2/2 - (x_{(3)} - x_{(1)})^2/2|]$, with $n = 3$, $\alpha = 2.25$, where $x_{(1)}, x_{(2)}, x_{(3)}$ are the smallest, median and largest order statistic of ten plant observations at a particular time point; choice of f in lowess regression in Fig. 21 are $f = 2/3$ for the central growth curve, and $f = 1/3$ for upper and lower curves in the band. Position of basic data points is also shown in Fig. 21. A remarkable feature observed is that the ten sampled plants visually show, as apparent from the above figures; almost similar longitudinal patterns present in the larger group of 180 plants, from which these were chosen by systematic sampling.

Data analysed is partly from ISI project 'Integrated nutrient management for sisal cultivation in laterite soil of Giridih, a subtropical plateau region of India'.

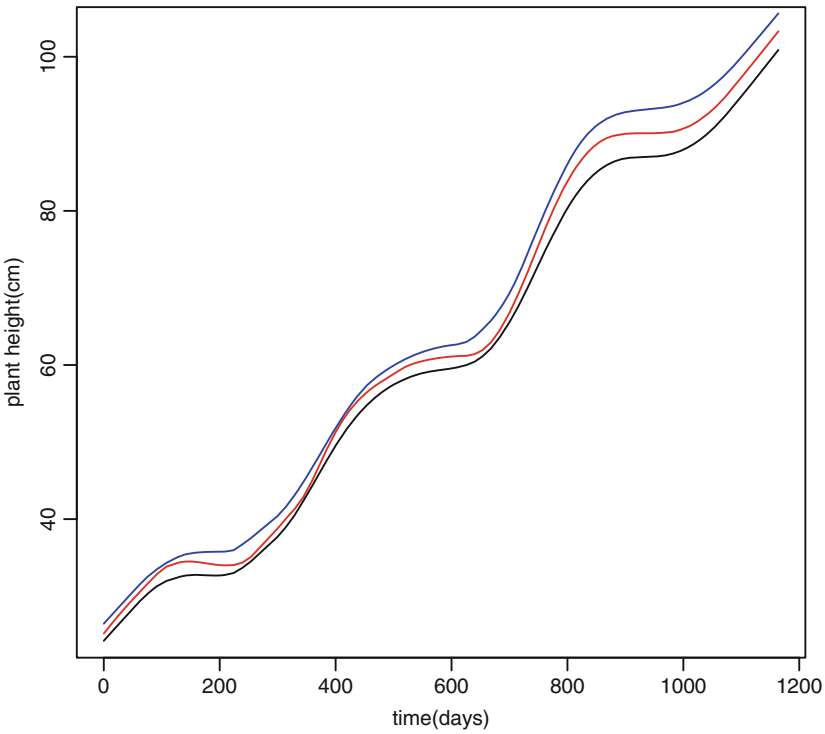


Fig. 19 Almost sure band for growth curve of Sisal plant height

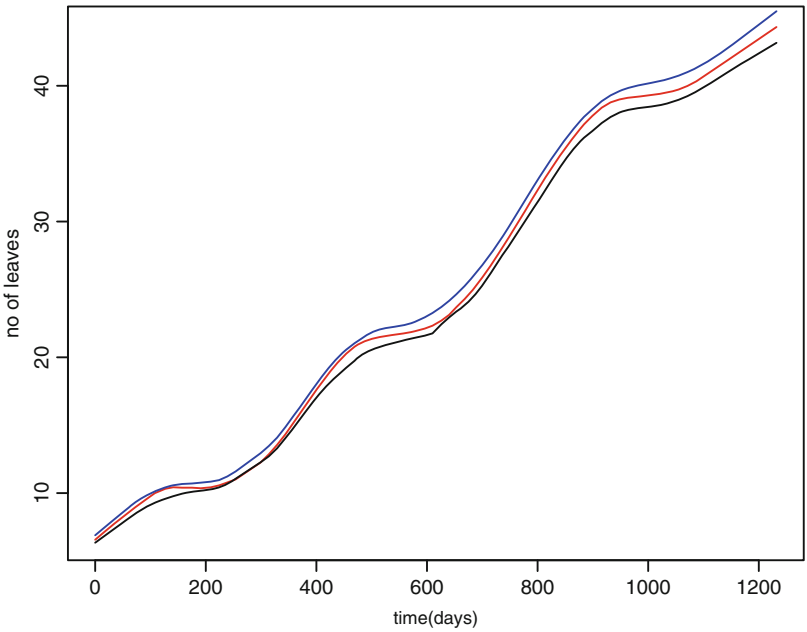


Fig. 20 Almost sure band for growth curve of leaf numbers in Sisal plant

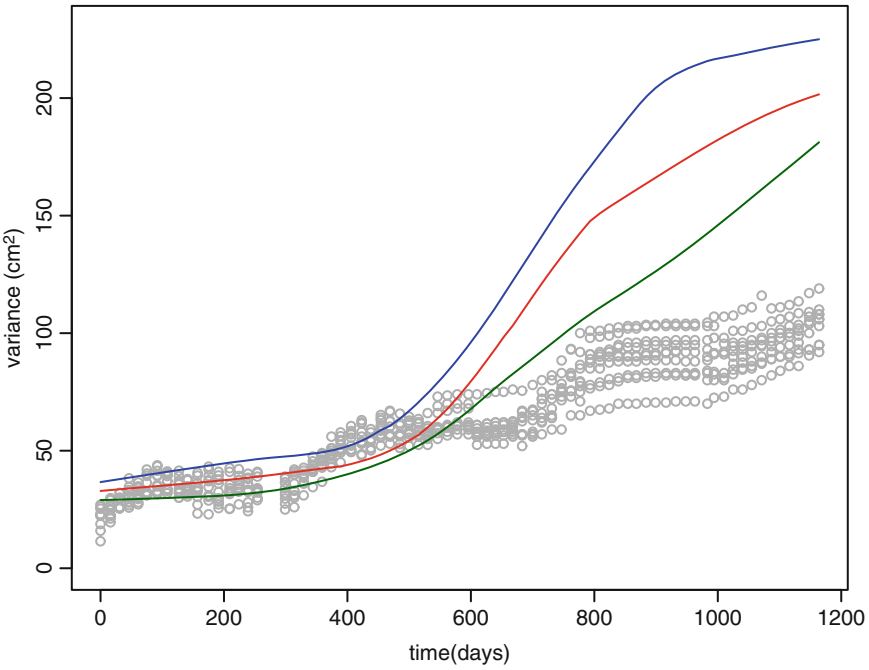


Fig. 21 Red curve in the band represent the central line of variance as time progresses. Basic data points of ten sisal plant height (in cm.) are also shown side by side to understand fluctuation of variance

In studies on number of leaves and height of the sisal plants, we observed a step function like increase in growth curve. This feature is due to seasonal variations. Growth of sisal is high in rainy season compared to other seasons, especially in winter the growth is retarded. The study indicates the land near rivulet to be more conducive for growth of sisal plants having economic potential, although the plant can adapt to harsh environment in Jharkhand.

References

Dasgupta R (1984) On large deviation probabilities of U-statistics in non iid case. Sankhyā 46: 110–116

Dasgupta R (1992) Rates of convergence to normality for some variables with entire characteristic function. Sankhyā A 54:198–214

Dasgupta R (2006) Nonuniform rates of convergence to normality. Sankhyā 68:620–635

Dasgupta R (2008) Convergence rates of two sample U-statistics in non iid case. CSA Bull 60: 81–97

Dasgupta R (2013) Non uniform rates of convergence to normality for two sample U-statistics in non iid case with applications, Chap 4. In: Advances in growth curve models: topics from the

Indian Statistical Institute. Springer Proceedings in Mathematics & Statistics, vol 46. Springer, New York, pp 61–88
Dasgupta R (2015a) Growth curve of elephant foot yam, one sided estimation and confidence band, Chap 5. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York
Dasgupta R (2015b) Growth of tuber crops and almost sure band for quantiles. Commun Stat Simul Comput. doi:10.1080/03610918.2014.990097
Gentry HS (1982) Agaves of Continental North America. University of Arizona press, Tucson
Ghosh M, Dasgupta R (1982) Berry–Esseen theorem for U-statistics in non iid case. In: Colloquia Mathematica Societatis Janos Bolyai, 32. Non parametric statistical inference, Hungery, vol. 1, North Holland, Amsterdam, pp 293–313
Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Ann Math Stat 19:293–325
Horsley V, Aliprantis AO, Polak L, Glimcher LH, Fuchs1 E (2008) NFATc1 Balances quiescence and proliferation of skin stem cells. Cell, 132:299–310
Inacio WP, Lopes FPD, Monteiro SN (2010) Diameter dependence of tensile strength by Weibull analysis: Part III sisal fiber. Matéria (Rio J) 15(2):124–130
Lock GW (1969) Sisal, 2nd edn. Longmans, Green and Co., London

Estimation of Animal Abundance Through Imperfectly Characterising Signatures

Debasis Sengupta

Abstract The problem of estimating population total of animals from imperfectly characterising animal signs poses a number of interesting statistical questions that are not addressed through conventional methods of estimating animal abundance. A case in point is the estimation of tiger population total from pugmark (footprint) measurements, which has been the traditional mode of tiger census in India for several decades. Usual methods based on distance sampling would not work well because, unlike dung produced by elephants or nests produced by birds, such signs are not produced at a steady rate. On the other hand, these signs may not carry as accurate and reliable characterising information as one expects from fingerprints. Is it still possible to estimate the population total precisely and accurately? If so, what should be the appropriate number of signs to be sampled? How can one cluster the signs so that each group of signs belongs to a distinct animal? Is good clustering a prerequisite for good estimation of population total? Is it possible to account for animals missed in the sample? In this article, we attempt to answer to these questions.

Keywords Population total • Clustering • Partially supervised learning • Training data • Data association • Maximum likelihood • Tiger census

1 Introduction

The methods of mark-recapture and distance sampling have been traditionally used for estimating animal abundance in a closed region (see Brochers et al. 2002 for a survey of methods). Sometimes it is difficult to use such methods of direct counting because of non-uniform distribution of animals, mobile populations or logistical constraints. In such cases, one can employ indirect methods of counting through animal signs. For example, elephants are counted through dung-piles, and birds through their nests (see, e.g., Liang et al. 2003). An estimate of the abundance of

D. Sengupta (✉)
Indian Statistical Institute, Kolkata, India
e-mail: sdebasis@isical.ac.in

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_4

animal signs can lead to an estimate of the abundance of the animal itself, if the rates of production and decay of these signs can be estimated accurately.

In India, pugmarks (footprints) of tigers have been used for estimation of abundance of tigers in reserve forests for over seven decades (see Champion 1929; Sharma et al. 2005). These signs are not produced by the tiger at a uniform rate. Therefore, estimation of abundance would depend crucially on the linkage between a pugmark and the tiger producing it. While there has been some statistical work towards establishing this linkage in the case of other cat families (see, e.g., Smallwood and Fitzhugh 1993, for work on mountain lions), there have been lack of similar work in the case of tigers. Gore et al. (1993), who showed that one can discriminate between male and female tigers through pugmark measurements, emphasised the need for quantification of the linkage between pugmarks and individual tigers for successful estimation of population total through pugmark data. In a scathing criticism of the Indian government's conduct of tiger census on the basis of pugmarks despite lack of proof of linkage, Karanth et al. (2003a,b) showed through a controlled experiment that one cannot rely on "expert opinion" to establish this linkage. The controversy received wider attention in January 2005 with publication of media reports (see, e.g., Mazoomdaar 2005) that the Sariska National Park, home to seventeen tigers according to the 2004 "Tiger Census," did not have any tiger at all—a fact that was subsequently acknowledged by the Government of India. Sharma et al. (2005) showed, on the basis of samples collected from tigers in reserve forests and zoological parks, that pugmarks can indeed be linked to individual tigers with reasonable degree of accuracy, especially when one has a set of replicated pugmarks of medium depth, collected from each sample trail. However, there has been no study to determine whether this level of accuracy is adequate for estimation of the population total. Indeed, no validated statistical method for this estimation exists in the literature.

Other methods that have been either used or proposed in connection with estimation of tiger population total include camera trap and genetic matching of bio-materials such as dung (see, for example, Karanth et al. 2004; Tiger Task Force 2005). Matching of images from camera trap (images captured automatically by hidden camera) is far from a foolproof exercise, particularly because different images may be taken from different angles. Genetic matching is also fraught with the risk of contamination and erroneous linkage/non-linkage.

While such sign-based enumeration methods rely heavily on a perfect linkage between animal sign and individual animal, available technology does not offer a perfect linkage. There have been efforts to improve the linkage further. However, there is no existing statistical method of estimating animal population on the basis of a collection of signs, while accepting imperfect linkage as a fact of life. The aim of the present article is to fill this void.

The problem of estimating the population total is as follows. The data consists of measurement vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ extracted from animal signs (e.g., pugmark image, camera-trap image, genetic features, etc.). If there are K animals represented in the sample, then there exists a "true" partition $P_1, \ldots, P_K$ of the index set $\{1, 2, \ldots, N\}$ (with $\cup_{i=1}^{K} P_i = \{1, 2, \ldots, K\}$ and $P_i \cap P_j = \phi$ for $i, j = 1, \ldots, K$,

$i \neq j$), such that the origin of the set of measurement vectors $\{\boldsymbol{y}_j\}_{j \in P_i}$ can be attributed to the ith individual animal. This partition is unknown however, and the task is to estimate the number K.

Note that the correct number K would be known if the correct partition (or cluster) is known. However, there can be many wrong partitions leading to the correct estimate of K, and many more wrong partitions leading to nearly correct estimate of K. Thus, estimation of K is relatively an easier task than identifying the correct partition.

While there is a "true" K in the present problem, the number of clusters is only a notion in the vast majority of clustering problems. The estimation of the "number" of clusters in these problems is akin to the selection of model order in parametric inference problems. This problem has been addressed by several researchers. Some of the early methods were summarised by Milligan and Cooper (1985). Kaufman and Rousseeuw (1990) presented a number of new methods, including partitioning around medoids (PAM), Divisive Analysis (DiAna) and the Silhouette method. Fraley and Raftery (1998, 2002) sought to estimate the number of clusters through model-based cluster analysis. Tibshirani et al. (2001) proposed the Gap statistic for this purpose. McLachlan et al. (2002) provided another solution through a mixture model-based approach to clustering.

Most of these methods attempt to ensure homogeneity (i.e., small variation) among units within a cluster and heterogeneity (i.e., large variation) among units belonging to different clusters. If there is no "true" number of clusters, then it is impossible to calibrate the method of determination of the number of clusters. If a method is applied without calibration to a clustering problem where there is a "true" number of clusters, there is no guarantee that the chosen objective function would have an optimum at or around the correct number of clusters.

An attractive feature of the present problem is that it may be possible to use a limited amount of training data. (For example, multiple pugmark samples can be collected from trails of distinct tigers.) Unlike in the problem of classification where samples from all candidate populations/groups are generally available, here the training data would consist of observations together with cluster labels in respect of some groups only. These groups may not even be represented in the eventual data that have to be clustered. In this sense, the problem corresponds to partially supervised learning. In the literature of partially supervised learning, there are instances where such "training data" are combined with unlabelled test data to determine the number of clusters (see, e.g., Schliep et al. 2003). However, the methods are usually very specific to the application at hand.

In this article, we propose a model-based method for estimating the number of clusters which makes explicit use of the information obtained from the training data. We leave aside the issue of variable/feature selection, noting that this selection can be made using available methods (see, e.g., Fowlkes et al. 1988) on the basis of the same training data. The proposed method does not require the selected features to uniquely identify the correct clusters, and may work reasonably well even if the clusters are not correctly estimated.

In Sect. 2, we present the model and develop the method for estimating the number of animals represented in the sign sample. We also extend this method to handle estimation of the population total (including animals not represented in the sign sample), under some assumptions on the sampling scheme. We study the performance of the estimation methods through simulation in Sect. 3, and propose bias-corrected confidence intervals through parametric bootstrap in Sect. 4. Analysis of a data set obtained from tiger "census" from the Sunderbans Tiger Reserve in the year 2004 is reported in Sect. 5. We conclude the article with some remarks in Sect. 6.

2 Model and Estimation

We assume a random effects model for the p-variate observation:

$$\boldsymbol{y}_j = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_j, \quad j \in P_i, \ \ i = 1, \ldots, K, \tag{1}$$

where $\{P_1, \ldots, P_K\}$ is a partition of the index set of observations $\{1, \ldots, N\}$ such that each set represents a distinct animal, $\boldsymbol{\mu}_i$ is the effect of the ith animal and the $\boldsymbol{\epsilon}_j$'s represent observation-specific random errors which are samples from a distribution. The effects $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ are samples from another distribution and are independent of the $\boldsymbol{\epsilon}_j$'s.

In this article we adopt a parametric approach and assume that the $\boldsymbol{\epsilon}_j$'s have the p-variate normal distribution $N(\boldsymbol{0}, \boldsymbol{W})$, and that the $\boldsymbol{\mu}_i$'s have the p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{B})$, where $\boldsymbol{W}$ and $\boldsymbol{B}$ are positive definite matrices.

2.1 Estimation of W and B from Training Data

The problem of estimation of the covariance matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ is that of estimating multivariate variance components in the possibly unbalanced one-way classified random effects model (1). Even in the univariate case, closed form unbiased nonnegative estimators of the two components do not exist. In the multivariate case, Calvin (1993) considered restricted maximum likelihood (REML) estimation for this problem, and proposed an iterative solution based on an EM algorithm. This method was extended by Calvin and Dykstra (1995) to handle parametric restrictions or inequality constraints on the variance components. This iterative estimator needs a good initial estimate. The following unbiased and closed-form estimate of $\boldsymbol{W}$ can be used for this purpose.

$$\hat{\boldsymbol{W}} = \left(\sum_{i=1}^{K} n_i - K \right)^{-1} \sum_{i=1}^{K} \sum_{j \in P_i} (\boldsymbol{y}_j - \bar{\boldsymbol{y}}_i)(\boldsymbol{y}_j - \bar{\boldsymbol{y}}_i)^T, \tag{2}$$

where $\bar{\boldsymbol{y}}_i = n_i^{-1} \sum_{j \in P_i} \boldsymbol{y}_j$,

and n_i is the cardinality of the set P_i. A possible initial estimator of $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}} = \left[\sum_{i=1}^{K} n_i - \left(\sum_{i=1}^{K} n_i^2\right) / \left(\sum_{i=1}^{K} n_i\right)\right]^{-1} \sum_{i=1}^{K} n_i (\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^T, \quad (3)$$

where $\bar{\boldsymbol{y}} = \left(\sum_{i=1}^{K} n_i\right)^{-1} \sum_{i=1}^{K} \sum_{j \in P_i} \boldsymbol{y}_j$.

It can be checked that this nonnegative definite estimator has expected value $\boldsymbol{B} + \left[\sum_{i=1}^{K} n_i - \left(\sum_{i=1}^{K} n_i^2\right) / \left(\sum_{i=1}^{K} n_i\right)\right]^{-1} (K-1)\boldsymbol{W}$. The bias can be shown to be bounded from above by $n_a^{-1}\boldsymbol{W}$, where n_a is the average size of all the partitions excluding the largest partition. The bias is small if the eigenvalues of $\boldsymbol{W}$ are much smaller than those of $\boldsymbol{B}$, or if n_a is large. Thus, the explicit and nonnegative definite initial estimates given by (2) and (3) may not be iterated upon when fast computation is needed and $\boldsymbol{B}$ has large eigenvalues.

2.2 Estimation of K from Test Data

For the purpose of estimating K, we proceed with the assumption that the matrices $\boldsymbol{W}$ and $\boldsymbol{B}$ are known, and look for the maximum likelihood estimate (MLE) of K and the partition $\{P_1, \ldots, P_K\}$, together with the nuisance parameter $\boldsymbol{\mu}$. In practice, $\boldsymbol{W}$ and $\boldsymbol{B}$ would be estimated from the training data, as indicated in Sect. 2.1.

If $\boldsymbol{F}\boldsymbol{F}^T$ is a rank factorisation of $\boldsymbol{W}^{-1}$, then one can carry out a spectral analysis of the matrix $\boldsymbol{F}\boldsymbol{B}\boldsymbol{F}^T$ to identify different linear combinations of variables (referred to as "features" in the field of machine learning) which are uncorrelated, have unit variance, and have successively decreasing discriminating capability. A subset of these uncorrelated features can then be used for clustering. Therefore, without loss of generality, we assume in the sequel that $\boldsymbol{W} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix of appropriate order.

The likelihood (with $\boldsymbol{W} = \boldsymbol{I}$) is

$$\int_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K} \prod_{i=1}^{K} \left[(|2\pi \boldsymbol{B}|)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \boldsymbol{B}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu})\right\} \times \prod_{j \in P_i} \left[(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T (\boldsymbol{y}_j - \boldsymbol{\mu}_i)\right\} \right] \right] d\boldsymbol{\mu}_1 \cdots d\boldsymbol{\mu}_K.$$

After explicit integration, the likelihood function simplifies to

$$(2\pi)^{-N/2}\prod_{i=1}^{K}|(\boldsymbol{I}+n_i\boldsymbol{B})|^{-1/2}\exp\Bigg[-\frac{1}{2}(\bar{\boldsymbol{y}}_i-\boldsymbol{\mu})^T(\boldsymbol{B}+n_i^{-1}\boldsymbol{I})^{-1}(\bar{\boldsymbol{y}}_i-\boldsymbol{\mu}) \\ -\frac{1}{2}\sum_{j\in P_i}(\boldsymbol{y}_j-\bar{\boldsymbol{y}}_i)^T(\boldsymbol{y}_j-\bar{\boldsymbol{y}}_i)\Bigg], \quad (4)$$

where n_i is the cardinality of the set P_i and $\bar{\boldsymbol{y}}_i = n_i^{-1}\sum_{j\in P_i}\boldsymbol{y}_j$.

The above likelihood has to be maximised with respect to the nuisance parameter $\boldsymbol{\mu}$, the partition $\{P_1,\ldots,P_K\}$ and the population total K. The first maximisation can be done explicitly. It can be shown that, for given K and given partitions $\{P_1,\ldots,P_K\}$, the likelihood (4) is maximised when

$$\boldsymbol{\mu}=\left(\sum_{i=1}^{K}(\boldsymbol{I}+n_i^{-1}\boldsymbol{B}^{-1})^{-1}\right)^{-1}\left(\sum_{i=1}^{K}(\boldsymbol{I}+n_i^{-1}\boldsymbol{B}^{-1})^{-1}\bar{\boldsymbol{y}}_i\right). \quad (5)$$

Substitution of (5) in (4) yields a likelihood which needs to be maximised with respect to the partition $\{P_1,\ldots,P_K\}$ (for fixed K), and subsequently, with respect to K.

Maximisation over all partitions of size K is computationally a very challenging task. There are K^{N-K} ways of partitioning a set of N objects into K clusters such that there is at least one unit in each cluster. Searching over all these clusters is a prohibitive task, even when N is only moderately large. As a computational shortcut, we can evaluate the likelihood (4) and (5) at the partition obtained from the k-means clustering algorithm with average linkage (instead of evaluating it at the maximum likelihood partition), and then maximise the resulting function with respect to K. We denote the maximising value of K by $\hat{K}$.

Note that when the partition itself is regarded as a parameter to be estimated, the corresponding parameter space is not compact, and hence the usual properties of the maximum likelihood estimator need not hold. It may also be argued that, when the main parameter of interest is K, the partition may be regarded as a nuisance parameter and one might work with a weighted sum of the likelihood with respect to a suitable distribution of the partition. However, we choose to estimate the partition (together with K) instead of working with a weighted sum of the likelihood. The reason for this choice is that evaluating the weighted sum of the likelihood involves computing the likelihood at a huge number of candidate partitions, which is a daunting task. Even if one seeks to approximate the weighted average by evaluating the likelihood at a set of sampled partitions, the approximation may not be good if the likelihood is very small (compared to its maximum value) at most of the candidate partitions.

2.3 Estimating Population Total

The method described in the foregoing section is meant for estimating the number of distinct animals from which a particular set of signs has originated. When the objective is to estimate the total number of animals in a closed region, it is necessary to extrapolate this estimate to cover animals that may have been missed in the sample.

Suppose that there is a total of K_0 animals in a closed region (that is, no animal moves into or out of it), and a sample of N signs ($N > K_0$) is collected in such a way that each sign has equal chance of coming from every animal. (The latter assumption is rather strong, but it may be quite reasonable for highly mobile animals in a not-too-large region.) It can be shown that the probability that exactly K animals are represented in the sample is

$$\binom{K_0}{K}(K/K_0)^N \sum_{i=0}^{K}(-1)^i \binom{K}{i}(1-i/K)^N,$$

and the expected number of animals represented in the sample is

$$n(N, K_0) = \sum_{K=1}^{K_0} K\binom{K_0}{K}(K/K_0)^N \sum_{i=0}^{K}(-1)^i \binom{K}{i}(1-i/K)^N. \tag{6}$$

The standard deviation of the number of animals in the sample can also be calculated from the above distribution.

Table 1 depicts the expected value $n(N, K_0)$ of the number of animals represented in a sign sample of size N, together with the standard deviation, for different combination of values of N and K_0. It is clear from this table that $N = 10K_0$ just about ensures inclusion of all animals in the sample. The expected number of animals represented is consistently more than 95 % of K_0 if the sign sample size is $3K_0$, and consistently more than 99 % of K_0, if the sign sample size is $5K_0$.

If the sign sample size is a few times larger than the animal population total, then one can use the relation (6) to obtain an estimator of the population total. Specifically, we propose to estimate the population total by

$$\hat{K}_0 = \frac{\hat{K}}{n(N, \hat{K})}/\hat{K},$$

where $\hat{K}$ is the estimated number of animals represented in the sign sample, to be computed in the manner described in the foregoing section.

In order to avoid confusion, we shall refer to $\hat{K}$ (defined in the previous section) as an estimate of the *number of animals in sign sample*, and to $\hat{K}_0$ as an estimate of the *animal population total* in a region.

Table 1 Expected value and standard deviation of the number of animals represented in a randomly drawn set of N signs collected from an area containing K_0 animals

True number of animals (K_0)	Sample size of animal signs (N)			
	$2K_0$	$3K_0$	$5K_0$	$10K_0$
5	4.463 (0.608)	4.824 (0.393)	4.981 (0.136)	5.000 (0.008)
10	8.784 (0.881)	9.576 (0.596)	9.948 (0.224)	10.00 (0.016)
20	17.43 (1.257)	19.08 (0.869)	19.88 (0.338)	20.00 (0.026)
50	43.37 (1.998)	47.59 (1.397)	49.68 (0.555)	50.00 (0.045)
100	86.60 (2.831)	95.10 (1.986)	99.34 (0.795)	100.0 (0.066)
150	129.8 (3.940)	142.6 (2.437)	149.0 (0.977)	150.0 (0.081)
200	172.9 (5.132)	190.1 (2.817)	198.7 (1.131)	200.0 (0.094)

2.4 Confidence Intervals

A theoretical study of the properties of the estimates of the number of animals represented in the sign sample ($\hat{K}$) and the population total ($\hat{K}_0$) is difficult. However, confidence intervals together with bias correction can be obtained through parametric bootstrap.

In order to carry out this bootstrap for $\hat{K}$, one can fix a trial value of K and simulate a "training data set" and a "test data set" (independent of one another) from the model (1) with the values of $\boldsymbol{B}$, $\boldsymbol{W}$ and $\boldsymbol{\mu}$ same as the corresponding estimates from the actual training data. The sizes of these two simulated data sets should match those of the actual training and test data sets, respectively. The number of distinct "animals" represented in the simulated training data should also match that of the actual training data. For the simulated test data set, one has to ensure that the number of "animals" represented in the test data is exactly equal to the fixed value of K.

The simulated training data set can be analysed to generate simulated estimates of $\boldsymbol{B}$ and $\boldsymbol{W}$. Principal components analysis of $\boldsymbol{F}\boldsymbol{B}\boldsymbol{F}^T$ (where $\boldsymbol{F}\boldsymbol{F}^T$ is a rank factorisation of the estimate of $\boldsymbol{W}^{-1}$) and selection of features can be made exactly as has been done in the case of the actual training data, to generate the chosen number of "normalised features". The number of features would be exactly the same as that used for the original estimation, but the features are specific to the current simulation run. The simulated test data can then be analysed on the basis of the simulated features. The likelihood described in (4), with $\boldsymbol{\mu}$ as in (5) and $\boldsymbol{B}$ as the estimated covariance matrix of the simulated features, can be maximised to generate a simulated estimate of $\hat{K}$.

For the chosen value of K and a fixed confidence coefficient α, the above simulation has to be repeated over a number of runs. The $\alpha/2$ percentile, median, and $1-\alpha/2$ percentile of the resulting (simulated) $\hat{K}$ may be identified as $l(K)$,

$m(K)$ and $u(K)$, respectively. This entire exercise has to be repeated for different trial values of K. Then, a bootstrap-revised estimate of K is

$$\tilde{K} = m^{-1}(\hat{K}). \tag{7}$$

The median of this estimate is approximately equal to the true value of K. A bootstrap confidence interval of K with approximate coverage probability $1 - \alpha$ is $(u^{-1}(\hat{K}), l^{-1}(\hat{K}))$.

A similar bootstrap procedure for $\hat{K}_0$ can be used. The steps are similar to those indicated above, with the following modifications: (a) while generating the test data, one only has to ensure that the underlying number of animals is fixed at a trial value of K_0, and need not bother about the number of "animals" represented in the test data set; the simulated $\hat{K}$ calculated from this test data has to be used to generate a simulated estimate of $\hat{K}_0$ through (2.3). With these modifications, one can calculate the $\alpha/2$ percentile, median, and $1 - \alpha/2$ percentile of the simulated $\hat{K}_0$ for every fixed K_0, and label them as $l_0(K_0)$, $m_0(K_0)$ and $u_0(K_0)$, respectively. A bootstrap-revised estimate of K_0 (with median approximately equal to the correct K_0) is

$$\tilde{K}_0 = m_0^{-1}(\hat{K}_0). \tag{8}$$

A bootstrap confidence interval of K_0 with approximate coverage probability $1 - \alpha$ is $(u_0^{-1}(\hat{K}_0), l_0^{-1}(\hat{K}_0))$.

2.5 Inclusion of Fixed Covariates

The model (1) can easily be expanded to include effects of fixed covariates such as soil type. After estimation of the parameters of the model through the usual analysis of a mixed effects model, one can proceed with the estimation of population total.

3 Simulation Results

A theoretical study of the properties of the estimates of the number of animals represented in sign sample ($\hat{K}$) and the population total ($\hat{K}_0$) mentioned in the foregoing sections is difficult. We study their behaviour through Monte Carlo simulations.

3.1 Estimation of Number of Animals in Sample

We run simulations with number of dimensions p chosen as 4. We choose the "true" within-animal covariance matrix $\boldsymbol{W}$ as the identity matrix and the between-animal covariance matrix $\boldsymbol{B}$ equal to a diagonal matrix with diagonal elements 500, 100,

50 and 20. We fix the number of signs (N) and the "true" number of animals (K_0) at different combinations of values. Subsequently, we carry out a parametric bootstrap for the estimate $\hat{K}$ as follows.

STEP 1. We simulate a training data set consisting of a total of 40 sign feature vectors from the sign feature model (1), such that each sign can originate from all animals with equal probability and there are exactly 20 animals represented in the sample.

STEP 2. We estimate $\boldsymbol{W}$ and $\boldsymbol{B}$ directly from (2) and (3), respectively, for simplicity of computation (as the eigenvalues of $\boldsymbol{B}$ are much larger than those of $\boldsymbol{W}$).

STEP 3. We simulate a test data set (independent from the training data) with N sign feature vectors from the sign feature model (1), such that each sign can originate from all animals with equal probability and there are exactly $K = K_0$ animals represented in the sample (different from those represented in the training data).

STEP 4. Using the estimators of $\boldsymbol{W}$ and $\boldsymbol{B}$ calculated in step 2, we estimate K from the test data set.

STEP 5. We repeat steps 1–4 a number of times, and store the resulting collection of simulated estimates of K. These simulated estimates indicate the pattern of randomness of the actual estimator $\hat{K}$. In particular, the approximate bias (in relation to the true value of K) and variance of the estimator can be calculated.

We repeat this exercise for different combinations of N and K. The findings are summarised in Table 2. Each number reported in this table is based on 100 identical runs of the simulation.

We observe from Table 2 that the estimate of K is generally good when N is three to five times as large as K. Very large values of N can lead to gross over-estimation.

3.2 Accuracy of Clusters

We use the adjusted Rand Index which was proposed by Hubert and Arabie (1985), for measuring the nearness of the true and estimated clusters. If $\{P_1, \ldots, P_L\}$ and $\{Q_1, \ldots, Q_M\}$ are the two clusters being compared, n_{ij}, $n_{i\cdot}$ and $n_{\cdot j}$ are the number of elements in the sets $P_i \cap Q_j$, P_i and Q_j, respectively, $i = 1, \ldots, L$, $j, \ldots, M$, and

Table 2 Mean and standard deviation of estimated number of animals in sign sample ($\hat{K}$) for various combinations of number of signs (N) and "true" number of animals represented in sign sample (K)

True value of K	Mean (and standard deviation) of estimated K when N is				
	$2K$	$3K$	$5K$	$10K$	$20K$
5	5.33 (0.47)	5.52 (0.50)	5.81 (0.44)	6.23 (0.53)	7.79 (1.63)
15	15.26 (0.63)	15.64 (0.59)	16.13 (0.56)	16.89 (1.41)	21.86 (4.24)
50	47.98 (1.65)	49.58 (1.33)	50.87 (1.15)	54.41 (3.91)	71.33 (12.07)

Table 3 Mean and standard deviation of adjusted Rand index between true cluster (K groups) and estimated cluster ($\hat{K}$ groups) for various combinations of number of signs (N) and "true" number of animals represented in sign sample (K)

True value of K	Mean (and standard deviation) of adjusted Rand index when N is				
	$2K$	$3K$	$5K$	$10K$	$20K$
5	0.943 (0.091)	0.924 (0.096)	0.931 (0.060)	0.906 (0.055)	0.804 (0.119)
15	0.965 (0.046)	0.960 (0.040)	0.967 (0.024)	0.957 (0.038)	0.858 (0.089)
50	0.940 (0.037)	0.960 (0.031)	0.966 (0.022)	0.957 (0.033)	0.867 (0.069)

$$n_{..} = \sum_{i=1}^{L} n_{i.} = \sum_{j=1}^{M} n_{.j} = \sum_{i=1}^{L}\sum_{j=1}^{M} n_{ij},$$

then the adjusted rand index for the nearness of these clusters is

$$\frac{\sum_{i=1}^{L}\sum_{j=1}^{M}\binom{n_{ij}}{2} - \sum_{i=1}^{L}\sum_{j=1}^{M}\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}/\binom{n}{2}}{\sum_{i=1}^{L}\binom{n_{i\cdot}}{2}/2 + \sum_{j=1}^{M}\binom{n_{\cdot j}}{2}/2 - \sum_{i=1}^{L}\sum_{j=1}^{M}\binom{n_{i\cdot}}{2}\binom{n_{\cdot j}}{2}/\binom{n}{2}}.$$

This measure is less than one in magnitude, attains the value 1 if and only if the two clusterings are identical and has the average value 0 when the two clusterings are completely independent. This measure had been compared with other measures through simulations, and had been recommended by Milligan and Cooper (1985).

The mean and standard deviation of the adjusted Rand index for the estimated clusters corresponding to the simulated $\hat{K}$'s reported in Table 2 are summarised in Table 3.

The index is generally high when N not more than $5K$.

3.3 Estimation of Animal Population Total

The steps for carrying out simulations for $\hat{K}_0$ are as in the case of the simulations for $\hat{K}$, with two differences: (a) in step 3, there is no need to ensure $K = K_0$ (i.e., some of the K_0 animals may not be represented in the sample), and (b) in steps 4–5, we have to calculate simulated estimates of K_0 to study properties of $\hat{K}_0$, instead of those of $\hat{K}$.

Table 4 shows the mean and standard deviation of the estimator $\hat{K}_0$ for various values of N and K_0, based on 100 simulation runs. A careful comparison with Table 2 would reveal that $\hat{K}_0$ generally has slightly larger standard deviation than $\hat{K}$. This deterioration is the price one has to pay in order to extrapolate from an estimate of K to an estimate of K_0.

As in the case of $\hat{K}$, $\hat{K}_0$ is also observed to have least error when N is three to five times K_0 (after bias and variance have both been taken into account).

Table 4 Mean and standard deviation of estimated animal population total for various combinations of number of signs (N) and "true" population total (K_0)

True value of K_0	Mean (and standard deviation) of estimated K_0 when N is				
	$2K_0$	$3K_0$	$5K_0$	$10K_0$	$20K_0$
5	5.41 (1.22)	5.39 (0.60)	5.84 (0.44)	6.17 (0.47)	7.57 (1.59)
15	15.45 (1.84)	15.72 (1.10)	16.04 (0.85)	17.13 (1.61)	22.45 (4.34)
50	47.10 (2.96)	49.02 (2.39)	51.32 (2.49)	55.03 (4.59)	70.51 (12.40)

4 Bias Correction and Confidence Interval

The observations of the previous section can be used to further improve the estimates of "animals in sample" ($\hat{K}$) and "population total" ($\hat{K}_0$), and to obtain confidence intervals.

The simulations of Sect. 3.1 can be run repeatedly for different trial values of true K, and the $\alpha/2$ percentile, median, and $1-\alpha/2$ percentile of the resulting (simulated) $\hat{K}$ for a specified K may be identified as $l(K)$, $m(K)$ and $u(K)$, respectively. Then, a revised estimate of K is

$$\tilde{K} = m^{-1}(\hat{K}). \tag{9}$$

The median of this estimate is approximately equal to the true value of K. A confidence interval of K with approximate coverage probability $1-\alpha$ is $(u^{-1}(\hat{K}), l^{-1}(\hat{K}))$.

Likewise, if $l_0(K)$, $m_0(K)$ and $u_0(K)$ are the $\alpha/2$ percentile, median, and $1-\alpha/2$ percentile of simulated $\hat{K}_0$ for a specified K_0, then a revised estimate of K_0 with approximately correct median is

$$\tilde{K}_0 = m_0^{-1}(\hat{K}_0), \tag{10}$$

and a confidence interval of K with approximate coverage probability $1-\alpha$ is $(u_0^{-1}(\hat{K}_0), l_0^{-1}(\hat{K}_0))$.

5 Analysis of Data From Sunderbans Tiger Reserve

The context of the present work has been the pugmark-based tiger "census" conducted by the Government of West Bengal in January 2004. The "census" actually consists of extensive collection of left hind pugmarks obtained through a comb search along the banks of all accessible waterways which are exposed during low tide, and should not be confused with a complete enumeration exercise. In order to use the method for estimation of the population total proposed in this paper, one has to have additional data where the pugmarks originating from each distinct tiger

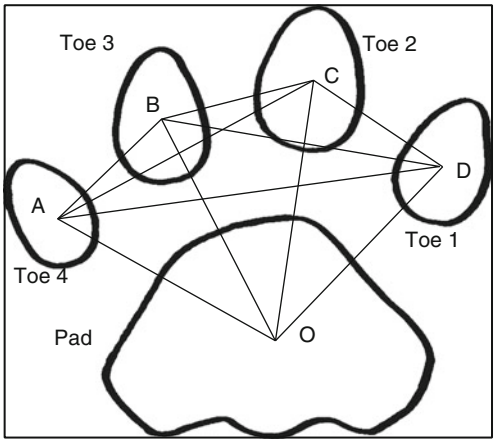


Fig. 1 Typical contour of a tiger pugmark

are clearly labelled. Replicated pugmarks of a particular tiger may be obtained from a single trail, and geographical separation of the trails and a narrow time window for data collection would ensure that there is no realistic chance of two trails having originated from the same tiger. The labelled data for the present exercise consisted of 76 observations, in which 20 distinct trails were represented. These included observations from a few right hind pugmarks that were reflected so that they become similar to left hind pugmarks. Following the convention followed during the tiger "census", these training samples were collected from three strata of soil: clay, loam and sandy. A total of 37 features were extracted from the contour of a tiger pugmark (see Fig. 1) drawn from a replica made of plaster of Paris, obtained from the pugmark impression. All these features are rotation and translation invariant, and are log-transformed so that these take values over the unrestricted real line.

The details of the computations will appear in a forthcoming paper. Here we give only a gist of the findings. At the outset, we eliminated the features that showed heterogeneity across soil types and/or across left and reflected right pugmarks, and also the features that did not discriminate between trails (individuals). This produced a pruned list of 24 features, for which we can use model (1). Five other features were eliminated in order to ensure numerical stability of inversion of the estimated matrix $\hat{\boldsymbol{W}}$. Normalised orthogonal features for clustering were obtained from eigenvectors of the matrix $\hat{\boldsymbol{F}}^T \boldsymbol{B} \hat{\boldsymbol{F}}$, where $\hat{\boldsymbol{F}} \hat{\boldsymbol{F}}^T$ is a rank factorisation of $\hat{\boldsymbol{W}}^{-1}$. Subsequently we used a lower rank approximation by minimising the cross-validation error-rate of the linear discriminant analysis (LDA) classifier for the 20 classes at hand. A rank-8 approximation was obtained in the process. The largest three eigenvalues of the rank-8 matrix accounted for more than 97 % of the trace, and therefore only the corresponding eigenvectors were used as linear combinations of variables for clustering.

The (unlabeled) test data consisted of entire collection of 1,059 pugmarks from 2004 Tiger "census" in the Sunderbans Tiger Reserve, which reduced to a total of 966 after screening for front and right hind pugmarks. Estimation was done as per

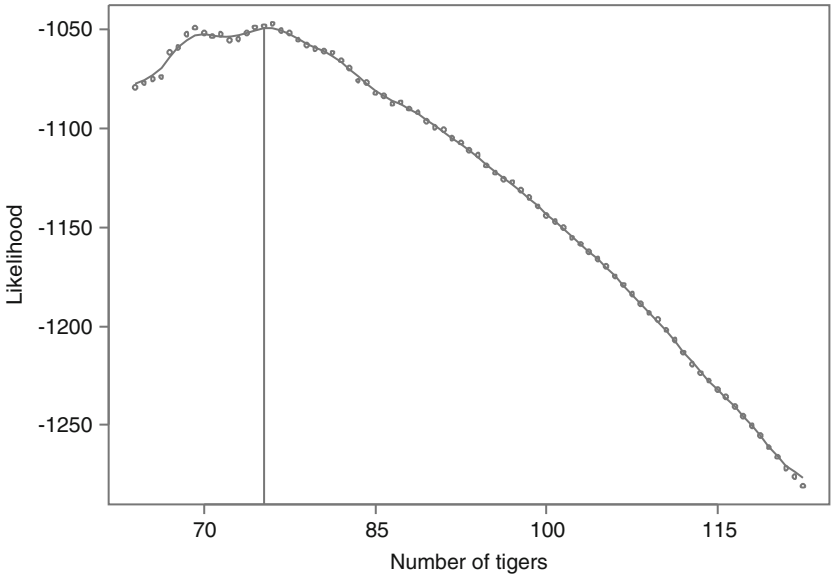


Fig. 2 Typical shape of likelihood function from bootstrap resamples

the method described in Sect. 2 with bootstrap based bias correction described in Sect. 4. The plot of the likelihood calculated from a typical resample is shown in Fig. 2.

The number of tigers represented in sample was estimated as 75, while a bootstrap confidence interval with 95 % coverage was 42–107. The estimated population total was also 75, with bootstrap confidence interval 45–106.

The confidence intervals are somewhat wide. This fact may be attributed to the excessively large size of the data in relation to the estimated population total. Going by the findings of the simulations reported earlier, collection of two to four hundred pugmark samples during the "census" operation would have been ideal.

6 Tracking Population Growth

Estimation of the population total and estimation of the correct cluster of the animal sign sample are related but distinct problems (see discussion of page 61). There may be an accurate solution of one problem but not the other. The quality of the estimator of the population total depends on the discriminating power of the chosen features, i.e., the size of the eigenvalues of the between-animals covariance matrix ($\boldsymbol{B}$) relative to the within-animals covariance matrix ($\boldsymbol{W}$). It appears from the work of Sharma et al. (2005) that, the level of contrast between $\boldsymbol{B}$ and $\boldsymbol{W}$ assumed in the simulations of Sect. 4 may very well be achievable in the case of tiger pugmarks.

Our simulation results indicate that, for the assumed values of these matrices, it should be possible to estimate the population total up to an accuracy of $\pm 10\,\%$, provided the size of the sign sample is reasonable (see last remark of Sect. 3.3).

In view of the fact that the proposed method is meant for estimating the population total, and not for tracking the presence of a particular animal in a sequence of studies, one has to make use of the trajectory of estimated population totals at different points of time in order to track population growth.

One can make use of location information for this purpose, in different ways. First, the method presented here can be used for animal signs obtained from sub-regions. However, the same animal may leave signs in more than one sub-region. Therefore, the sum of the estimated totals in different sub-regions may be more than the estimate obtained for the entire region. Even so, the growth of population in a sub-region may be tracked by using the proposed method on data obtained from that sub-region at different points of time. Second, positional data on the location of each sign can be used as additional features. In order for these data to be valuable for the present purpose, the entire data needs to be collected over a short span of time, and there should be information about the maximum possible geographical separation between two signs originating from the same animal.

Another quantity relevant for population growth is the number of cubs. For the tiger pug-mark data (and similar other examples), it may be possible to use the proposed method to exclusively estimate the cub population total. By tracking this estimate across studies conducted at various times, one may get a clearer idea about the growth of population.

References

Brochers DL, Buckland ST, Zucchini W (2002) Estimating animal abundance: closed populations. Springer, London

Calvin JA (1993) REML estimation in unbalanced multivariate variance components model using an EM algorithm. Biometrics 49:691–701

Calvin JA, Dykstra RL (1995) REML estimation of covariance matrices with restricted parameter spaces. J Am Stat Assoc 90:321–329

Champion FW (1929) Tiger tracks. J Bombay Nat Hist Soc 33:284–287

Fowlkes EB, Gnanadesikan R, Kettering JR (1988) Variable selection in clustering. J Classif 5:205–228

Fraley C, Raftery AE (1998) How many clusters? Which clustering methods? Answers via model-based cluster analysis. Comput J 41:578–588

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631

Gore AP, Paranjape SA, Rajgopala G, Kharshikar AV, Joshi NV, Watve MG, Gogate MG (1993) Tiger census: role of quantification. Curr Sci 64:711–714

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Karanth KU, Nichols JD, Seidenstricker J, Dinerstein E, Smith JLD, McDougal C, Johnsingh AJT, Chundawat RS, Thapar V (2003) Science deficiency in conservation practice: the monitoring of tiger populations in India. Anim Conserv 6:141–146

Karanth KU, Raghunandan S, Chundawat RS, Nichols JD, Samba Kumar N (2003) Estimation of tiger densities in the tropical dry forests of Panna, Central India, using photographic capture-recapture sampling. Anim Conserv 7:285–290

Karanth KU, Nichols JD, Kumar NS (2004) Photographic sampling of elusive mammals in tropical forests, in WL Thompson, ed., Sampling Rare and Elusive Species, Island Press, Covelo, CA, 229–247

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley Series in Probability and Mathematical Statistics. Wiley, New York

Liang SE, Buckland ST, Burn RW, Lambie D, Amphlet A (2003) Dung and nest surveys: estimating decay rates. J Appl Ecol 40:1102–1111

Mazoomdaar J (2005) Have you seen a tiger at Sariska since June? If yes, you're the only one. The Indian Express, 21 January 2005 (http://www.indianexpress.com/res/web/pIe/archive_full_story.php?content_id=632809)

McLachlan GJ, Bean R, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:413–422

Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179

Schliep A, Schönhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19:(Suppl. 1) i255–i263

Sharma S, Jhala Y, Sawarkar VB (2005) Identification of individual tigers (Panthera tigris) from their pugmarks. J Zool 266:1–10

Smallwood KS, Fitzhugh EL (1993) A rigorous technique for identifying individual mountain lions (Felis concolor) by their tracks. Biol Conserv 65:51–59

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of data clusters via the Gap statistic. J R Stat Soc Ser B 63:411–423

Tiger Task Force, Government of India (2005) Joining the dots: the report of the tiger task force (http://projecttiger.nic.in/TTF2005/pdf/full_report.pdf9)

Growth Curve of Elephant Foot Yam, One Sided Estimation and Confidence Band

Ratan Dasgupta

Abstract Almost sure convergence of an estimator to a real valued parameter from a particular side (above/below), termed one sided estimation, is of interest in conservative estimation. Such estimator converges to the parameter from a particular direction almost surely (a.s.) for all large sample size n. We consider one sided estimation of growth curve for Elephant foot yam from experimental data and examine the resultant confidence band of the curve. Almost sure upper and lower bounds for yield over time may be used as conservative estimates of crop yield with probability 1. Deviation probabilities and convergence rates in central limit theorem for proposed estimators are studied under the set-up of U-statistics. Probability bounds of tail events concerning error in approximation are shown to be exponentially decaying. Some special types of L statistics relevant to one sided convergence are also considered for computing rates in CLT. Implications of results are discussed in the context of Yam growth curve estimation with live data. Associated a.s. confidence band for estimated growth curve may be narrow even for a widely dispersed data, as the goal of constructing confidence band here is different from including all data points inside the band. Upper and lower estimate of the growth curve of yam are seen to cover the mean response curve in general. Confidence band for variance of yam yield over time exhibits similar coverage properties. Presence of an upward spike is observed in the growth curve of yam yield towards the end of yam plant lifetime. The spike is prominent when yam is harvested at the end of second season in a 2-year study, instead of harvesting the crop after first year.

Keywords Elephant foot yam • Growth curve • U-statistics • L-statistics • Order statistics • One sided estimates • Berry–Esseen theorem • Moderate deviation • Large deviation

MS Subject classification: Primary 62P10; Secondary 62J02, 62G05, 60F05, 60F10, 60F15, 62G20, 62G30

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_5

1 Introduction

Consider the problem of estimating confidence band of growth curve for a tuber crop like Elephant foot yam. A conservative confidence band may be constructed following the idea of one sided estimation. Let $X_1, X_2, \cdots$ be a sequence of iid random variables and let h be a symmetric kernel of m arguments. Let $\theta = Eh(X_1, \cdots, X_m)$ be the parameter of interest. For example, with $m = 1$ and $h(x) = x$, θ is the population mean; with $m = 2$ and $h(x_1, x_2) = (x_1 - x_2)^2/2$, θ represents the population variance.

There are situations when one needs to estimate the parameter θ either from above or from below. That is, for an estimator θ_n of θ, one may require $\theta_n \to \theta$ almost surely, and $\theta_n \geq \theta$ a.s. for all large n. When this happens we say that θ_n is an upper estimate of θ, and write $\theta_n \to_+ \theta$. Convergence from below may be defined in a similar fashion, and we denote $\theta_n \to_- \theta$. Such estimators may be required when the loss due to over and under estimation are not the same and conservative estimates are preferred. For example, one may like to have an underestimate of the strength of a dam rather than have an overestimate of it. As another example, consider estimation of maximum level of flood water from above to be in safer side.

The problem of estimating the *mean* of a population in such a manner was first considered in Gilat and Hill (1992). An upper estimate was derived by taking the weighted average of the sample observations with more weights to the higher order statistics and lesser weight to the lower order statistics. Subsequently, Bose and Dasgupta (1994) showed that this estimate has an underlying U-statistics structure which may be utilised for estimating $\theta = Eh(X_1, \cdots, X_m)$. We denote by $U_n(g)$ the U-statistic based on the symmetrised version of the kernel g. Consider the function $D(X_1, \cdots, X_{2m}) = |h(X_1, \cdots, X_m) - h(X_{m+1}, \cdots, X_{2m})|$ and define the estimator

$$\theta_n = U_n(h) + a_n U_n(D) \tag{1}$$

where a_n is a suitable sequence of positive constants converging to zero. Indeed a_n is selected in such a way that it overcomes the maximum fluctuation of $U_n(h)$ around θ as specified by LIL, and pulls the estimator to the right of θ. We establish the uniform and nonuniform Berry–Esseen bound for the one sided estimate θ_n in Theorems 2.1 and 2.2, exploiting the U statistics structure of our proposed estimates. These results are analogous to the sample mean given in Katz (1963) and Ghosh and Dasgupta (1978). Similarity of the results indicates that the new estimator inherits the nice properties of sample mean, thus opening up the possibility of real life applications; even after suitable modifications so as to converge from a particular side. Probabilities of deviation results are provided in the second part of Theorems 2.2, Remark 1 and Theorem 2.3. Variance and MSE of the proposed estimates are also computed, see Remark 2.

Assuming that data on plant lifetime versus yam yield recorded in a small time domain to be iid observations, one may estimate the growth from above or below. Large variability in agricultural yield data on yam requires appropriate technique

of estimation and we adopt nonparametric lowess regression for estimating the mean response curve and its confidence band from cross sectional studies. Almost sure band for growth curve is estimated from live yam data arising from several experiments conducted in different types of soil at Indian Statistical Institute (ISI) Giridih farm. Presence of an upward spike in the growth curve of yam yield is generally observed towards the end of plant lifetime. The spike is prominent when yam is harvested at the end of second season in a 2-year study, rather than harvesting the crop at the end of first season.

The large sample properties of modified estimators for one sided convergence are seen to be similar to the original U statistics, since the amount of perturbation converges to zero with increase in sample size.

In Sect. 2 we study convergence rates and deviation probabilities of one sided estimators. Subsequently, based on these estimators, confidence bands are constructed. Section 3 deals with nonuniform CLT rates for general U statistic under moment assumption on the kernel h that ensures m.g.f. of h necessarily exists but h may not be bounded. These sharp results are then extended to the proposed one sided estimators θ_n, justifying its applicability. Some special types of L statistic relevant to one sided convergence are also considered in Sect. 4. In Sect. 5 almost sure nonparametric confidence bands for growth curve are computed with five sets of farm data collected during the production season 2013–2014 on yam. In Sect. 6 we search for parametric models to explain underground yam growth, in view of drying up of quantifiable biomass aboveground and simultaneous fast yam deposition underground towards the end of plant lifetime.

2 Convergence Rates of One Sided Estimator

To state the results recall that $\theta = Eh(X_1, \cdots, X_m)$ and θ_n as in (1).

2.1 Speed of Convergence and Probabilities of Moderate Deviations for θ_n

To compute the rates of convergence of standardised θ_n, we use the representation of θ_n as a U statistic with a kernel that depend on n. This will enable us to use the theory of U statistics to obtain uniform and nonuniform rates of convergence to normality for standardised θ_n.

Extend h to a function of $2m$ arguments by defining

$$h^*(X_1, \cdots, X_{2m}) = [h(X_1, \cdots, X_m) + h(X_{m+1}, \cdots, X_{2m})]/2$$

Observe that for $n \geq 2m$, the estimator θ_n defined in (1) may also be expressed as

$$\theta_n = U_n(h^*) + a_n U_n(D) = U_n(h^* + a_n D) \tag{2}$$

Define

$$\Delta = E|h(X_1, \cdots, X_m) - h(X_{m+1}, \cdots, X_{2m})|$$

and observe that

$$E(\theta_n) = \theta + a_n \Delta$$

Therefore,

$$\begin{aligned} \theta_n - E(\theta_n) &= U_n(h^* + a_n D) - (\theta + a_n \Delta) \\ &= \hat{U}_n(h^* + a_n D) - (\theta + a_n \Delta) + R_n \\ &= \hat{U}_n(h^*) - \theta + a_n(\hat{U}_n(D) - \Delta) + R_n = V_n + R_n \end{aligned} \tag{3}$$

where $\hat{U}_n$ is Hájek's projection and R_n is the corresponding remainder.

V_n may also be written as

$$V_n = \frac{m}{n} \sum_{i=1}^{n} \tilde{h}_1(X_i) + a_n \frac{2m}{n} \sum_{i=1}^{n} \tilde{D}_1(X_i) = \frac{2m}{n} \sum_{i=1}^{n} [\tilde{h}_1(X_i)/2 + a_n \tilde{D}_1(X_i)] \tag{4}$$

where $\tilde{h}_1(x) = Eh(x, X_2, \cdots, X_m) - \theta$ and

$$\tilde{D}_1(x) = E|h(x, X_2, \cdots, X_m) - h(X_{m+1}, \cdots, X_{2m})| - \Delta$$

V_n and R_n are orthogonal and $E(R_n^2) = O(n^{-2})$, if $Eh^2 < \infty$. See, for example, Serfling (1980, page 188).

Under the assumption that $E|h|^{2+c} < \infty$, where $0 < c < \infty$, following the idea of Funk (1970) or Grams and Serfling (1973), one may show that

$$E[n^{1/2} R_n]^{2m_0} = O(n^{-m_0}) \tag{5}$$

for any integer m_0 satisfying $c < 2m_0 \leq c + 2$.

Observe that although the kernel corresponding to θ_n is changing with n, order of moment bound of R_n are not affected by this. This is because the sequence a_n is bounded (in fact a_n is going to zero). Assume that

$$\inf_n \sigma_n^2 > 0 \tag{6}$$

where $\sigma_n^2 = \text{var}[\tilde{h}_1(X_i)/2 + a_n \tilde{D}_1(X_i)]$. To study rates of convergence of the estimator, one needs to consider $T_n^* = \frac{n^{1/2}(\theta_n - \theta)}{2m\sigma_n}$. However, to apply martingale central limit theorems directly, it is convenient to consider $T_n = \frac{n^{1/2}(\theta_n - E(\theta_n))}{2m\sigma_n}$.

Note that $T_n - T_n^* = \frac{n^{1/2}(\theta - E(\theta_n))}{2m\sigma_n} = O(n^{1/2}a_n)$. Thus the results on T_n may easily be converted to a corresponding result on T_n^*.

From Friedrich (1989, Remark 5; p. 175), see also Theorem 1 of Ghosh and Dasgupta (1978) and Lee (1990), it is possible to obtain the following Berry–Esseen theorem:

Theorem 2.1. *Let* $E|\tilde{h}_1(X)|^{2+c} < \infty$, $E|h|^{(4+c)/3} < \infty$, *for some* $c > 0$ *and (6) holds. Then for some constant* $K > 0$,

$$\sup_x |P(T_n \le x) - \Phi(x)| \le Kn^{-\min(c,1)/2}$$

Using the general results of Ghosh and Dasgupta (1978), see pages 363–364, one can also obtain nonuniform Berry–Esseen bound and moderate deviation probabilities of standardised θ_n. In view of (5) above, (3.18) and (3.21) of Ghosh and Dasgupta (1978) hold with $u(x) = 1$. Thus using (3.1) of the above paper and (5), we have the following theorem.

Theorem 2.2. *Let* $E|h|^{2+c} < \infty$ *for some* $c > 0$. *Then*

(i) for some constant $K > 0, g(c) > 0$ *and for all real* x

$$|P(T_n \le x)) - \Phi(x)| \le Kn^{-\min(c,1)/2}(\log n)^{g(c)}(1 + |x|^{2+c})^{-1}$$

(ii) if $x_n \to \infty$ *such that*

$$x_n^2 \le c \log n + 2(c+1) \log |x_n| + M$$

for some $M > 0$, *we have*

$$1 - P(T_n \le x_n)) \sim \Phi(-x_n) \sim P(T_n \le -x_n).$$

Observe that the range of x_n *in (ii) above may also be written as*

$$x_n^2 \le c \log n + 2(c+1) \log \log n + M$$

Remark 1. Theorem 2.2(i) and (ii) provide a nonuniform Berry–Esseen bound and probability of moderate deviation, respectively.

It is also possible to obtain such results for $c = 0$, when no ordinary moment higher than 2 exists; see Theorems 2–3 of Dasgupta (2008), see also Remark 2 and (3.16) of Ghosh and Dasgupta (1978). In such case we have the following. Proofs are similar.

Let $Eh^2u(h) < \infty$ and $u(h)$ be a slowly varying even function. Then there exist a constant $b > 0$ dependent only on u such that
$| P(T_n \le x) - \Phi(x) | \le b[u(\sqrt{n})]^{-1}(1 + x^2), \quad -\infty < x < \infty.$
Further, if
$x_n^2 \le [2 \log | x_n | + 2 \log u(rn^{1/2}\sigma_n x_n)] + M, 0 < r < 1/2, M > 0$; then
$1 - P(T_n \le x_n)) \sim \Phi(-x_n) \sim P(T_n \le -x_n);\ x_n \to \infty.$

For $u(x) = \log^a(1 + | x |), a > 0$, the above normal approximation zone becomes
$x_n^2 \le 2a \log\log n + 2a \log\log\log n + M.$

This normal approximation zone is smaller than moderate deviation zone, since no ordinary moment higher than 2 exists for h.

Sharper nonuniform bounds and probabilities of deviations of higher order for the estimator are possible by using the results of Dasgupta (1989) under stronger moment assumptions. In Dasgupta (2008, 2013a) deviation probabilities of two sample U-statistics are considered under different moment assumptions. Probabilities of deviations are useful in computing efficiencies of test statistics. Chernoff type large deviations are discussed next.

2.2 Chernoff Type Large Deviation Probabilities

We discuss the case when the order $m = 1$. Further, without loss of generality, we assume that $h(x) = x$.

Then we have

$$\theta_n - E(\theta_n) = (\bar{X}_n - \mu) + a_n U_n(\tilde{D}) \tag{7}$$

where

$$\tilde{D}(x_1, x_2) = |x_1 - x_2| - E|X_1 - X_2| = |x_1 - x_2| - \Delta$$

Thus the large deviation probabilities of $\theta_n - E(\theta_n)$ will be the same as $(\bar{X}_n - \mu)$ provided the second term in (7) is negligible.

Assume that for some $s_0 > 0$, $E \exp(s_0|X|) < \infty$. Then it follows that for $0 < s < s_0$,

$\psi_{\tilde{D}}(s) = E \exp(s\tilde{D}(X_1, X_2)) \le \exp(-s\Delta)[E(s_0|X|)]^2 \le R^2$ say

Using this and Lemma C of Serfling (1980, page 200), with $k = [n/2], t = sk$ and $0 < s < s_0$, it follows that

$$\begin{aligned} P(a_n U_n(\tilde{D}) > \delta_n a) &\le \exp(-t a_n^{-1}\delta_n a) E[\exp(t U_n(\tilde{D}))] \\ &\le \exp(-t a_n^{-1}\delta_n a)[\psi_{\tilde{D}}(s)^k] \\ &\le \exp(-t a_n^{-1}\delta_n a)(R^2)^k = \exp(-s[n/2]a_n^{-1}\delta_n a)(R^2)^{[n/2]} \end{aligned}$$

A similar bound holds for $P(-a_n U_n(\tilde{D}) > \delta_n a)$. Choosing $\delta_n \to 0$ such that $b_n = a_n^{-1}\delta_n \to \infty$, we have for some $c_1 > 0$,

$$P(a_n|U_n(\tilde{D})| > \delta_n a) \leq \exp(-c_1 n b_n) \tag{8}$$

We then have the following theorem

Theorem 2.3. *Assume that* $E\exp(s_0|X|) < \infty$, *for some* $s_0 > 0$. *Further assume that* $P(X - \mu > a) > 0$. *Then*

$$n^{-1}\log P(\theta_n - E(\theta_n) \geq a) \to \log \rho_a$$

where,

$$\rho_a = \inf_{t\geq 0} e^{-at} E[e^{t(X-\mu)}]$$

Proof. The proof is easy by using relation (8) and the following fact from Bahadur (1971, page 9, relation (3.12))

$$\frac{1}{n}\log P(\bar{X}_n - \mu \geq (1-\delta_n)a) \to \log \rho_a.$$

Chernoff type large deviation for U statistics when the order $m > 1$ may be obtained under more stringent assumption of "strong-orthogonality" of the kernel ϕ as mentioned in Dasgupta (2008), Remark 1. In the set-up of independent random variables, not necessarily iid, the restriction on kernel ϕ *which should ensure that* $E\overline{\psi}^{(2)}(X_{i_1}, X_{j_1})\cdots\overline{\psi}^{(2)}(X_{i_{2q}}, X_{j_{2q}}) = 0$, unless each pair of suffixes (i_k, j_k) is repeated at least twice in the above expansion, provides a sharp bound on the remainder; consequently, Chernoff type large deviation behaviour of U statistics in such cases is similar to that of independent random variables appearing in Hájek's projection $\hat{U}$, see Dasgupta (1984, 2008). As the perturbation $a_n \to 0$, one may apply the result from Bahadur (1971, page 9, equation (3.12)) to projection $\hat{U}_n$ to show that the same large deviation property holds for the proposed one sided estimator (2) based on U statistics under this stringent assumption. □

Remark 2 (Variance of the Estimator θ_n*).* To obtain approximation for the variance and hence the mean square error of θ_n, the representation (3), (4), and (5) with $m_0 = 1$, are useful. Using these relations we have when $Eh^2 < \infty$ and $a_n^{-1} = O(n)$,

$$\begin{aligned}\mathrm{Var}(n^{1/2}\theta_n) &= m^2\mathrm{Var}(\tilde{h}_1(X_1)) + 4m^2 a_n \mathrm{Cov}(\tilde{h}_1(X_1), \tilde{D}_1(X_1)) \\ &\quad + 4m^2 a_n^2 \mathrm{Var}(\tilde{h}_1(X_1)) + O(n^{-1}) \\ &= m^2\mathrm{Var}(\tilde{h}_1(X_1)) + O(a_n)\end{aligned}$$

One sided estimators are biased, $E(\theta_n) = \theta + a_n\Delta$, because of the stringent restrictions imposed on these to converge from a particular direction. Hence

$$\mathrm{MSE}(\theta_n) = a_n^2\Delta^2 + \frac{m^2}{n}\mathrm{Var}(\tilde{h}_1(X_1)) + O(n^{-1}a_n)$$

3 Rates of Convergence for U Statistics

In this section we prove some general results regarding nonuniform rates of convergence in CLT for U statistics when m.g.f. of the kernel h necessarily exist. Similar results may be proved for one sided estimator θ_n exploiting the U statistics structure (1) and (2).

Recall the notations: $X, X_1, \cdots, X_n$ are iid with distribution function F, $U_n(h)$ is a U statistic based on kernel $h = h(x_1, \cdots, x_m)$, and $\theta = E_F h$.

Following the steps used in Proposition 1 of Dasgupta (2008), it is possible to obtain the following. We adopt the notations of Serfling (1980).

Theorem A. *Let U_n be a U statistic based on a symmetric kernel h where $E|h|^v < \infty$, for some real $v > 1$. Then*

$$U_n - \theta = \hat{U}_n - \theta + R_{2n} + \cdots + R_{mn} \tag{9}$$

where

$$E|R_{jn}|^v \le L^v n^{-jv/2} e^{jv\log v} E|h|^v \tag{10}$$

and $L > 1$ is a constant that does not depend on n and v.

When the first conditional variance of h is positive i.e., $\zeta_1 > 0$, then $\hat{U}_n - \theta = \frac{m}{n}\sum_{i=1}^n \tilde{h}_1(X_i)$, where $\tilde{h}_1(x) = Eh(x, X_2, \cdots, X_m) - \theta$. When first $(c-1)$ conditional variances are zero and the cth one is positive i.e., $\zeta_1 = \cdots = \zeta_{c-1} = 0 < \zeta_c$, then one may define

$$\hat{U}_n - \theta = \frac{m(m-1)\cdots(m-c+1)}{n(n-1)\cdots(n-c+1)} \sum_{1\le i_1<\cdots<i_c\le n} \tilde{h}_c(X_{i_1}, \cdots, X_{i_c}) \tag{11}$$

see Serfling (1980, page 180). This reduces to the usual projection taking $c = 1$.

We shall prove a bound for $P(\sup_{i\ge n}|U_i - \hat{U}_i| > t)$ and use this to obtain nonuniform CLT bound for standardised U statistic in the nondegenerate case. The following result is nontrivial for all values of t and is of independent interest.

Theorem 3.1. *Let U_n be a U statistic based on a symmetric kernel h of degree m. Let $\zeta_1 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. Define $\hat{U}_n$ as in (11), then under the following moment bound*

$$E|h|^v \le L^v e^{\gamma v\log v}, \forall v \ge 2, L > 0 \text{ and } \gamma \ge 0, \tag{12}$$

one has

$$P(\sup_{i\geq n}|U_i-\hat{U}_i|>t)\leq \exp[-\alpha(n^{(c+1)/2}t)^{1/(\gamma+m)}] \tag{13}$$

where $\alpha>0$ *is a constant depending only on* L,γ,m *and* c.

For a nondegenerate U statistic, $\zeta_1>0$ and $c=1$ in the above.

Proof. Note that $(U_n-\hat{U}_n)$ is itself a U statistic and using reverse martingale property of U statistics (see page 189 of Serfling 1980), we can write from (10) and Minkowski's inequality

$$\begin{aligned} P(\sup_{i\geq n}|U_i-\hat{U}_i|>t) &\leq t^{-v}E|U_n-\hat{U}_n|^v \\ &\leq t^{-v}L^v n^{-(c+1)v/2e^{(m+\gamma)v\log v}}E|h|^v \end{aligned} \tag{14}$$

under (12), considering higher power of terms over index j in the product. In the above $L>0$ represents a generic constant.

Taking logarithm of the term in r.h.s. above and differentiating it with respect to v, we find the optimal choice of v as

$$v=e^{-1}\{n^{(c+1)/2}tL^{-1}\}^{1/(m+\gamma)}$$

The minimum value of the r.h.s. of (14) corresponding to above v is

$$\exp[-(m+\gamma)v]=\exp[-(m+\gamma)e^{-1}\{n^{(c+1)/2}tL^{-1}\}^{1/(m+\gamma)}]$$

Hence the theorem. □

Remark 3. A slight improvement of Theorem 3.1 is possible. Rewrite (14) as

$$P^*:=P(\sup_{i\geq n}|U_i-\hat{U}_i|>t)\leq t^{-v}L^v\sum_{j=c+1}^{m}n^{-jv/2}e^{(j+\gamma)v\log v}$$

under (12). Observe that the common ratio of the geometric series in the above sum over j is $(v/n)^{v/2}=r$, say. Therefore in the case $r\geq 1$, replacing the sum by $(m-c)$ times the mth term, one gets

$$\begin{aligned} P^* &\leq t^{-v}L^v(m-c)n^{-mv/2}e^{(m+\gamma)v\log v} \\ &\leq t^{-v}L^v n^{-mv/2}e^{(m+\gamma)v\log v} \end{aligned}$$

for some $L>1$. Minimising the r.h.s. with respect to v one gets

$$P^*\leq \exp[-\alpha(n^{m/2}t)^{1/(\gamma+m)}], \text{ for some } \alpha>0.$$

Similarly, considering the term corresponding to $j = c + 1$ when $r \leq 1$, one may write

$$P^* \leq \exp[-\alpha(n^{(c+1)/2}t)^{1/(\gamma+c+1)}], \text{ for some } \alpha > 0.$$

Combining the above two inequalities for P^*, we get an overall bound.

$$P^* \leq \exp[-\alpha n^{(c+1)/2(c+1+\gamma)} t^{1/(\gamma+m)}], \text{ for some } \alpha > 0. \tag{15}$$

This is an improvement over (13), the bound given in Theorem 3.1. The results (13), (15) provide sharp rate of strong convergence for U statistics under stronger assumptions.

Consider a nondegenerate U statistic. So $\zeta_1 > 0$, and write

$$\begin{aligned} \frac{n^{1/2}(U_n - \theta)}{m\zeta_1^{1/2}} &= \frac{n^{1/2}(\hat{U}_n - \theta)}{m\zeta_1^{1/2}} + \frac{n^{1/2}(U_n - \hat{U}_n)}{m\zeta_1^{1/2}} \\ &= (n\zeta_1)^{-1/2} \sum_{i=1}^{n} \tilde{h}_1(X_i) + R_n^*, \text{ say} \end{aligned} \tag{16}$$

Representation (16) is clearly of the type (4.1) of Dasgupta (2006). From (13) one may get the following

$$\begin{aligned} P(|R_n^*| > a_n(t)) &= P(|U_n - \hat{U}_n| > n^{-1/2} m\zeta_1^{1/2} a_n(t)) \\ &\leq \exp[-\alpha(n^{1/2}a_n(t))^{1/(\gamma+m)}] \end{aligned} \tag{17}$$

for some $\alpha > 0$, letting $c = 1$ in (13). This bound is similar to (4.2) of Dasgupta (2006), with $\beta = 0$ and $\delta = m$.

It may be mentioned that $k = k(\gamma)$ in (4.6) of Dasgupta (2006) may be taken arbitrarily large for $\gamma + \delta \in (1/2, 1)$, with the notations used therein. Supplemented by the results of Dasgupta (1989) [see, e.g., (2.27) of Dasgupta (1989), with λ_2 arbitrarily large] while computing the first term in the r.h.s. of (4.5) in Dasgupta (2006), it is evident that the coefficient k of $|t|^{2\wedge 1/(\gamma+\delta)}$ in Theorem 4.1 of Dasgupta (2006) may be taken arbitrarily large for an *extended* zone $\gamma + \delta \in (1/2, \infty)$.

Now proceeding like Theorem 4.1 of Dasgupta (2006), one may obtain the following theorem on nonuniform rates in CLT for U statistics.

Theorem 3.2. *Let U_n be a U statistic based on a symmetric kernel h of degree m. Let $\zeta_1 > 0$ and (12) hold for $\gamma \geq 0$. Let k be an arbitrary large constant. Then there exists a constant $b > 0$ depending on k and γ such that for $t \in (-\infty, \infty)$,*

$$|P(\frac{n^{1/2}(U_n - \theta)}{m\zeta_1^{1/2}} \leq t) - \Phi(t))| \leq bn^{-1/2}(\log n)^{\gamma+m} e^{-k|t|^{1/(\gamma+m)}}$$

4 Rates of Convergence in CLT for Some Special L Statistics

Let $X_1, \cdots, X_n$ be iid copies of a random variable X from a distribution with $E|X| < \infty$. Also let $X_{(1)}, X_{(2)} \cdots, X_{(n)}$ be the ordered observations.

Consider an L statistic of the form

$$L_n(k) = \frac{k}{n^k} \sum_{i=1}^{n} i^{k-1} X_{(i)}, \text{where } k \text{ is a positive integer.} \tag{18}$$

Then $L_n(k) \to E(\max(X_1, \cdots, X_k)) = M_k$ a.s.

L statistic of this form is useful in one sided estimation, see, e.g., Gilat and Hill (1992).

$L_n(k)$ can be related to a U statistic U_n with kernel $h = h_k = h(x_1, \cdots, x_k) = \max(x_1, \cdots, x_k)$. Note that for $k = 1$, $L_1(1) = n^{-1} \sum_{i=1}^{n} X_{(i)} = U_n(h_1)$, the sample mean.

For $k = 2$,

$$\begin{aligned} U_n(h_2) &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \max(X_{(i)}, X_{(j)}) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} X_{(j)} \\ &= \binom{n}{2}^{-1} \sum_{1 \le j \le n} (j-1) X_{(j)} = \frac{n}{n-1} L_n(2) - \frac{2}{n-1} L_n(1) \end{aligned}$$

Therefore,

$$L_n(2) - U_n(h_2) = \frac{2}{n} U_n(h_1) - \frac{1}{n} U_n(h_2)$$

In a similar fashion, it can be shown that

$$L_n(k) - U_n(h_k) = a_1 U_n(h_1) + a_1 U_n(h_1) + \cdots + a_k U_n(h_k) \tag{19}$$

where the coefficients $a_i, i = 1, \cdots, k$ have order of magnitude at most $1/n$.

The above representation (19) enables us to obtain an estimate of $E|L_n(k) - U_n(h_k)|^v, v > 1$. We show the following.

Proposition 4.1. *Let $L_n(k)$ defined in (18) be an L statistic based on iid random variables $X_1, \cdots, X_n$. Let a $U_n = U_n(h_k)$ be a U statistic with kernel $h = h_k = h(x_1, \cdots, x_k) = \max(x_1, \cdots, x_k)$, and $\hat{U}$ be Hájeck's projection of U. Then*

$$E|L_n(k) - \hat{U}_n(h_k)|^v \le B^v n^{-v} e^{kv \log v} E|X|^v \tag{20}$$

for all $v > 1$, where $B > 0$ is a constant which may depend on k.

Proof. In what follows, let $B > 1$ denote a generic constant. Observe that $\max(x_1, \cdots, x_k) \leq \sum_{i=1}^{k} |x_i.|$ Therefore

$$E|h(m)|^v = E|\max(x_1, \cdots, x_k)|^v \leq B^v E|X|^v; m = 1, \cdots, k \tag{21}$$

In view of (9), (10) and (21), one gets

$$E|U_n(h_m) - \hat{U}_n(h_m)|^v \leq B^v \sum_{j=2}^{m} n^{-jv/2} e^{jv \log v} E|X|^v \tag{22}$$

Recall a moment bound for general stochastic processes including martingales stated in Dasgupta (1993) (there is an error in page 151, as

$E \max(|S_n|^{v-2}X_n^2, |S_n^*|^{v-2}X_n^2) < E(|S_n|^{v-2}X_n^2 + |S_n^*|^{v-2}X_n^2)$. Thus an additional multiplicative factor 2 appears in the r.h.s. of (4) therein, and the modified value is $c_v = \{2(v-1)\delta\}^{v/2}$ in (1) therein).

With an application of said bound to the sum of iid random variables, one may write

$$E|\hat{U}_n(h_m) - M_m|^v \leq B^v n^{-v/2} e^{v \log v} E|X|^v; m = 1, \cdots, k \tag{23}$$

where $M_m = E \max(X_1, \cdots, X_k) \leq mE|X| \leq kE|X|$.

Next write

$$\begin{aligned} E|U_n(h_m) - M_m|^v &\leq (E|U_n(h_m) - \hat{U}_n(h_m)|^v + E|\hat{U}_n(h_m) - M_m|^v) B^v \\ &\leq B^v \textstyle\sum_{j=0}^{m} n^{-jv/2} e^{jv \log v} E|X|^v \end{aligned} \tag{24}$$

for all $v > 1$, from (21) and (23). Therefore from (19) and (24), recalling that $a_i = O(1/n), i = 1, \cdots, k$, one gets

$$E|L_n(k) - U_n(h_k)|^v \leq n^{-v} B^v \sum\nolimits_{j=0}^{k} n^{-jv/2} e^{jv \log v} E|X|^v \tag{25}$$

From (22) with $m = k$, and (25) we get

$$\begin{aligned} E|L_n(k) - \hat{U}_n(h_k)|^v &\leq B^v \Big(\sum_{j=2}^{k} n^{-jv/2} e^{jv \log v} + n^{-v} \sum_{j=0}^{k} n^{-jv/2} e^{jv \log v}\Big) E|X|^v \\ &\leq B^v \sum_{j=2}^{k} n^{-jv/2} e^{jv \log v} E|X|^v \\ &\leq B^v n^{-v} e^{kv \log v} E|X|^v \end{aligned}$$

This completes the proof. □

Now assume the following moment bound for the basic random variable X.

$$E|X|^v \le L^v e^{\gamma v \log v}, \forall v > 1, \text{ and for some } L > 1 \tag{26}$$

Then from (25) one may write

$$E|L_n(k) - \hat{U}_n(h_k)|^v \le B^v n^{-v} e^{(k+\gamma)v \log v} \tag{27}$$

Define

$$L_n^*(k) := \frac{n^{1/2}(L_n(k) - M_k)}{k\zeta_1^{1/2}} = \frac{n^{1/2}(U_n(h_k) - M_k)}{k\zeta_1^{1/2}} + \frac{n^{1/2}(L_n(k) - U_n(h_k))}{k\zeta_1^{1/2}} \tag{28}$$

where $\zeta_1 = \text{var}[h(X)], h(x) = E \max(x, X_2, \cdots, X_k)$.

When the distribution of X is nondegenerate, $\zeta_1 > 0$; and from (27) we observe that a representation like (4.1) of Dasgupta (2006) holds for the standardised L statistic $L_n^*(k)$. From (27), we further observe that the remainder term in the representation (28) satisfies (4.2) of Dasgupta (2006) with $\beta = 0$ and $\delta = k$. Hence the following nonuniform CLT bound on L statistics holds, this is similar to Theorem 3.2.

Theorem 4.1. *Let the distribution a random variable X be nondegenerate and (26) hold. Then for the standardised L statistic $L_n^*(k)$ defined in (28) the following holds for $k \ge 2$*

$$|P(L_n^*(k) \le t) - \Phi(t)| \le bn^{-1/2}(\log n)^{\gamma+k} e^{-\alpha|t|^{1/(\gamma+k)}}, t \in (-\infty, \infty)$$

where $b = b(k, \gamma) > 0$ and $\alpha > 0$ may be taken arbitrarily large.

Remark 4. A linear combination of $L_n(j), j = 1, \cdots, k$ of the following form, where the coefficients α might depend on n is also an L statistic for which the above results may be extended. Consider

$$L_n' := \alpha_1 L_n(1) + \alpha_2 L_n(2) + \cdots + \alpha_k L_n(k), \alpha_k = \alpha_{nk} \ne 0 \tag{29}$$

This may be expressed as a linear combination of U statistics, in view of representation (19).

The estimator proposed in Gilat and Hill (1992) is

$$\hat{X}_n = \sum_{i=1}^{n} (\frac{1}{n} - \frac{n+1}{2n^\alpha} + \frac{i}{n^\alpha}) X_{(i)} = a_{1n} L_n(1) + a_{2n} L_n(2)$$

$\alpha \in (2, 5/2)$, is of the form (29). Let

$$U_n' = \alpha_1 U_n(h_1) + \alpha_2 U_n(h_2) + \cdots + \alpha_n U_n(h_n), \ \hat{U}_n' = \sum_{j=1}^{k} \alpha_j \hat{U}_n(h_j) \tag{30}$$

Following the steps to prove Proposition 4.1, one can show that

$$E|L_n' - \hat{U}_n'|^v \leq B^v n^{-v} e^{kv \log v} E|X|^v \tag{31}$$

Observe that $\hat{U}_n'$ is expressible as sample mean of independent random variables to which standard theory applies; see also (3.4) of Dasgupta (2006). Define the standardised version $\tilde{L}_n$ of L_n' as

$$\tilde{L}_n = (L_n' - \sum_{j=1}^{k} \alpha_j M_j)/(\text{var}(\hat{U}_n')^{1/2}) \tag{32}$$

In view of (31) and following the steps to prove Theorems 3.2 and 4.1, see also Theorem 4.1 of Dasgupta (2006), it is possible to extend the results of Theorem 4.1 for $\tilde{L}_n$ replacing L_n^*.

5 Almost Sure Confidence Band for Yam Growth

Data analysed below refers to experiments on Elephant foot yam conducted in ISI Giridih Farm in the years 2013–2014. Average seed corm weight is 500 g in Experiments 1, 2, 3 and 5; for Experiment 4, the average seed weight is 20 g. Farmers sometimes do not harvest yam at the end of a production season, keep this underground for another year to sprout again. The yield at the second year is much higher even if the initial seed size was small, due to accumulated affect of 2 years.

Lowess, a local polynomial regression estimator with smooth tricubic kernel and variable bandwidth based on k-th nearest neighbour, employs weighted least square criterion that assigns less weights to distant observations, to have a robust estimate of response curve insensitive to large-residual outliers, by down-weighting these over several iterations; see Cleveland (1979, 1981). However, lowess does not provide an explicit functional form of response variable with predictor variables.

A broad idea about the (mean) growth of yam over plant lifetime is explained via these techniques.

Lowess regression is done with the proportion of the data in the smoothing window $f = 2/3$ for all concerned plots, except for Figs. 7, 12, 13 and 17 where $f = 1/3$, the number of iteration performed is 3 in all cases. The goal here is to estimate a continuous curve by gaining strength from adjacent points.

With 96 observations of Experiment 1, the yam growth curve is estimated by nonparametric lowess regression in Fig. 1. This experimental plot is of lateritic soil and full of gravels. One seed corm did not germinate and 3 observations arising from weak plants were not considered out of seed plantations in 100 pits. At each observed point on lifetime t of yam plant, we compute the perturbation

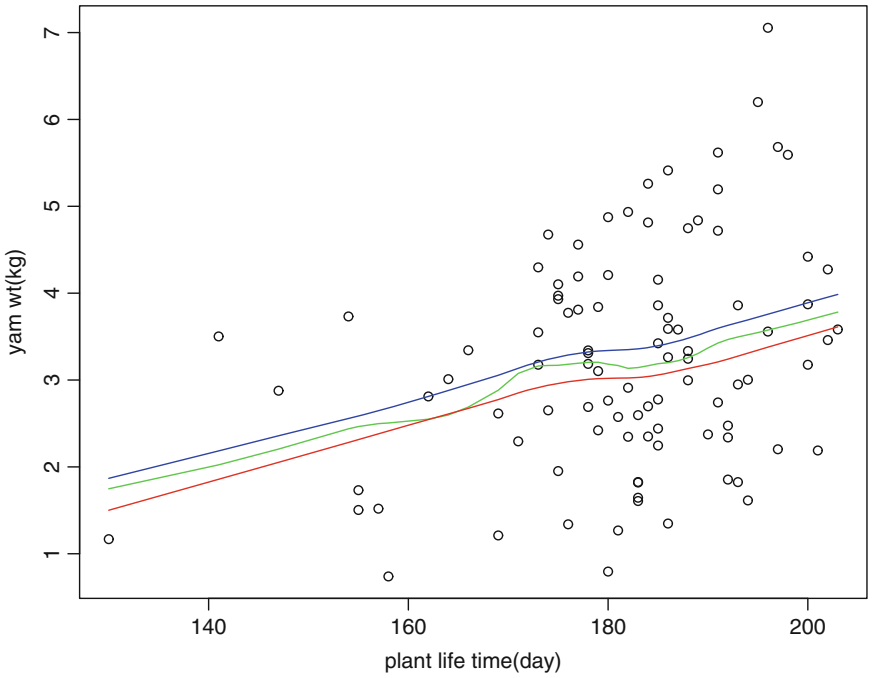


Fig. 1 Almost sure band for growth curve: Expt1

$\frac{1}{4n^\alpha}\sum_{i,j=1}^{n}|X_i - X_j|$, $\alpha \in (2, 2.5)$ as proposed in Gilat and Hill (1992). We take two immediate points above and below t and consider the corresponding yam yield $X = X(t)$ to compute the perturbation part with $n = 3, \alpha = 2.25$. The points so obtained by addition/subtraction from lowess curve fall above/below the lowess growth curve, and the points falling in a particular side of the curve are again lowess smoothed to obtain approximate a.s. upper/lower confidence bands as seen in Fig. 1.

We next obtain estimate of variance by considering the kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$. At each observed point of time t, a U statistic is computed out of 3 observed times points viz., a particular point along with immediate above and below lifetime observations from that particular time point. This results in $\binom{3}{2} = 3$ such values of observed kernel for the middle time point. Mean of these values i.e., value of $U(h)$ so computed, is assigned to the middle lifetime point. The procedure is carried out for all observed lifetime, except for the smallest and largest plant lifetime, where no value is available to the left/right of the smallest/largest lifetime. By lowess regression to these U values we then compute the central curve of variance estimate. The perturbation part $\frac{1}{4n^\alpha}\sum_{i,j=1}^{n}|X_i - X_j|$ is computed after X values are replaced by h values and choosing $n = 3, \alpha = 2.25$. Apart from a multiplicative constant, the U statistic $U(D)$, corresponding to symmetrised version of the kernel $D(x_1, x_2, x_3) = |h(x_1, x_2) - h(x_2, x_3)|$, is used for computing

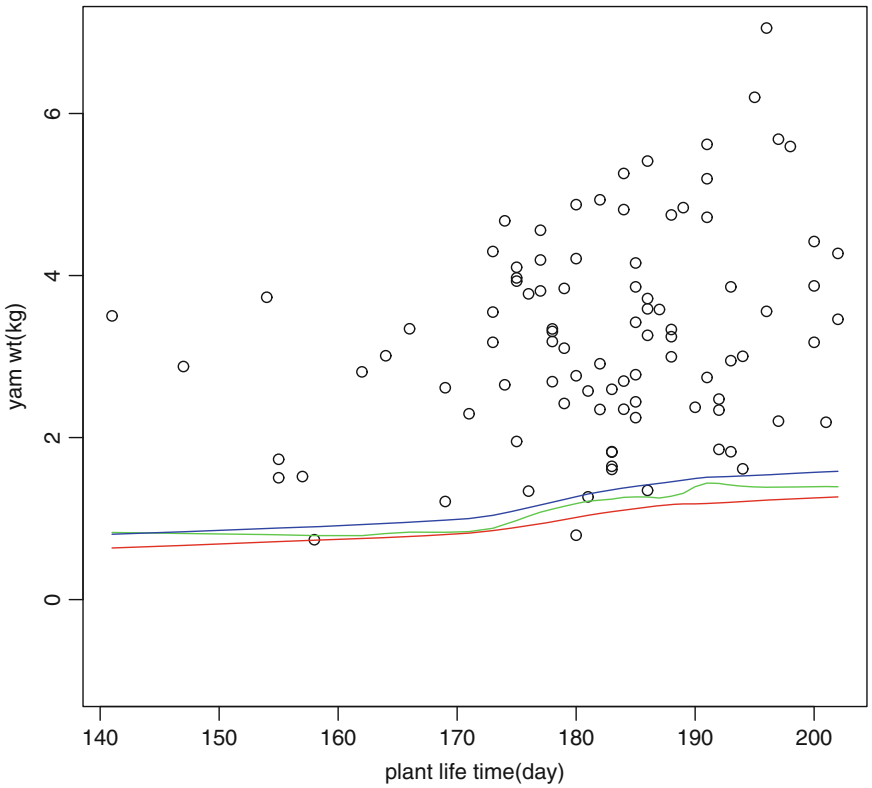


Fig. 2 Almost sure band for variance curve: Expt1

perturbation component. The a.s. upper and lower confidence band for variance are obtained by lowess regression on these shifted points above and below, respectively, of the central curve. These three curves, lower, central and upper, are shown in Fig. 2, variance seems to increase slightly over lifetime. To understand the nature of variation, these curves (with measurement unit kg^2) are plotted along with data-scatter.

Since the upper and lower points are lowess smoothed to gain strength from adjacent points, central line may not always remain equidistant from upper and lower band as seen in Figs. 1 and 2.

Scatterplot of data related to crop yields often show high scattering. Growth curve shown in Fig. 1 may be further smoothed if the observations are locally averaged before lowess regression is made to the observed data on plant lifetime and yield $(t_j, y_j); 1 \leq j \leq n$. At time point $t_j, 1 < j < n$ we assign the average yield $\tilde{y}_j = (y_{j-1} + y_j + y_{j+1})/3$ to smooth out local irregularities. The central line and the bandwidth points $(t_j, \tilde{y}_j), 1 < j < n$ following the similar procedure used to obtain Fig. 1 are shown in Fig. 3.

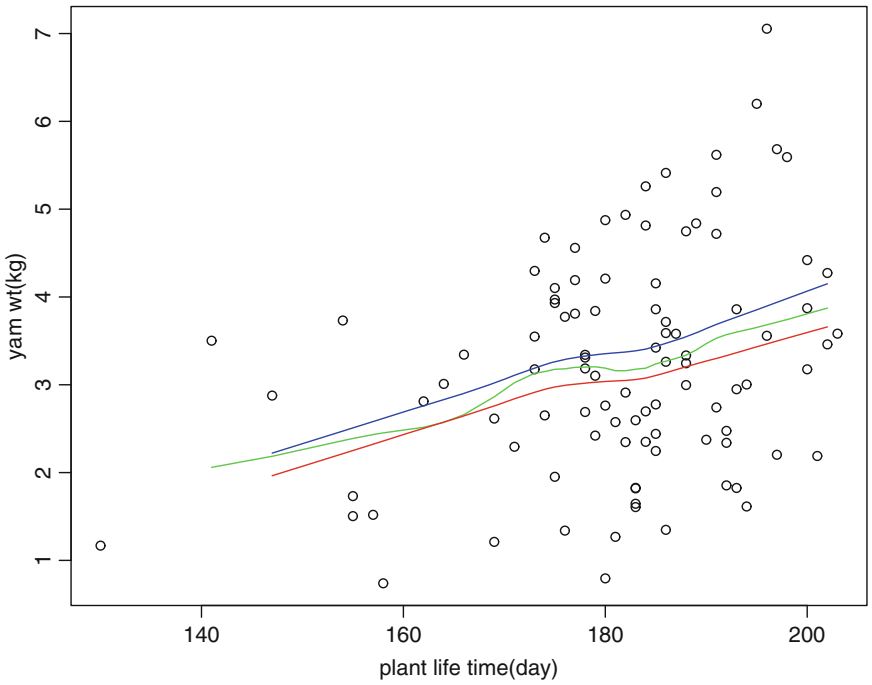


Fig. 3 Almost sure band for smoothed growth curve: Expt1

Figures 4, 5, 6, and 7 refer to Experiment 2, conducted in a slightly elevated experimental plot in a fertile land exposed to harsh summer, and the plot was partly subjected to water logging during rainy season. Out of 100 seed corms, one corm did not germinate. Organic fertilisers namely mustard oil-cake, vermicompost, etc. were administered before starting the experiment, and several times during the experiment. On harvest, some yams were seen damaged as these were infested by small white ants and worms. To deal with initial irregularities in Experiment 2 we plot plant lifetime after yam plants cross a height of 15 cm. The growth curve and the band are shown in Fig. 4. The initial point at lifetime as 52 days with weight 170 g pulls down the curve at start, and a hump in growth is observed around plant lifetime of 165 days. The variance curve is shown in Fig. 5, this seems to exhibit an increasing trend.

A smoothed version of growth curve for this experiment is obtained in Fig. 6, following local averaging as adopted to obtain Fig. 3. The features noted in Fig. 6 are similar to that of Fig. 4, except that the central line now remains within the band in smoothed version.

We may examine the effect of ignoring the point corresponding to lowest lifetime in the scatter. Figure 7 shows the resultant growth curve along with confidence bandwidth 98 data points, the central line lies within band most of the times and the curves are lifted up in the start compared to Figs. 4 and 6.

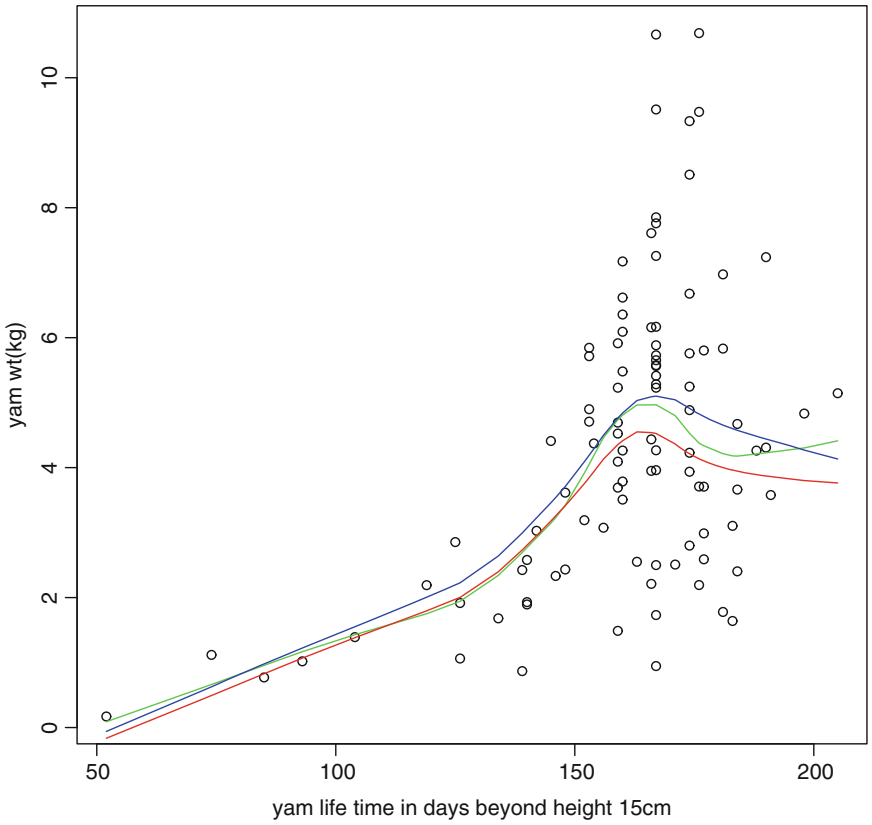


Fig. 4 Almost sure band for growth curve: Expt2

In Figs. 8, 9, and 10, we plot the yam data from Experiment 3 in the Rose villa campus of ISI. Two corms did not germinate out of 100 pits made in an unfertile plot of land exposed to harsh weather of Jharkhand. The growth curve in the middle of Fig. 8 shows that around plant life of 80 days the mean yield is more than double the seed size, then a slow upward growth continues till plant life of 140 days; a change in rate of growth occurs afterwards and the peak growth is seen around 172 days. Upper and lower a.s. confidence band exhibit a similar pattern. For monetary reasons, farmers sometimes opt for early harvest around 90 days, the doubling time. The variance curve of Fig. 9 exhibits an upward trend, except around 160 days, where this seems to be more or less steady with a little bit of fluctuation within band. Growth curve after locally averaging, as obtained in Fig. 3 and Fig. 6 for Experiment 1 and 2, respectively, is shown in Fig. 10 for Experiment 3. Figure 10 seem to ignore the effects of three data points seen in low right corner; of these two points, (187, 0.92) and (187, 0.917) are overlapping. A similar effect like Fig. 10 is seen in Fig. 11, when these two overlapping points are not considered for lowess regression.

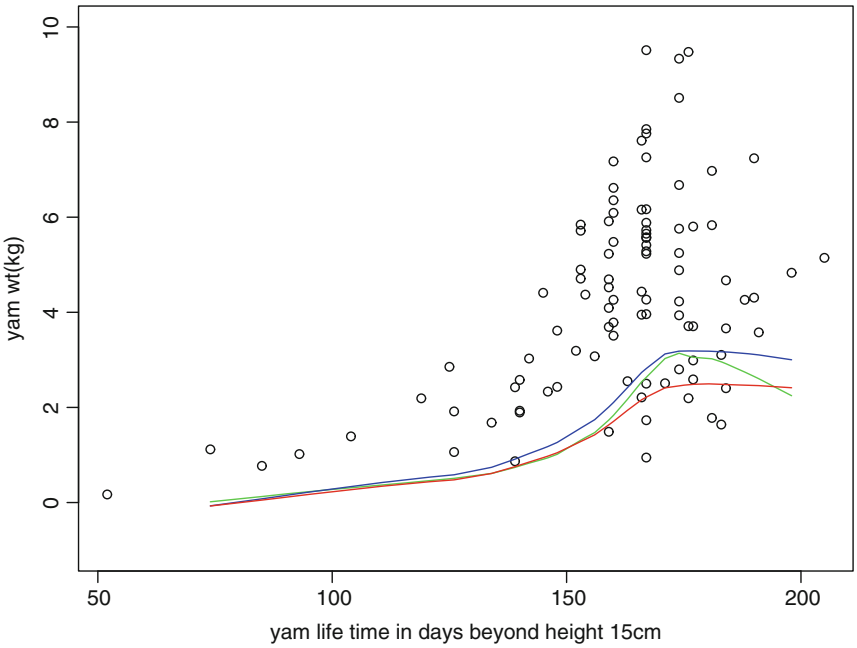


Fig. 5 Almost sure band for variance curve: Expt2

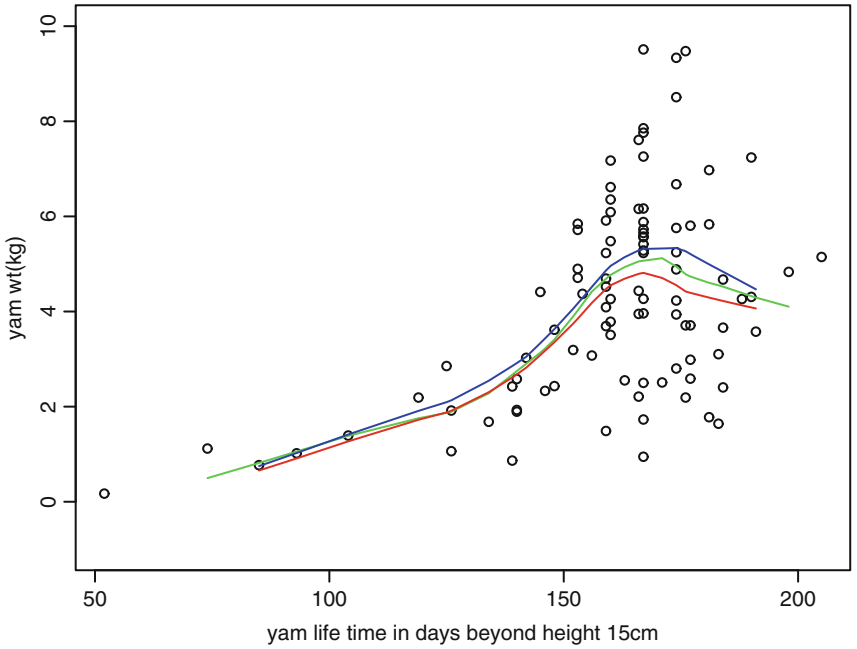


Fig. 6 Almost sure band for smoothed growth curve: Expt2

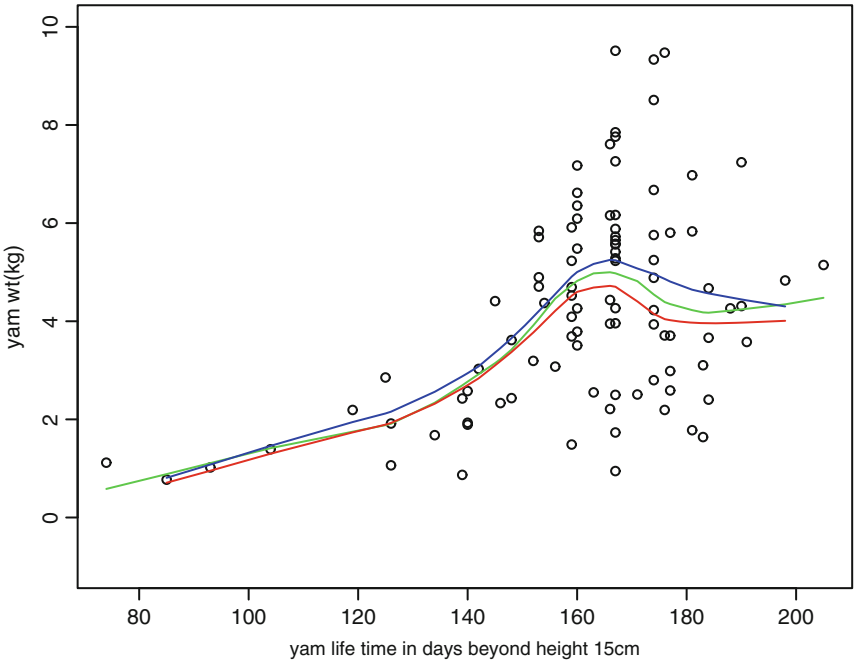


Fig. 7 Almost sure band for growth curve deleting lowest lifetime: Expt2

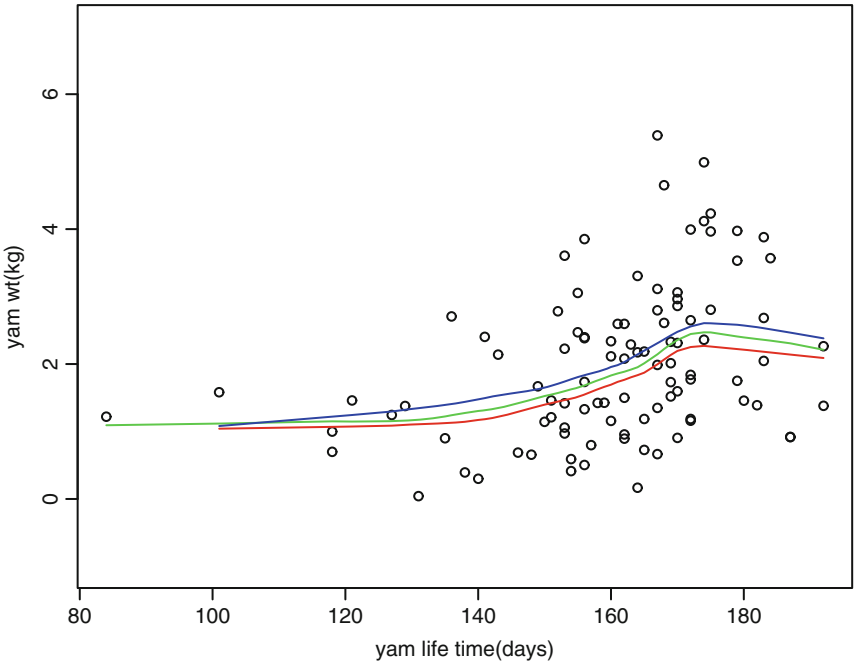


Fig. 8 Almost sure band for growth curve: Expt3

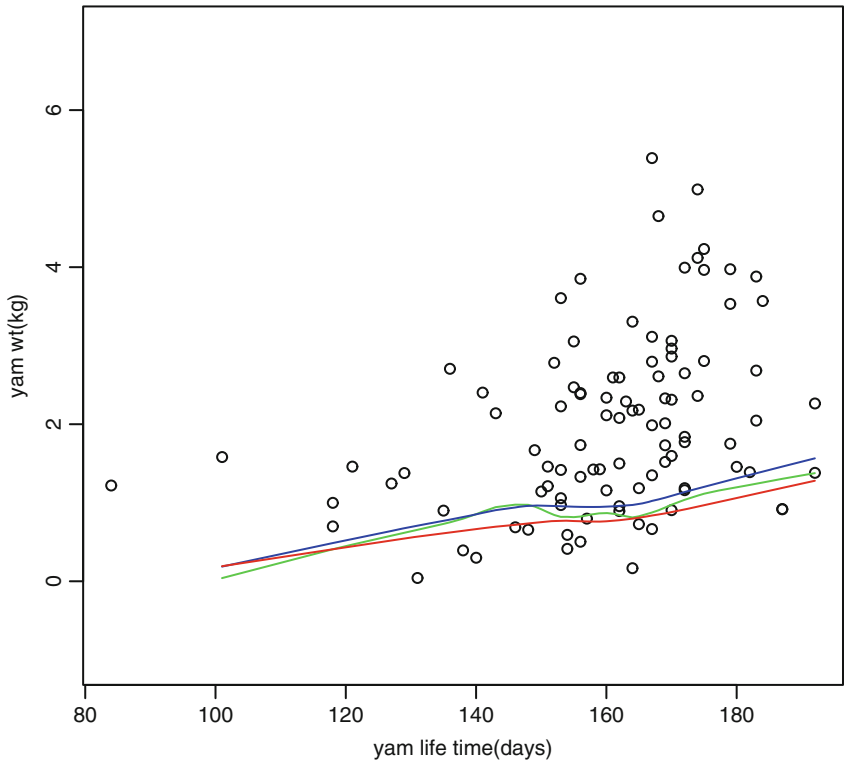


Fig. 9 Almost sure band for variance curve: Expt3

Experiment 4 refers to a 2-year study where the average seed weight is 20 g. In this experiment, yams were not harvested at the end of first year, but were allowed to remain underground that sprouted again; these were harvested at the end of second harvesting season. Figure 12 distinctly shows the presence of an upward spike in the growth curve of yam yield at the end of second season. Upper and lower confidence bounds are nearly equidistant from central growth curve. A yield of 3.7 kg was observed corresponding to a yam seed weight of 15 g. Smoothed version of the growth curve shown in Fig. 13 is similar to Fig. 12. The variance curve and upper and lower bands are seen in Fig. 14. This shows increase in variance towards higher values of plant lifetime.

In the harsh climate of Jharkhand, Experiment 5 was conducted in a shady region under the shade of Delonix regia, a species of flowering plant in the family Fabaceae; also known as *Krishnachura* tree to protect the plants from strong sun in summer. Direct and strong sunlight for long time may sometimes adversely affect plant growth. This experiment was disturbed over a large segment of yam plantation due to fall of a heavy branch from a tree above around middle of the growing season. One seed corm did not germinate out of 100. Increasing features of mean response curve

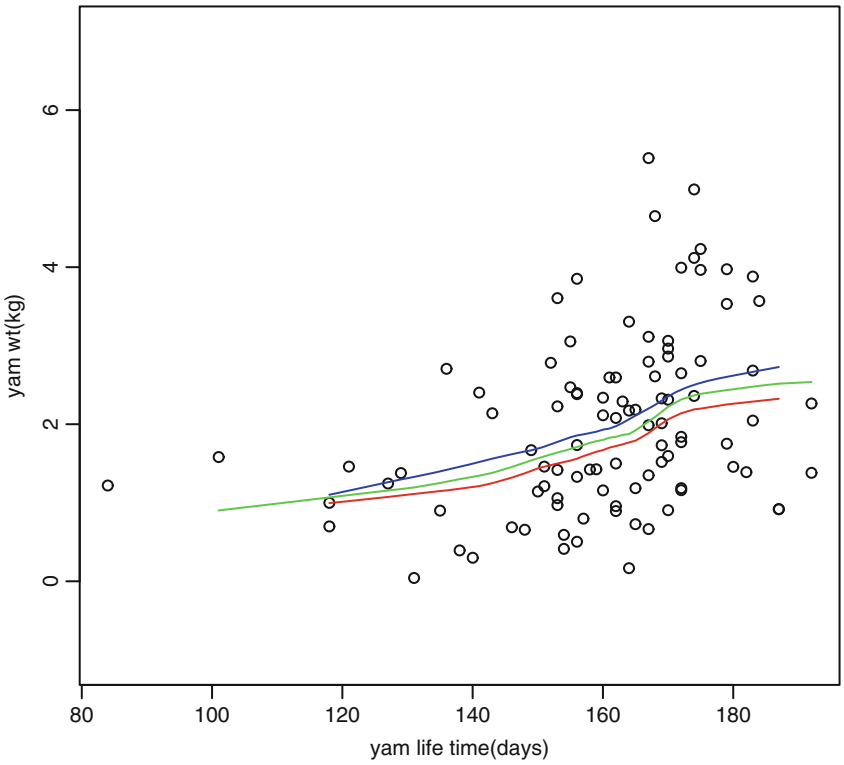


Fig. 10 Almost sure band for smoothed growth curve: Expt3

and the band are nicely revealed in Fig. 15 with 99 observations. The increasing trend is slightly dampened towards end when one considers the smoothed version of the growth curve as shown in Fig. 16, other features remain almost the same as those of Fig. 15. Variance of yield observations, as an increasing function of time is shown in the middle curve in Fig. 17. The upper and lower confidence bands are also shown; these exhibit increasing features while these stabilise towards end of yam lifetime with slight decrease.

Here we are estimating the response curve from data, and a band for response curve; the procedure works even though the data may be widely dispersed. Confidence bands constructed seem astonishingly narrow taking the large spread of data into account. The goal of constructing band here is different from including all data points inside it. Upper and lower estimate of yam growth curve cover the mean response curve in general, even for small values of n. Confidence band for variance of yam yield over time exhibits similar properties. Theoretical results indicate that the procedures suggested can be confidently used for applied purposes.

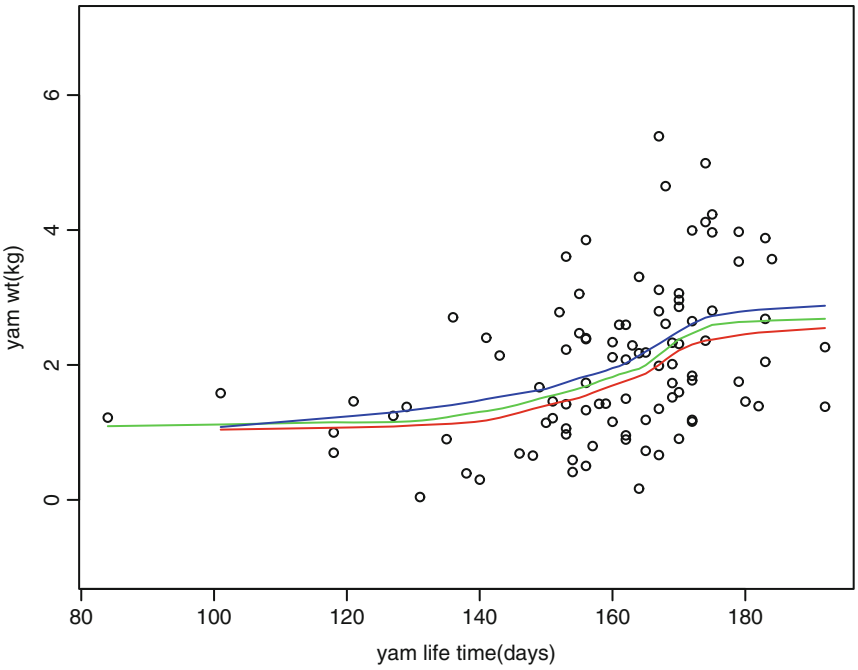


Fig. 11 Almost sure band for growth curve deleting two observations: Expt3

6 Search for a Parametric Model for Yam Growth

Nonparametric estimates of yam growth exhibit an upward spike towards end of plant lifetime, as seen in repeated studies. Traditional parametric curves like logistic, Gompertz lack this feature. Logistic and Gompertz growth curves can be deduced as a limiting form of a model that has exponentially decaying proliferation rate (Dasgupta 2013b).

The observed growth of yam stems to some extent resembles Gompertz curve

$$y(t) = a \exp(b \exp(ct)); a > 0, b < 0, c < 0. \tag{33}$$

In a confined space where the availability of nutrients is limited, growth rate is high in the beginning and then it slows down due to competition for nutrients. However, underground yam deposition is faster towards end of plant lifetime. Physiological processes governing the matter partitioning to the tubercles seem rapid towards end.

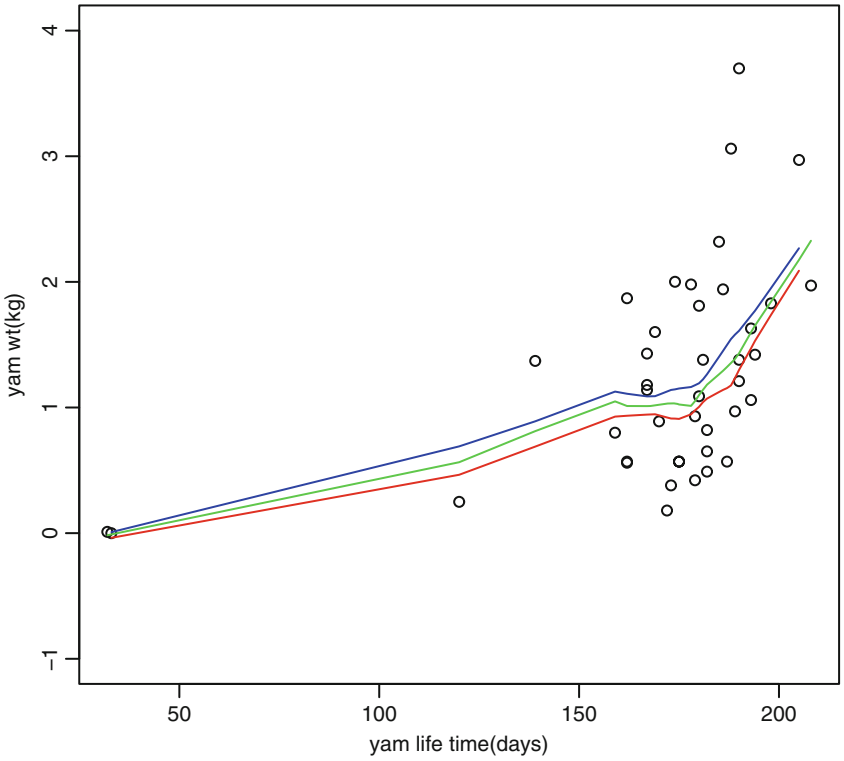


Fig. 12 Almost sure band for growth curve: Expt4 (small seed)

Growth pattern may sometimes vary in different time segments. Yam deposition is rapid towards end while the mature plant is shrinking and drying up. Yam growth $Y(t)$ beyond a time point t_0 may be modelled as a mirror reflection of the concave part of stem growth curve near its highest value a attained at time t^* say, the convex reflection being on the line $\ell(t)$ connecting the points origin and (t^*, a). Usually lifetime of an yam plant in a production season $t^* \leq 210$ days.

At time $t(> t_0)$, consider the stem height y following a Gompertz model (33). Then vertical distance between stem height and ℓ at time t is $y(t) - \ell(t)$. For yam yield the corresponding point is below the line ℓ and at same distance. Thus $Y(t) = \ell(t) - (y(t) - \ell(t)) = 2at/t^* - y(t) = a(2t/t^* - \exp(b\exp(ct)))$ may serve as an empirical parametric model for underground yam deposition at time $t(> t_0)$, depending on the growth pattern of stem observed above ground. A rescaled version of the above is

$$z(t) = w(2t/t^* - \exp(b\exp(ct))); w > 0, b < 0, c < 0, t_0 < t \leq t^* \qquad (34)$$

where w is the maximum attainable (plant specific) yam weight.

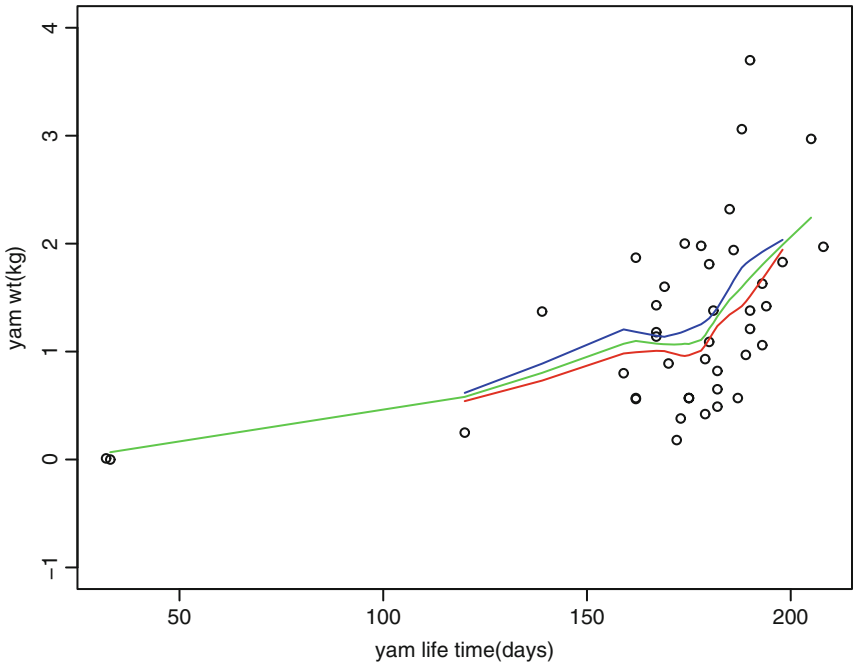


Fig. 13 Almost sure band for smoothed growth curve: Expt4 (small seed)

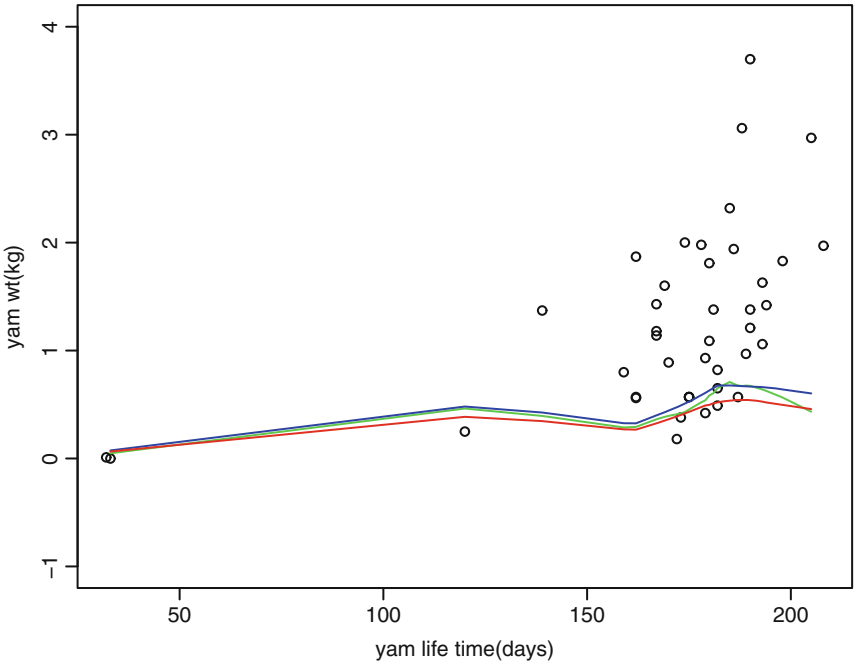


Fig. 14 Almost sure band for variance curve: Expt4 (small seed)

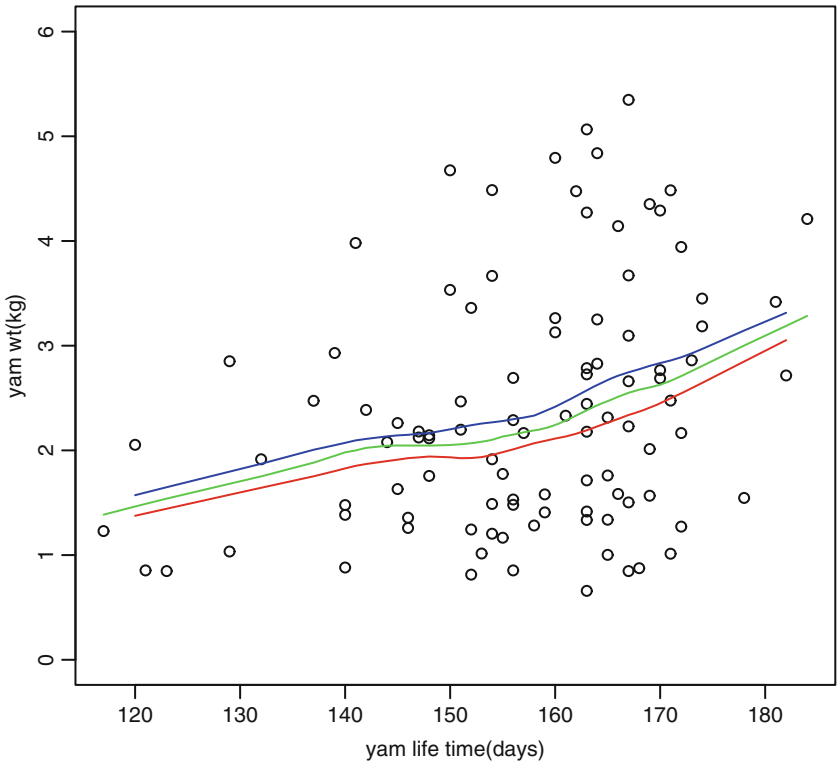


Fig. 15 Almost sure band for growth curve: Expt5

Model (34) may explain yam growth curve with sharp upward turn towards the end of plant lifetime. For $t \leq t_0$, a Gompertz model (33) may be appropriate for underground yam deposition.

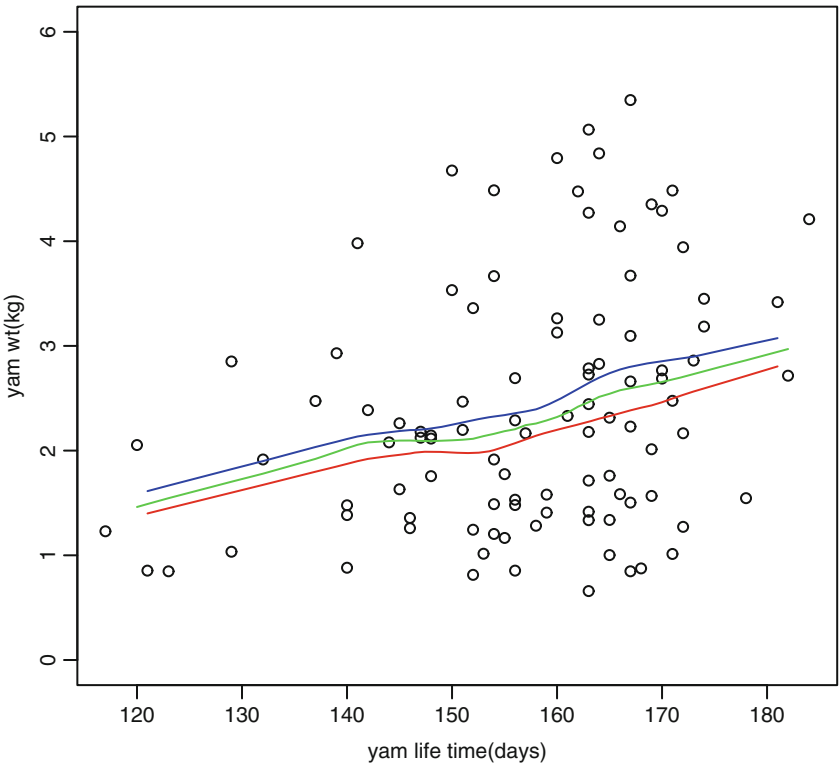


Fig. 16 Almost sure band for smoothed growth curve: Expt5

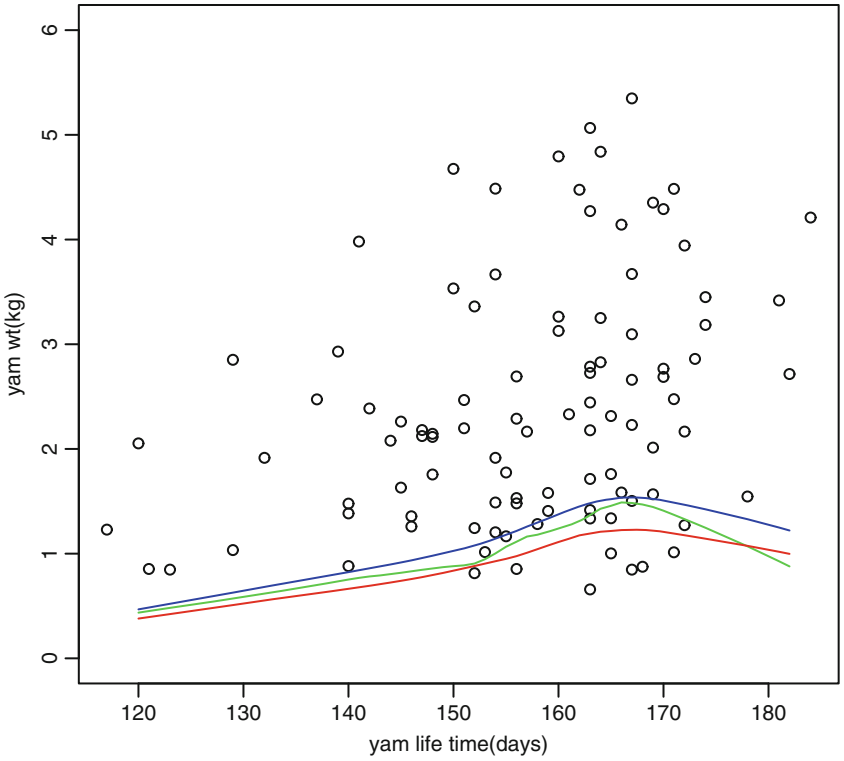


Fig. 17 Almost sure band for variance curve: Expt5

References

Bahadur RR (1971) Some limit theorems in statistics. SIAM, Philadelphia

Bose A, Dasgupta R (1994) On some asymptotic properties of U statistics and one-sided estimates. Ann Probab 22:1715–1724

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74(368):829–836

Cleveland WS (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54

Dasgupta R (1984) On large deviation probabilities of U-statistics in non iid case. Sankhyā A 46:110–116

Dasgupta R (1989) Some further results on nonuniform rates of convergence to normality. Sankhyā A 51(2):144–167

Dasgupta R (1993) Moment bounds for some stochastic processes. Sankhyā A 55:150–152

Dasgupta R (2006) Nonuniform rates of convergence to normality. Sankhyā 68:620–635

Dasgupta R (2008) Convergence rates of two sample U-statistics in non iid case. CSA Bull 60: 81–97

Dasgupta R (2013a) Non uniform rates of convergence to normality for two sample U-statistics in non IID case with applications, Chap 4. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 61–88

Dasgupta R (2013b) Optimal-time harvest of elephant foot yam and related theoretical issues, Chap 6. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 101–129
Friedrich KO (1989) A Berry-Esseen bound for functions of independent random variables. Ann Stat 17:170–183
Funk GM (1970) The probabilities of moderate deviations of U-statistics and excessive deviations of Ko1mogorov-Smirnov and Kuiper statistics. Ph.D. Dissertation, Michigan State University
Ghosh M, Dasgupta R (1978) On some nonuniform rates of convergence to normality. Sankhyā A 40:347–368
Gilat D, Hill TP (1992) One-sided refinements of the strong law of large numbers and the Glivenko-Cantelli theorem. Ann Probab 20:1109–1602
Grams WF, Serfling RJ (1973) Convergence rates for U-statistics and related statistics. Ann Stat 1:153–160
Katz ML (1963) Note on the Berry-Esseen theorem. Ann Math Stat 34:1107–1108
Lee AJ (1990) U-statistics, theory and practice. Marcel Dekker, New York
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York

Interrelationship Between Economic Growth and Income Inequality: The Indian Experience

Sattwik Santra and Samarjit Das

Abstract Kuznets's "inverted U" hypothesis postulates that inequality initially rises with economic prosperity and after reaching a certain maximum, falls thereafter. The recent literature on growth and inequality suggests a possible relationship on how income distribution affects economic growth. Keeping these two literatures in perspective, we attempt to examine the Kuznets's inverted U hypothesis in a two-way error component panel data framework. We argue that the per capita real consumption expenditure is expected to be endogenous, the feature which is surprisingly missing in the literature on Kuznets's hypothesis and try to account for this possible endogeneity between inequality and per capita real consumption expenditure. We use panel level data of Indian states on consumption expenditures. Constructing a suitable price index we establish a relationship between income inequality and per capita real consumption expenditure. To account for the possible endogeneity between these two variables, we formulate appropriate instruments for our fixed-effects model. From the findings, it is evident that the dynamics of inequality in India do not support the hypothesis as suggested by Kuznets. Indeed, we find that, initially inequality falls as the per capita real consumption expenditure rises and after reaching a certain minimum, it increases with the per capita real consumption expenditure.

Keywords Inequality • Kuznets's hypothesis • Endogeneity

JEL Classification C23 • O40

S. Santra
Centre for Training and Research in Public Finance and Policy, Center for Studies in Social Sciences, Calcutta, R-1, Baishnabghata Patuli Township, Kolkata 700094, India

S. Das (✉)
Economics Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India
e-mail: samarjit@isical.ac.in

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_6

1 Introduction

In developing countries, the effect of economic performance on inequality is a topic of interest to development economists and policy makers alike. This is so because economic growth and income distribution are closely related to people's lives and to social stability. Although the relationship between inequality and economic performance was first discussed by Kuznets (1955), the whole literature on Kuznets's hypothesis fails to account for an inherent problem of endogeneity between these two macroeconomic aggregates, which is well documented in the literature on endogenous growth theory. This present work studies this particular and important relationship in the Indian context where we put special emphasis on the issue of endogeneity. We attempt to find some valid instruments based on economic theory and then try to estimate the relationship based on instrumental variable method. We follow the exact definition of inequality as used by Kuznets as opposed to the more popular inequality measures like Gini, etc. We construct the inequality measure based on unit level data. Since unit level income data is not available in India, we use the data on consumption expenditure. As there are no price indices for rural and urban area separately, we are forced to independently construct new price indices. Here it may be noted that the popularly used Consumer Price Indices for agricultural labourers and industrial workers are only crude approximations of the rural and urban price indices, respectively. Thus we construct Laspeyres price indices for each panel level from the consumption expenditure data and use it to deflate our nominal variables.

According to Kuznets (1955), the relationship between inequality and per capita income may be described by a curve in the shape of an inverted "U". In other words, Kuznets's hypothesis advocates that, as an economy prospers income inequality first increases and after a certain "turning point" declines thereafter. Kuznets argues that this is due to a shift of labour from low-productivity to high-productivity sectors in the early stage of development, which results in an increasing disparity in wages. Later, however, the high-productivity sector comes to dominate the economy, and wage inequality decreases (for alternative explanations, see Acemoglu and Robinson 2002). There have been a large number of studies regarding the above hypothesis with contradicting and inconclusive conclusions (see Lecaillon et al. 1984 for a survey of literature). Studies in the 1960s and 1970s in general supported the hypothesis. The centerpiece of such studies comprises of articles by Ahluwalia (1976) and Ahluwalia et al. (1979). However, this hypothesis has been challenged and several empirical studies found that there is no significant relationship between inequality and per capita income (see, for example, Anand and Kanbur 1993). Li et al. (1998a, b) find that Kuznets curve works better for a cross-section of countries at a particular point of time rather than for the evolution of inequality over time within countries. There are few studies on Indian economy as well (see, for example, Andrew and Pal 2004 and the references therein). A number of econometric concerns are quite evident in cross-section based studies. In particular, cross-section based methods fail to allow for unobserved (and persistent) differences across countries/states, and they are susceptible to endogeneity biases. Therefore, it

necessitates testing the hypothesis using panel data as panel data models circumvent all these well-known problems involved in studies based on cross-sectional data.

Earlier empirical studies in the literature investigating this relationship primarily focus on unidirectional causality, that is to say, how a country's economic growth influences its income distribution. However, after the endogenous economic growth theory has been introduced since the mid-1980s, economists' interests have altered to the opposite direction, that is how income distribution affects economic growth. Recent studies on income distribution and endogenous growth by Alesina and Perotti (1993), Bertola (1991), Perotti (1993, 1994), Persson and Tabellini (1994), Persson and Guido (1994) and Forbes (2000) return to the old debate. This new literature looks at the impact of inequality on growth rather than the reverse as was the case with the earlier literature influenced by Kuznets. In fact, economic performance and income distribution both may be endogenous variables in the empirical model. Treating one as a dependent variable and other one as an independent variable could lead to biased and inconsistent estimation. In the light of these research works, the present paper tries to revisit the hypothesis for Indian economy using panel data models that allows for possible endogeneity between the variables of interest. The results suggest that after accounting for the possible endogeneity and "controlling for" a host of other factors, inequality initially falls with economic performance and after reaching a certain minimum, it increases thereafter thus suggesting a possible U relationship.

The plan of the paper is as follows. Section 2 provides a brief description of data and also discusses the econometric model. Some naked eye observations based on descriptive statistics are provided. The construction of the price indices and inequality measure are also discussed in this section. Empirical findings are depicted next in Sect. 3 which also highlights the issues that need to be addressed via suitable policy prescriptions. Section 4 concludes the paper summarising the major findings.

2 Data and Model

This study is based on the last six major rounds of survey on "Household Consumer Expenditure" provided by the National Sample Survey Organization (NSSO) of India. The data covers 32 states and Union territories[1] of India and was collected in the years 1987–1988 (43rd round), 1993–1994 (50th round), 1999–2000 (55th round), 2004–2005 (61st round), 2009–2010 (66th round) and 2011–2012 (68th round) comprising of various socio-economic characteristics of a household. Apart from the household specific characteristics, data is also provided on the localization of the sampled households. The localization of a sampled household includes the

[1]The states and the Union territories are: Andaman & Nicobar Islands, Andhra Pradesh, Arunachal Pradesh, Assam, Bihar, Chandigarh, Dadra & Nagar Haveli, Daman & Diu, Delhi, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu & Kashmir, Karnataka, Kerala, Lakshadweep, Madhya Pradesh, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Orissa, Pondicherry, Punjab, Rajasthan, Sikkim, Tamil Nadu, Tripura, Uttar Pradesh, West Bengal.

sector (rural or urban), state, region (a subdivision of each state based on certain broad geographical features) and the district in which the household resides. For our purpose, we have considered only the sector (rural or urban), and the state of the sampled households.

For our present study, we have considered the sampled households' principal occupation, social group, religion, amount of land possessed and the amounts of various items consumed together with the associated expenditure adjusted to a suitable reference period. Apart from the household specific observations, we have also considered the age, sex and education of each household members. Using this data, we have constructed various state-sector specific aggregates which constitute our individual panel unit. With the data in hand, we have a panel with 62 state-sector combinations ($N = 62$) and six time periods ($T = 6$) with some gaps in the availability of the data, resulting in a dataset having 367 observations.

These aggregates of course include the two variables of interest: a state-sector specific inequality index and the corresponding aggregate per-capita real consumption expenditure. The index of inequality (following Kuznets 1955) is calculated as the difference between the aggregate share of monthly consumption expenditures of the households above the top (fifth) consumption expenditure quintile to the aggregate share of monthly consumption expenditures of the households below the bottom (first) quintile residing in the state and sector of interest. As briefed in the introduction, to be thorough with our analysis, we compute our own price indices suited to the panel structure of the data. For any given time point and panel unit, we aggregate the quantity and value of consumption of all commodities excluding the durables. From this data, we readily compute the item wise prices and use it to construct our Laspeyre's price index with the year 2004 as the base period. We use this index to obtain the state-sector specific per-capita real expenditure.

Various statistics related to the distributions of these two constructed variables across the combinations of the state-sector is tabulated in Tables 1, 2, 3, 4, 5, 6, 7 and 8 for each of the six NSSO rounds. Various graphical representations of these two variables are also depicted in Figs. 1 and 2. Figure 1a, and b depict how the rural and urban income distributions have progressed over time respectively. It is quite evident that mean incomes both for rural and urban area have more or less increased over time. Tables 3 and 4 suggest that both for rural and urban area, each section (quarter) of the population has been better off in terms of real income. Moreover, there is

Table 1 Descriptive for the state wise inequality measure for rural India

Year	Mean	Standard deviation	First quantile	Median	Third quantile	Skewness	Kurtosis
1987	0.3337	0.0141	0.3282	0.3353	0.3435	−1.6487	10.3792
1993	0.3199	0.0199	0.3082	0.3186	0.3382	−0.6923	3.2988
1999	0.2979	0.0216	0.2805	0.3033	0.3172	−0.2876	2.6334
2004	0.3175	0.0267	0.3060	0.3265	0.3315	−0.6073	3.6419
2009	0.2991	0.0271	0.2953	0.3023	0.3071	−0.3380	3.1281
2011	0.2979	0.0242	0.2897	0.3069	0.3111	−1.0239	3.4919

Table 2 Descriptive for the state wise inequality measure for urban India

Year	Mean	Standard deviation	First quantile	Median	Third quantile	Skewness	Kurtosis
1987	0.3506	0.0229	0.3423	0.3475	0.3661	−0.4783	6.5376
1993	0.3488	0.0229	0.3326	0.3446	0.3571	0.1106	3.5664
1999	0.3348	0.0316	0.3102	0.3370	0.3507	0.4697	2.9214
2004	0.3660	0.0233	0.3638	0.3714	0.3774	−1.6986	6.7518
2009	0.3643	0.0282	0.3447	0.3702	0.3949	−0.5439	3.5767
2011	0.3605	0.0269	0.3419	0.3582	0.3850	−0.9000	4.6525

Table 3 Descriptive for the state wise monthly per-capita real consumption expenditure for rural India

Year	Mean	Standard deviation	First quantile	Median	Third quantile	Skewness	Kurtosis
1987	393.3884	112.5156	328.2205	368.6754	432.2861	2.0373	8.4451
1993	393.9833	98.0205	315.9030	363.7263	447.3277	1.8717	9.2935
1999	492.7826	95.2384	438.8170	492.4627	507.5273	1.6423	7.2100
2004	586.0752	134.4901	539.2877	575.6547	601.6321	1.8084	6.3929
2009	553.6098	173.4677	407.6331	538.8271	606.3124	1.4121	4.6791
2011	600.1614	196.6000	425.1906	574.3254	722.5880	1.1302	3.8018

Table 4 Descriptive for the state wise monthly per-capita real consumption expenditure for urban India

Year	Mean	Standard deviation	First quantile	Median	Third quantile	Skewness	Kurtosis
1987	715.7302	187.8960	678.0958	718.5422	763.2227	2.1350	10.5304
1993	749.5914	171.7440	683.2816	753.5291	833.4937	1.2849	7.6688
1999	849.1603	197.7756	710.2099	883.4963	933.9968	0.3859	4.5695
2004	1108.2491	178.7918	944.5705	1158.9690	1228.4292	−0.0175	3.5689
2009	1150.3853	251.6907	968.5101	1240.3491	1325.6998	−0.3225	4.2712
2011	1202.7876	271.0289	1054.9183	1319.5826	1406.4446	−0.6413	2.6433

Table 5 Intertemporal Spearman rank correlation for the state wise inequality measure for rural India (figures in brackets indicate p-values)

Year	1987	1993	1999	2004	2009	2011
1987	1.0000					
1993	0.4302 (0.0157)	1.0000				
1999	0.5758 (0.0007)	0.3242 (0.0752)	1.0000			
2004	0.3262 (0.0733)	0.2786 (0.1291)	0.5415 (0.0017)	1.0000		
2009	0.3629 (0.0448)	0.0794 (0.6710)	0.5488 (0.0014)	0.7093 (0.0000)	1.0000	
2011	0.2794 (0.1279)	0.3702 (0.0404)	0.5044 (0.0038)	0.6093 (0.0003)	0.5665 (0.0009)	1.0000

Table 6 Intertemporal Spearman rank correlation for the state wise inequality measure for urban India (figures in brackets indicate p-values)

Year	1987	1993	1999	2004	2009	2011
1987	1.0000					
1993	0.4359 (0.0142)	1.0000				
1999	0.0879 (0.6382)	0.4242 (0.0174)	1.0000			
2004	0.5895 (0.0005)	0.4431 (0.0125)	0.4238 (0.0175)	1.0000		
2009	0.4532 (0.0105)	0.3935 (0.0285)	0.4681 (0.0079)	0.8173 (0.0000)	1.0000	
2011	0.3137 (0.0857)	0.4411 (0.0130)	0.4536 (0.0104)	0.7081 (0.0000)	0.7137 (0.0000)	1.0000

Table 7 Intertemporal Spearman rank correlation for the state wise monthly per-capita real consumption for rural India (figures in brackets indicate p-values)

Year	1987	1993	1999	2004	2009	2011
1987	1.0000					
1993	0.7774 (0.0000)	1.0000				
1999	0.8347 (0.0000)	0.7423 (0.0000)	1.0000			
2004	0.7786 (0.0000)	0.6262 (0.0002)	0.8565 (0.0000)	1.0000		
2009	0.7855 (0.0000)	0.9327 (0.0000)	0.7677 (0.0000)	0.6673 (0.0000)	1.0000	
2011	0.7625 (0.0000)	0.9202 (0.0000)	0.7133 (0.0000)	0.6468 (0.0001)	0.9633 (0.0000)	1.0000

Table 8 Intertemporal Spearman rank correlation for the state log monthly per-capita real consumption for urban India (figures in brackets indicate p-values)

Year	1987	1993	1999	2004	2009	2011
1987	1.0000					
1993	0.8093 (0.0000)	1.0000				
1999	0.7065 (0.0000)	0.7173 (0.0000)	1.0000			
2004	0.7726 (0.0000)	0.6677 (0.0000)	0.8165 (0.0000)	1.0000		
2009	0.6798 (0.0000)	0.7919 (0.0000)	0.6234 (0.0002)	0.6044 (0.0003)	1.0000	
2011	0.7347 (0.0000)	0.8052 (0.0000)	0.6383 (0.0001)	0.6597 (0.0001)	0.9722 (0.0000)	1.0000

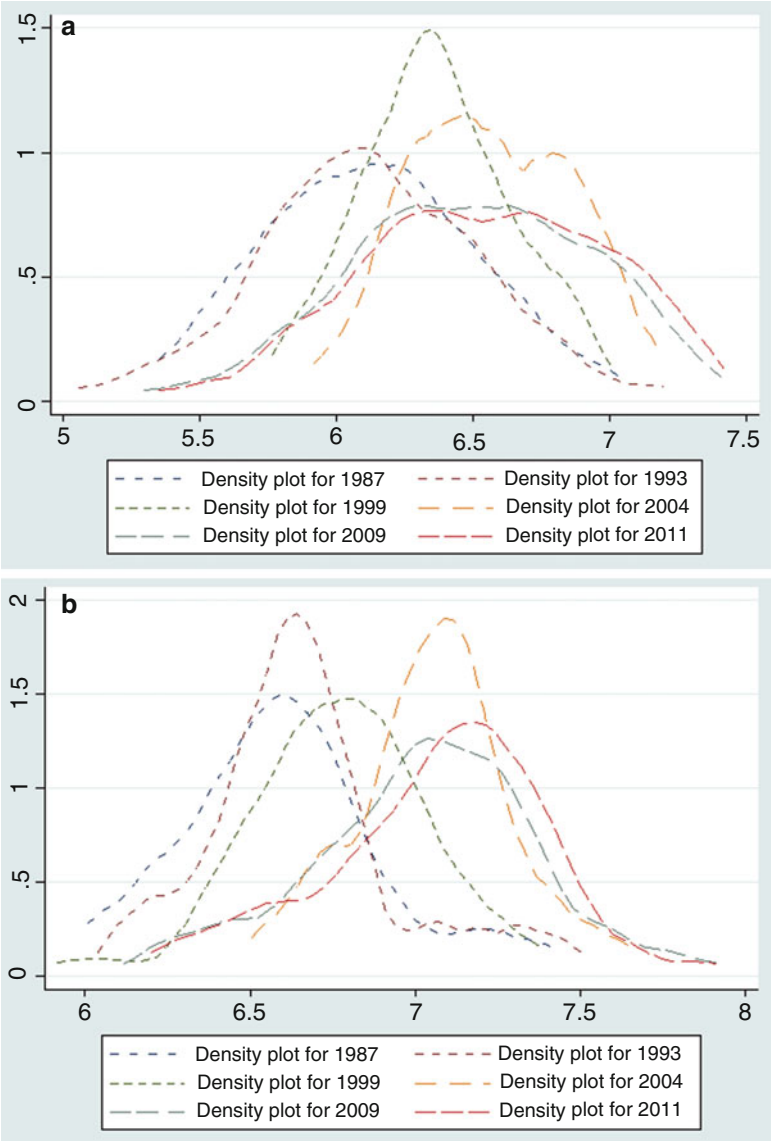


Fig. 1 (**a**) Density plot of log monthly per-capita real consumption expenditure for rural India for the various rounds. (**b**) Density plot of log monthly per-capita real consumption expenditure for urban India for the various rounds

some amount of churning within the distribution. The extent of this churning can be summarised by using Spearman rank correlation (given in Tables 5, 6, 7 and 8). These tables show that the churning is more prominent for inequality as compared to per capita consumption expenditure.

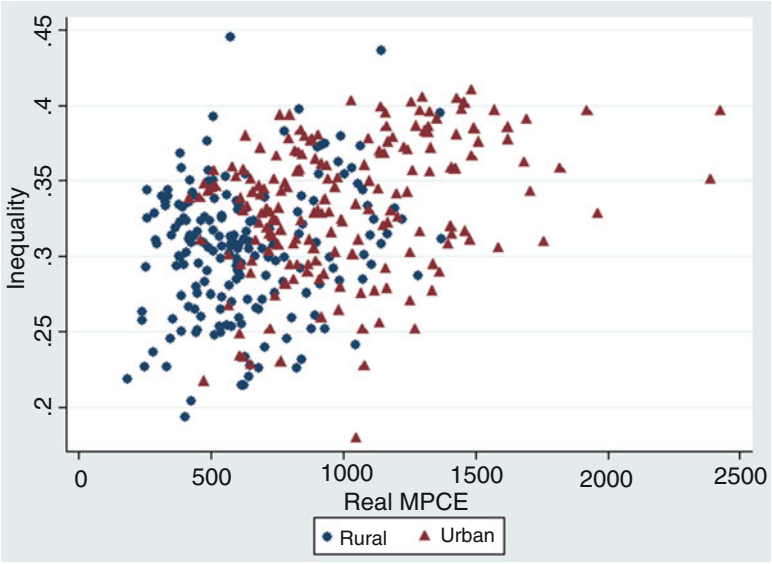


Fig. 2 Overall scatter of the inequality against the log monthly per-capita real consumption expenditure with different markers differentiating the rural and urban state combinations

Apart from these two variables, we have also considered other quantities specific to each panel unit. These include the total land possessed by the households, the amount of cultivable land possessed by the households, the proportion of households belonging to three broad principal occupation types (professional, technical, administrative, executive, managerial and related workers dubbed as occupation group 1, clerical, sales, service, farmers, fishermen, hunters, loggers, production and related works, transport equipment operators and labourers clubbed into occupation group 2 and workers not classified by occupations including unemployed labourers, grouped as 3), the overall population, the working population (defined as the number of people with ages between 18 and 62), the number of people belonging to four different classes of education (illiterate, literate but below secondary level of education, secondary and higher secondary level of education and above secondary level of education), the proportion of people belonging to the various social (scheduled tribes, scheduled castes and others) and religious (Hinduism, Islam, Christianity, Sikhism, Jainism, Buddhism, Zoroastrianism and others) groups. These variables are used either in their level values or as ratios with other variables as controls for our empirical model or as instruments to the per-capita real consumption expenditure.

The basic form of the Kuznets's hypothesis suggests a quadratic relation between income inequality and economic performance, in which inequality increases with real income at early stages and starts declining after reaching a peak. A natural specification in panel data with a very general form can be hypothesised as:

$$Ineq_{it} = \beta_0 + \beta_1 Y_{it} + \beta_2 Y_{it}^2 + \alpha_i + \lambda_t + \theta Z_{it} + \epsilon_{it} i = 1..N; t = 1, 2, 3 \quad (1)$$

where Y_{it} denotes the per capita real monthly consumption expenditure transformed in logarithmic scale, Z_{it} is a vector of controls, α_i and λ_t are the state and time specific unobserved heterogeneity effects, respectively; and u_{it} is the disturbance term. Initially, we have considered a model where we do not account for the possible endogeneity between inequality and per-capita real consumption expenditure. However, Hausman test as presented in Table 10 strongly suggests for the use of instruments for our fixed effect model.

In order to find suitable instruments for the above empirical model, we turn to a constant returns to scale production function for possible indications. A constant returns to scale production function is given by: $Y = F(K, H, L)$,where Y denotes the aggregate production, K the physical capital inputs, H the human capital inputs and L denotes the labour employed in the production. Because of constant returns to scale we can rewrite the above relationship as: $y = F(k, h, 1)$ where the capitalization is removed to indicate the respective variables as a ratio of the amount of labour employed. Based on this simple yet general production formulation, for any panel unit of a particular round we take the respective total land, cultivated land and the number of people belonging to the two highest categories of education qualifications as proportions of the working population as the suitable instruments for the per capita real monthly consumption expenditure. The results from the various model specifications illustrated thus, are detailed in the next section.

3 Empirical Findings

At the outset, it is worthwhile to mention that, finding suitable well-specified model is of utmost importance. To this end, we have tried several specifications for the Eq. (1) including the cubic and fourth order income variables on the right-hand side. It turns out that both cubic and fourth order variables are insignificant. Throughout the whole exercise, we have considered two-way error component model to take into account both unobserved individual and time specific effects. It may be noted that the panel fixed effect model is considered without loss of generality. The individual heterogeneity parameter, α_i, is invariant across the time and accounts for state-sector specific unobserved structural heterogeneity, state and rural–urban specific fiscal, monetary and industrial policies, labour laws, etc. On the other hand, the unobserved time effect, λ_t, is invariant across the panel units and accounts for time-specific effects, such as common shocks, the impact of central fiscal, monetary and industrial policies, labour laws, etc.

Now we turn to examine Kuznets's hypothesis. Here it may be noted that the economic theory on the growth-inequality nexus is quite inconclusive. The reason of positive relationship may be found in Aghion et al. (1999). The primary reason for such positive relationship is that if the growth rate is positively related to the proportion of national income that is saved, more unequal economies are bound to grow faster than economies with a high level of income distribution, since the marginal propensity to save of the rich is higher than that of the poor. In a

Table 9 Fixed effect regression results

Dependent variable: inequality			
Independent variables	Coefficient	Robust standard error	P-value
ln(MPCE)	−0.4888***	0.1196	0.0000
ln(MPCE) squared	0.0411***	0.094	0.0000
Social group			
1	0.1377*	0.0741	0.0640
2	0.0908	0.0691	0.1910
Religion			
2	0.0240	0.0779	0.7580
3	0.0786	0.1467	0.5920
4	0.1173	0.2338	0.6160
5	−0.8760	0.5546	0.1150
6	0.4351**	0.1865	0.0200
7	3.6676	4.4579	0.4110
8	(Omitted)		
9	0.3571	0.2228	0.1100
Occupational inequality	−0.0787	0.0581	0.177
R square	0.8483		

Note that *, ** and *** indicate significance at 1, 5 and 10 % levels, respectively

democratic political set-up, the inverse relationship between income inequality and growth would be expected to be stronger if the income distribution is tilted to the left, giving lower-income groups more political power (Persson and Tabellini 1994). From Table 9, it is found that the slope coefficient corresponding to the per-capita real monthly consumption expenditure variable is significantly negative. On the other hand, the slope coefficient corresponding to square of the variable is significantly positive. The signs of both the slope coefficients are contrary to the conjecture of Kuznets. Thus, the empirical results indicate that in the context of India, the relationship between inequality and economic well being measured by per-capita real monthly consumption expenditure exhibits a U relationship contradicting Kuznets's inverted U hypothesis. We have also used several forms of robust standard errors. It is needless to mention that overall conclusion remains unchanged across various forms of standard errors. It may be interesting to look into the plot given in Fig. 3 which is derived from the panel regression after removing the contributions of control variates. The plot (Fig. 3) visibly and unambiguously shows that growth-inequality relationship is in U shape and strongly evidences against the Kuznets's hypothesis. The empirical findings and the fitted plot strongly signify the presence of some nontrivial dynamics of economic growth and income distribution.

It may be worthwhile to scrutinise the above findings in the light of endogeneity. As we have already mentioned earlier that there is a possibility of presence of endogeneity. Studies by Alesina and Perotti (1993), Bertola (1991), Perotti (1993, 1994), and Persson and Tabellini (1994) and Forbes (2000) look at the impact of inequality on growth rather than the reverse as was the case with the earlier literature influenced by Kuznets (1955). In fact, economic performance and income

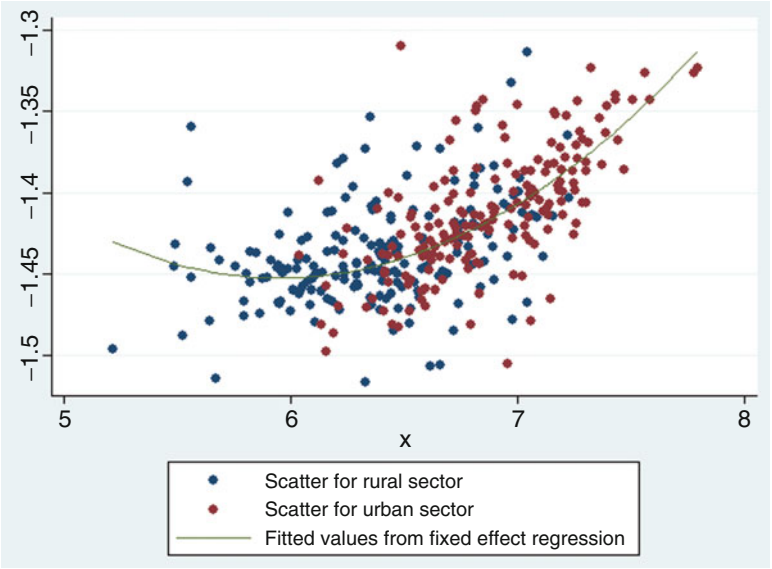


Fig. 3 Plot of the inequality against net of the effects of control variates against the log monthly per-capita real consumption expenditure with different markers differentiating the rural and urban state combinations. The fitted line corresponds to the predicted values of inequality constructed from the log monthly per-capita real consumption expenditure and its square as depicted in Table 9

distribution both may be endogenous variables in the empirical model. Treating one as a dependent variable and other one as an independent variable could lead to biased and inconsistent estimation. In the light of these research works, it is necessary to revisit the above findings incorporating this endogeneity issue. To this end, we apply the Hausman test for endogeneity. The results given in Table 10 strongly suggest that growth variable is an endogenous variable. The weak identification test (Cragg-Donald Wald F statistic) statistic is far away from the generally accepted critical value of 10, indicating that the considered instruments are quite strong. Furthermore, Hansen's J test for over-identification strongly suggests that the instruments are valid. The IV-based parameter estimates are given in Table 10. The results corresponding to Table 10 strongly corroborate with the findings of Table 9. In other words, even after taking care of endogeneity, Kuznet's hypothesis fails to exist for the Indian context. The plot (Fig. 4), which is derived from the panel instrumental variable regression after removing the contributions of control variates, visibly and unambiguously shows that growth-inequality relationship is in U shape and strongly evidences against the Kuznets's hypothesis.

Our finding has definite policy implications. Kuznets's argument behind the inverted "U" association between income inequality and real income follows from the nature of adoption of new technology. Technological progress which results in the growth of an economy initially favours only a few, therefore augmenting inequality, but as the technology is adapted more and more, the inequality decreases.

Table 10 Fixed effect regression using instrumental variables estimated using iterated GMM

Dependent variable: inequality Independent variables	Coefficient	Robust standard error	P-value
ln(MPCE)	−0.5126*	0.2880	0.0760
ln(MPCE) squared	0.0431*	0.0222	0.0530
Social group			
1	0.1496*	0.0777	0.0550
2	0.0673	0.0683	0.3250
Religion			
2	0.0258	0.0787	0.7430
3	0.0356	0.1401	0.8000
4	0.1732	0.2196	0.4310
5	−1.1141**	0.5115	0.0300
6	0.4341**	0.1878	0.0210
7	3.3309	4.4351	0.4530
9	0.3175	0.2212	0.1520
Occupational inequality	−0.0847	0.0715	0.2370
Uncentred/centred R2	0.4559		
Hausman test for endogeneity	42.1100		
$Chi^2(16)$ P-value	0.0004		
Weak identification test (Cragg-Donald Wald F statistic):	21.2360		
Hansen J statistic (overidentification test of all instruments)	3.9470		
$Chi^2(3)$ P-value	0.2672		

Note that *, ** and *** indicate significance at 1, 5 and 10 % levels, respectively

In other words, this argument presupposes a dynamics in technological progress which ultimately leads an economy to "mends its self" ensuring an equitable distribution of resources. Our empirical results calls into question the validity of such dynamics in the context of India and instead vindicates a tradeoff between growth and inequality. From a normative standpoint, since a significant amount of the population still live in poverty, there is a need for policy interventions to tackle the socio-economic fallout of inequality which seems to persist in India despite the growth of the economy. But this does not necessarily entail the cost of a lower rate of growth for the economy. The literature on endogenous growth has produced models (see, for example, Sarkar 1998) that demonstrate how the size of the middle class plays an extremely important role in determining the rate of growth of an economy. In our exercise, the lower middle class registers the least amount of inequality. This opens up the possibility of government policies aimed towards the sufficiently large Indian middle class which might lead to a sustainable growth prospect for India in the near future.

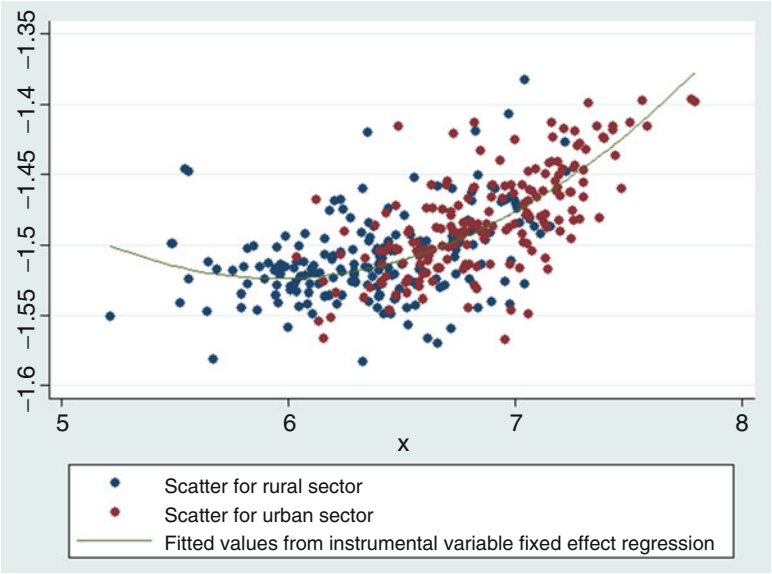


Fig. 4 Plot of the inequality against net of the effects of control variates against the log monthly per-capita real consumption expenditure with different markers differentiating the rural and urban state combinations. The fitted line corresponds to the predicted values of inequality constructed from the log monthly per-capita real consumption expenditure and its square as depicted in Table 10

4 Concluding Remarks

In this paper we test the Kuznets's U curve hypothesis with balanced panel data of the 32 states and Union territories[2] of India covering the time span of 1987–2011 comprising of various socio-economic characteristics of a household. Based on the estimated model, we find that one should reject Kuznets's U curve hypothesis. Income inequalities are highly explained by level of real income along with state and time specific factors. The fitted curve evidences that initially inequality decreases as income level increases up to a limiting point, where after it starts to increase as income increases further.

Future research may commence to unearth the more specific causes for such findings. It may also be useful to employ various other measures of indicators of economic performance and also various income inequalities along with other dimensional inequalities to examine robustness of such empirical findings.

[2]The states and the Union territories are: Andaman & Nicobar Islands, Andhra Pradesh, Arunachal Pradesh, Assam, Bihar, Chandigarh, Dadra & Nagar Haveli, Daman & Diu, Delhi, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu & Kashmir, Karnataka, Kerala, Lakshadweep, Madhya Pradesh, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Orissa, Pondicherry, Punjab, Rajasthan, Sikkim, Tamil Nadu, Tripura, Uttar Pradesh, West Bengal.

References

Acemoglu D, Robinson J (2002) The political economy of Kuznets curve. Rev Dev Econ 6: 183–203

Aghion P, Caroli E, Garcia-Penalosa C (1999) Inequality and economic growth: the perspective of the new growth theories. J Econ Lit 37:1615–1660

Ahluwalia M (1976) Inequality, poverty and development. J Dev Econ 3:307–342

Ahluwalia M, Carter N, Chenery H (1979) Growth and poverty in developing countries. J Dev Econ 6:299–341

Alesina A, Perotti R (1993) Income distribution, political instability and investment. NBER working paper no 4486

Anand S, Kanbur S (1993) The Kuznets process and the inequality-development relationship. J Dev Econ 40:25–52

Andrew M, Pal S (2004) Relationships between household consumption and inequality in the Indian states. J Dev Stud 40:65–90

Bertola G (1991) Market structure and income distribution in endogenous growth models. NBER working paper no 3851, National Bureau of Economic Research, Cambridge

Forbes K (2000) A reassessment of the relationship between inequality and growth. Am Econ Rev 90:869–887

Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45:1–28

Lecaillon J, Paukert F, Morrisson C, Germidis D (1984) Income distribution and economic development: an analytical survey. International Labour Office, Geneva

Li S, Squire L, Zou H (1998a) Income inequality is not harmful for growth: theory and evidence. Rev Dev Econ 2:318–334

Li H, Squire L, Zou H (1998b) Explaining international and inter temporal variations in income inequality. Econ J 108:26–43

Perotti R (1993) Political equilibrium, income distribution, and growth. Rev Econ Stud 60:755–776

Perotti R (1994) Income distribution and investment. Eur Econ Rev 38:827–835

Persson T, Guido T (1994) Is inequality harmful for growth? Theory and evidence. Am Econ Rev 84:600–621

Persson T, Tabellini G (1994) Is inequality harmful for growth? Am Econ Rev 84:601–621

Sarkar A (1998) Endogenous growth and the size of the market. Keio Econ Stud 35(1):29–44

Growth Curve Reconstruction in Damaged Experiment via Nonlinear Calibration

Ratan Dasgupta

Abstract Consider the problem of estimating growth curve of a bulb crop over time in an agricultural experiment. Harsh environment in experimental land of farm may result in an incomplete and damaged data set. Under certain mild assumptions we model the original data by nonlinear calibration, estimate the growth curve and derive almost sure confidence band for the curve by adopting a technique of Dasgupta (Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York, 2015). A yield-environment model is proposed and properties of the maximum likelihood estimator therein are investigated. A technique of computing derivative of response curve by nonparametric regression is shown to be strongly consistent for higher order derivatives. Growth data set on a bulb crop is analysed. First derivative, proliferation rate and second derivative curves of calibrated growth for garlic crop are estimated. Associated errors in estimation of the growth curve, its derivative and proliferation rate from composite function, attributed to individual component functions are studied. To a first degree of approximation, error in estimation is symmetric in individual errors, but asymmetric in individual functions.

Keywords Growth curve • Bulb crop • Allium sativum • Lowess

MS Subject classification: Primary: 62J02, secondary: 62P10.

1 Introduction

Garlic (*Allium sativum*) is a year round crop grown in moderate climates. Garlic plant cannot withstand extreme temperature. Exposure to dormant cloves or young plants to temperature of around 20 °C or lower for a period hastens subsequent bulbing. In dry weather conditions, with increase in evaporation rate during Indian summer, plant growth may be substantially affected. The maximum summer

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_7

temperature can be as high as 47 °C in Jharkhand, India. We consider estimating growth curve of garlic yield in an experiment conducted during February–May in the farm of Indian Statistical Institute (ISI), Giridih; Jharkhand. Hot summer in later period of cultivation caused dehydration of the crop, making estimation of the growth curve from damaged experiment a difficult task. We estimate the growth curve from such experiment by nonlinear calibration of available data on yield. Techniques developed involving composite function may be used for estimating growth curve arising in similar situations. Associated errors in estimation by such calibration are derived in large samples. We obtain almost sure confidence band for the growth curve based on one-sided estimators, following a technique of Dasgupta (2015). A simple model for proportionate gain in crop yield with dominant environment effect is proposed.

Previous studies, experimental set-up and data. Growth of garlic plants are studied in Diriba-Shiferaw et al. (2013). Apart from culinary use, garlic has medicinal value in reducing blood glucose, antibacterial effects, etc., see, e.g., Jelodar et al. (2005) and Matthew et al. (2007). Tissue culture produces virus-free clones that are more productive and keep the desired traits of the cultivar of garlic. Scotton et al. (2013) analysed the in vitro regeneration of eight marketable cultivars of garlic using root segments as explants.

In the present study, one hundred garlic clove seedlings were planted in an experimental plot at ISI Giridih farm on 12 February 2014, in winter season. The plot had topsoil eroded, this is part of a barren land having sandy soil composition mixed with "dhoincha" (Sesbania bispinosa) plant compost manure, so as to make survival of plants easier in the unfertile plot of land. In each row there were ten plantations. Plant to plant distance was 15 cm. There were ten rows; distance between rows was 30 cm. A little bit of vermicompost manure was also provided in the experimental plot. Out of 100 plantations, 87 resulted in healthy garlic plants having positive yields on maturity. For remaining 13 plants, there were no yields. In Fig. 1 we plot the plant lifetime (in day) and weight of 87 dehydrated crops (in gram). A nonparametric regression curve for mean response is also shown. Garlic yield data indicate high scattering especially for higher values of plant lifetime.

The paper is arranged as follows. In Sect. 2 we analyse the data set on growth experiment conducted in Giridih, where summer maximum temperature reached as high as 43.5 °C during the experiment. We obtain growth curve for garlic before and after calibration of data obtained from the experiment conducted in this extreme environment. Following the technique described in Dasgupta (2015), almost sure band of the growth curve is derived from the properties of one sided estimators, where convergence to the parameter is restricted to a particular direction viz., from above/below; see Gilat and Hill (1992). In Sect. 3, we model proportionate gain in yield with dominant environmental affect, study properties of the proposed model and analyse observed data. The model explains some uncommon behaviour of the maximum likelihood estimate. Consistency of derivative estimates are shown in

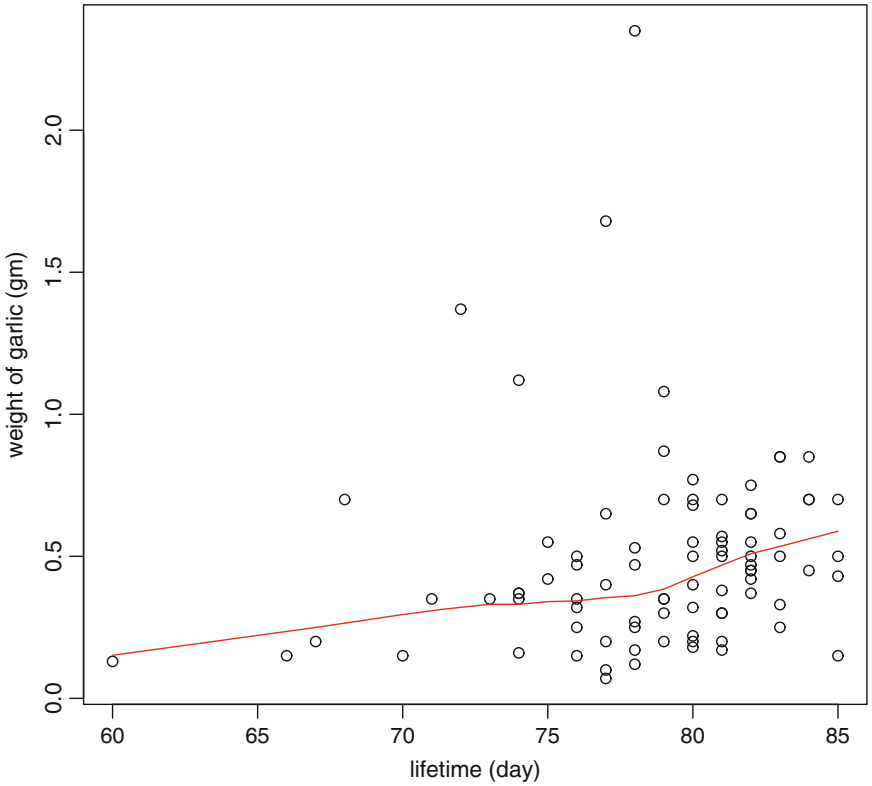


Fig. 1 Growth curve of a bulb crop (dehydrated). Growth curve of garlic is estimated by lowess regression with $f = 2/3$. The curve shows an upturn after a lifetime of 79 days

Sect. 4, these are computed from yield data. Section 5 deals with accuracy of growth curve estimation and resultant proliferation rate based on composite function in large samples.

2 Data Analysis and Some Comments

We start with available data on dehydrated garlic. Growth curve by lowess regression, see Cleveland (1981), with $f = 2/3$ for yield data in Fig. 1 shows a marked increasing trend after a lifetime of 79 days.

To estimate the garlic weights before dehydration, we observed that the outer structures of the dehydrated crop remained intact. Thin and dry membranous outer scales still resemble original shape of the crop that developed underground, making it possible to have an estimate for volume of the crop.

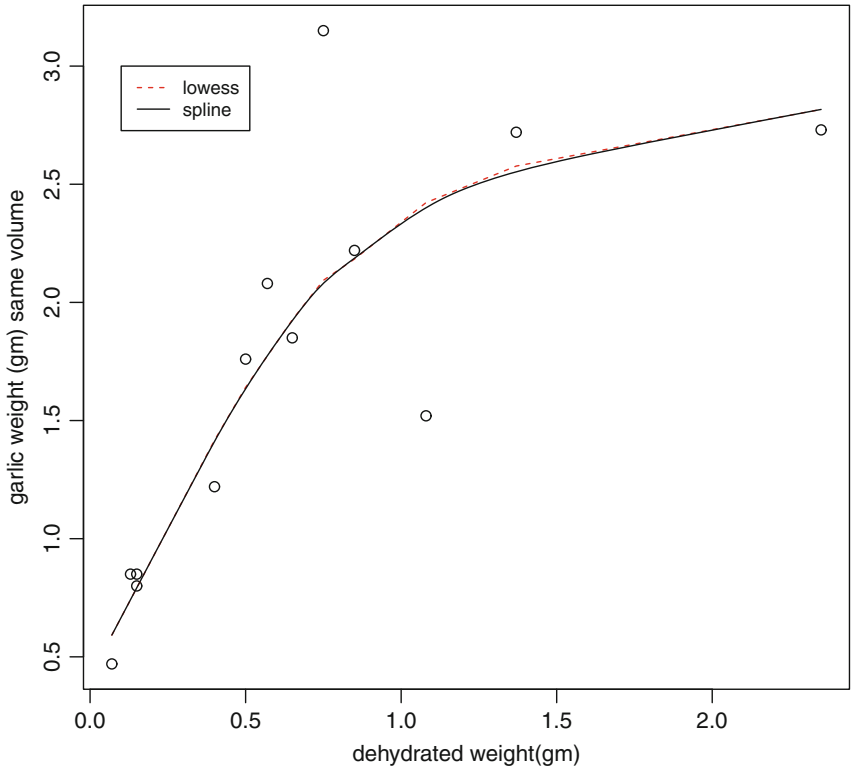


Fig. 2 Calibration factor obtained by lowess and spline regression. Calibration factor shown by lowess and spline regression are almost overlapping

Next we find weight of healthy garlic corresponding to similar volume of dehydrated garlic, dry garlic being selected evenly from the weight range of crop yield. Lowess regression with $f = 0.8$ based on weight of 13 dried garlic ($\approx$15 %) selected from yield observations, versus weight of healthy garlic of similar volume are shown in Fig. 2. Spline regression, with shape parameter 1, is also shown in the same figure; smoothing spline is almost overlapping with the lowess regression. One may select either of the curves for a valid calibration of data points.

Concave graphs of Fig. 2 indicate that for a garlic of small weight having small volume, dehydration is relatively high resulting in more weight loss. For large weights, losses due to dehydration seem less. This is in conformity with the fact that harsh and extreme summer affects the bulb of less weight through hot soil. For heavier bulbs, hot soil may not be able to affect innermost core region of crop, saving underground garlic from much weight loss.

For each weight of dried garlic, a calibrating factor is now available from the lowess graph to predict the original weight of garlic yield, which may be obtained by multiplying the dehydrated weight with corresponding calibrating factor. Such calibration techniques are also used in Dasgupta and Pan (2015).

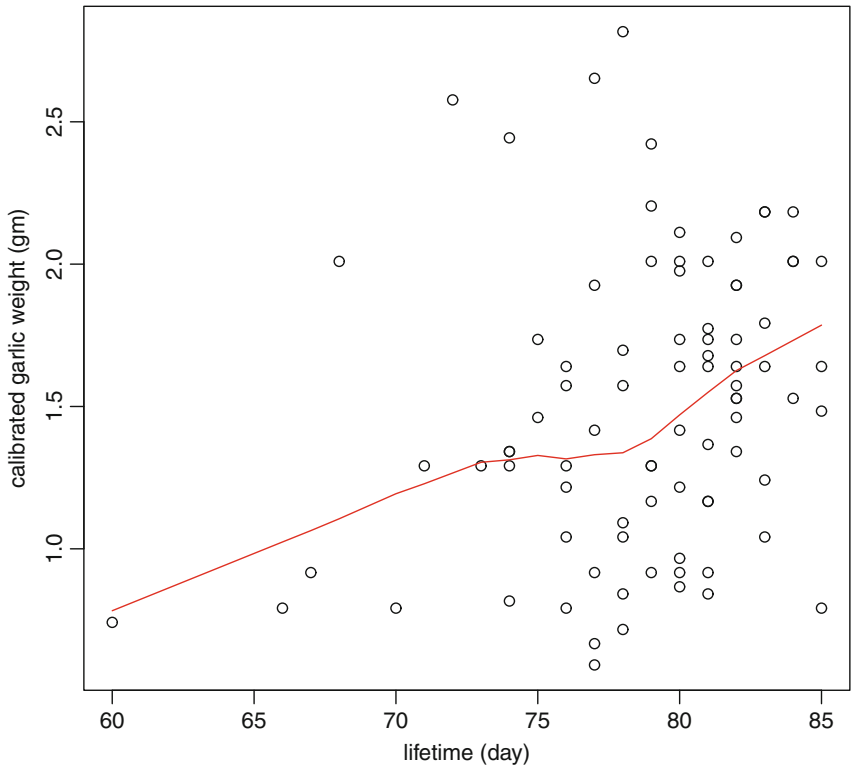


Fig. 3 Growth curve (lowess) based on calibrated yield. Growth curve of calibrated garlic weight is estimated by lowess regression with $f = 2/3$. The curve shows an upturn after a lifetime of 78 days

The scatter diagram of plant lifetime versus modified weight of the crop is shown in Fig. 3. Lowess regression (with $f = 2/3$) to the scatter shows a distinctive increasing trend in the growth curve after a lifetime of 78 days. Curves shown in Figs. 1 and 3 show similarity, the latter curve is pulled up, especially towards large values of lifetime, due to calibration. Presence of spikes is also observed in the upward turn of the growth curves towards end of the lifetime. It appears that the plants may have a hunch when their lifetime is going to be over, as a result plants may accumulate food in the bulb faster towards end.

To obtain an estimate of percentage gain in cultivation of the crop in Giridih, we plot the scatter diagram of planted initial weight versus modified weight of the crop yield in Fig. 4. The slope of the least square regression line passing through origin is 1.67, whereas a ratio estimate of the multiplicative factor obtained as ratio of total yield by total initial weight is 1.71.

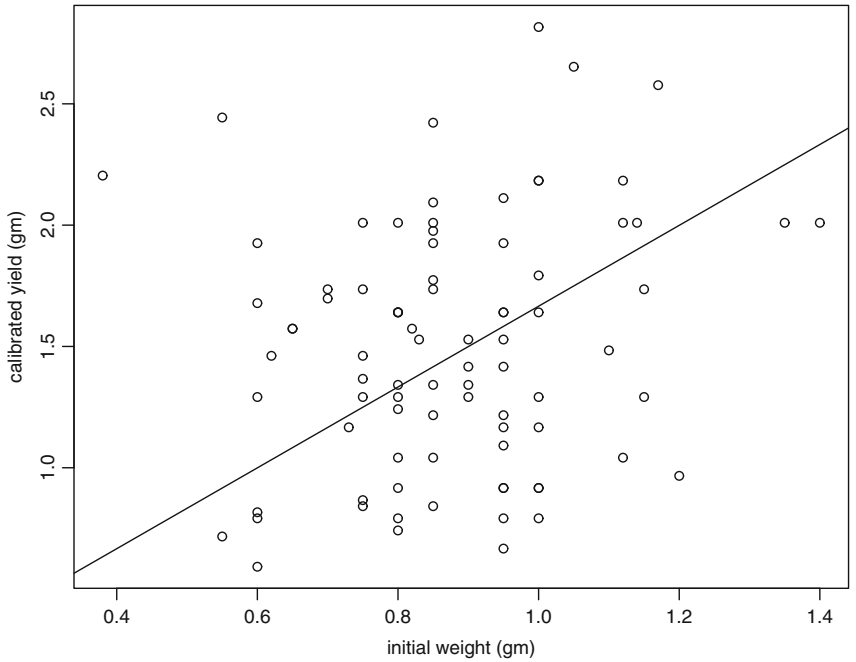


Fig. 4 Least square regression line through origin for initial vs final weight. Least square regression line through $(0, 0)$ for initial (seed) and final (calibrated) weight of garlic is shown. Slope of the line is 1.67

Thus, on an average, the gain is about 70 % for garlic cultivation in Giridih farm.

In Fig. 5 we obtain almost sure confidence band for the growth curve based on dehydrated garlic data, following the technique described in Dasgupta (2015). At each observed point on lifetime t of garlic plant, we compute the perturbation $\frac{1}{4n^{\alpha}} \sum_{i,j=1}^{n} |X_i - X_j|, \alpha \in (2, 2.5)$ as proposed in Gilat and Hill (1992). We take two immediate points above and below a lifetime t and consider the corresponding garlic yield $X = X_i$ to compute the perturbation part with $n = 3, \alpha = 2.25$. Exception is made for the lowest time point, where we consider another additional repeat point to draw the band from start. The points so obtained by addition/subtraction from lowess curve fall above/below the lowess growth curve, and the points falling in a particular side of the curve are again lowess smoothed to obtain approximate a.s. upper/lower confidence bands as seen in Fig. 5. The curve in blue/black, lying on the top/bottom of the central curve in red for garlic growth, is the upper/lower confidence curve.

Confidence band based on calibrated weights of garlic is shown in Fig. 6. The almost sure bands cover the central curves in both the figures.

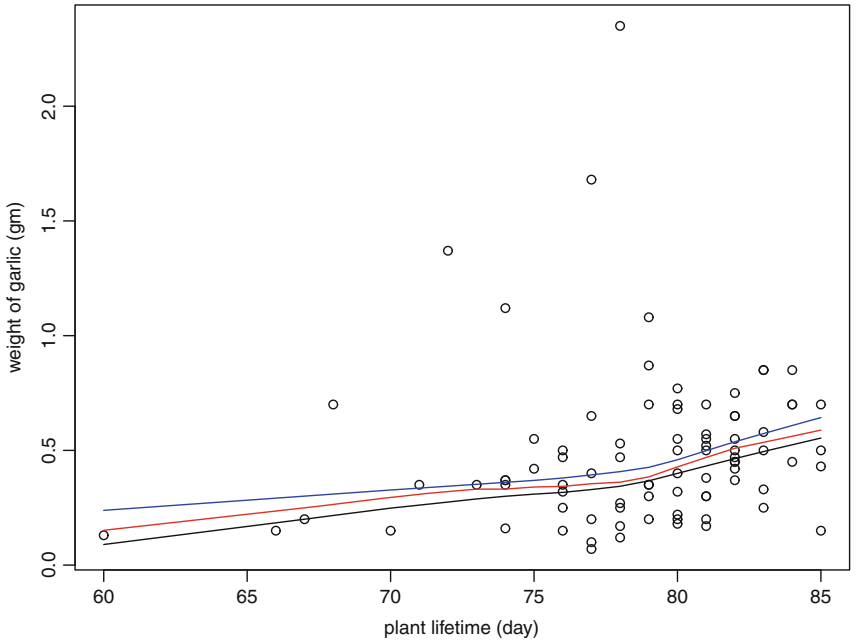


Fig. 5 Almost sure band for growth curve of a bulb crop (dehydrated). With 87 observations, the garlic growth curve is estimated by nonparametric lowess regression in Fig. 1. At each observed point on lifetime t of garlic plant, we compute the perturbation $\frac{1}{4n^{\alpha}} \sum_{i,j=1}^{n} |X_i - X_j|, \alpha \in (2, 2.5)$ as proposed in Gilat and Hill (1992). We take two immediate points above and below t and consider the corresponding garlic yield $X = X_i$ to compute the perturbation part with $n = 3, \alpha = 2.25$. Exception is made for the lowest time point, where we consider another additional repeat point to draw the band from start. The points so obtained by addition/subtraction from lowess curve fall above/below the lowess growth curve, and the points falling in a particular side of the curve are again lowess smoothed to obtain approximate a.s. upper/lower confidence bands

3 Modeling Crop Yield With Dominant Environment Affect

Extreme weather in general adversely affects crop production, e.g., see Gourdji et al. (2013). As seen from the experiment described above, healthy garlic plants result in positive yield in the experiment, and on an average a single clove of garlic produced 1.7 times yield in summer of Jharkhand. In other words, average number of additional garlic clove from a single clove planted is 0.7, on harvest. This estimate of additional clove is indirect, as this is based on weight. Later we shall examine a direct estimate.

Consider a situation where condition of weather, a random variable x is regulated by the parameter θ in a scale of $\theta \in [0, 1]$, indicating weather conduciveness for good yield in a region having generally harsh environment for agriculture. For θ near 0, i.e., in extreme weather there may not be any significant additional gain in yield.

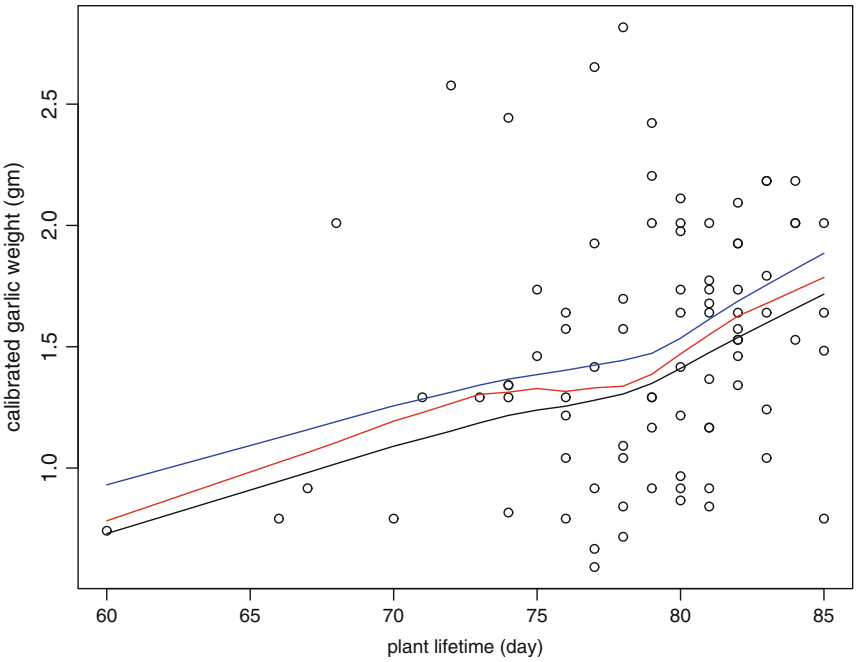


Fig. 6 Almost sure band for growth curve on calibrated yield. Almost sure confidence bands are drawn for growth curve based on calibrated weight (shown in Fig. 3), following a similar technique that is adopted for Fig. 5

Let the modeled *proportionate gain* in yield be θ, i.e., gain is θ times the seed weight, where θ is the true value of parameter. Then on an average, the number of additional clove from a single garlic clove planted is θ. Modeled crop production is approximately the same as the amount of seed planted in that region, if the weather is extremely harsh ($\theta \approx 0$). A suitably scaled garlic production on an average may be twice the initial investment in ideal condition ($\theta \approx 1$). Let the observed weather index x be a point binomial variable with parameter θ. The maximum summer temperature can be as high as 47 °C in that place. Damage to the major grain crops begins when temperatures rise above 30 °C during flowering. In rice, wheat, and maize, grain yields are likely to decline by 10 % for every 1 °C increase over 30 °C. At about 40 °C, yields are drastically reduced to zero; see Halweil (2014).

We consider temperature $\geq$38 °C as extremely harsh for garlic. Proportion of days where maximum temperature is 38 °C and above is 0.08696, during the experiment. This proportion can be considered as average of x values, where $x_i = 1$, if the maximum temperature on ith day is $\geq$ 38 °C, and $x_i = 0$, otherwise. The proportion 0.08696 may be considered as a priori value of θ, in absence of any additional information on relevant variables, e.g., on yield.

Let the observed number of cloves be $(1 + y)$, grown out of a single garlic clove planted in the experiment, where $y(\geq 0)$ is a Poisson variable with parameter θ. We find maximum likelihood estimate (m.l.e.) of θ on the basis of independent samples on (x, y).

Likelihood based on plant characteristics and field environment $(x_i, y_i), i = 1, \cdots, n$ is

$$L(\theta) \propto e^{-n\theta}\theta^{(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i)}(1-\theta)^{(n-\sum_{i=1}^{n} x_i)},\ \theta \in [0,1] \tag{1}$$

From (1), log likelihood based on $(x_i, y_i), i = 1, \cdots, n$ is

$$\log L(\theta) \propto -n\theta + (\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i)\log\theta + (n - \sum_{i=1}^{n} x_i)\log(1-\theta), \theta \in (0,1) \tag{2}$$

One of the roots of resulting quadratic equation in θ in log likelihood equation $\frac{d}{d\theta}\log L(\theta) = 0$ from (2) lies outside the parametric space; the other root is

$$\hat{\theta} = 1 - [\{4(1-\bar{x}) + (\bar{y})^2\}^{1/2} - \bar{y}]/2 \tag{3}$$

This lies within the parametric space and is strongly consistent for θ. Second derivative of log likelihood is negative, $\frac{d^2}{d\theta^2}\log L(\theta) = -\frac{n(\bar{x}+\bar{y})}{\theta^2} - \frac{n(1-\bar{x})}{(1-\theta)^2}$. The m.l.e. depends on the coordinates of (x, y) values via their mean $\bar{x}$ and $\bar{y}$.

Note that $\hat{\theta} \to 1$ as $\bar{x} \to 1$ and/or $\bar{y} \to \infty$. Equation (3) represents a three-dimensional relation for $\hat{\theta}$ on $\bar{x} \in [0,1], \bar{y} \in [0, \infty)$.

Figure 7 shows the m.l.e. of θ as a function of $\bar{y}$, plotted for different values of $\bar{x}$. Initial values of m.l.e. are quite sensitive to $\bar{x}$.

For the present case, $\bar{x} = 0.08696$ and $\bar{y} = 0.7$. An estimate of the parameter θ from Eq. (3) provides $\hat{\theta} = 0.3324$.

Now consider the case $\bar{x} = 1$. The term involving $(1 - \theta)$ vanishes in the likelihood function given by (1), and the function L is increasing as $\theta \uparrow 1$, $\frac{d}{d\theta}\log L(\theta) = -n + \frac{n+n\bar{y}}{\theta} > 0$.

Thus, for $\bar{x} = 1$, m.l.e. of θ is 1, irrespective of the y values. The m.l.e. then relies on x values only.

If the weather condition is conducive over several seasons, then under the proposed model $\hat{\theta} = 1$, irrespective of the observed value of crop yield.

Farmers may expect twice the initial investment as yield, in such a situation.

Behaviour of the m.l.e. in this case has some similarity with Bayesian viewpoint in data analysis. If the joint distribution involving prior and data is such that a segment of information is convincing, then the remaining part becomes redundant. In the proposed environment-yield model, information on $\bar{y}$ is ignored for computation of m.l.e., in a part of sample space of (x, y); where $\bar{x} = 1$.

Instead of taking the ratio of total seed weight and final weight of garlic produced, one may consider the ratio of number of cloves that resulted in healthy plants with positive yield, and total number of cloves in harvested garlic, in order to calculate gain in crop yield. In the present case the ratio is $98/87 = 1.1267$, thus gain in yield

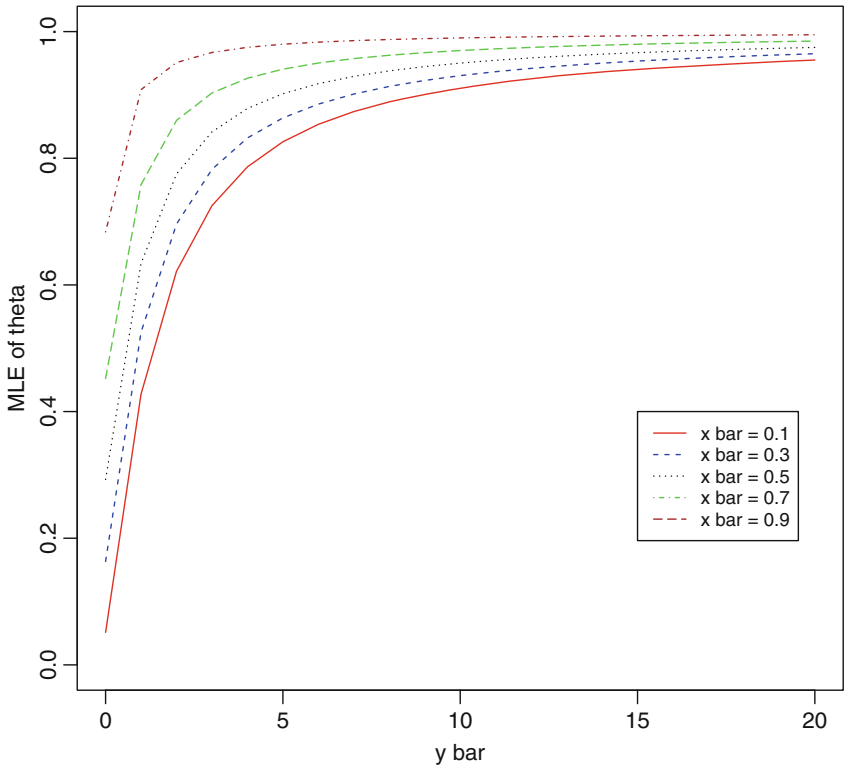


Fig. 7 MLE of theta. Maximum likelihood estimate of θ of environment-yield model is shown as a function of $\bar{y}$, the mean of additional yield; plotted for different values of $\bar{x}$, the mean of environment index. Initial values of the m.l.e are quite sensitive to $\bar{x}$

is 0.1267; this may provide a new estimate of θ from Eq. (3) with $\bar{x} = 0.08696$ and $\bar{y} = 0.1267$, leading to a modified estimate $\tilde{\theta} = 0.10572$.

Estimate of θ with a prior value 0.08696, and different yield input values 0.7 and 0.1267, as described above, turns out to be 0.3324 and 0.10572, respectively.

In view of low return of crop yield in harsh environment, garlic does not seem to be a worthwhile cultivable crop in Giridih, Jharkhand, unless adequate fertilizers e.g., DAP, organic manure etc. are administered and additional cares like regular irrigation, loosening the soil near plants are undertaken.

In a follow-up study to be reported later, the growth scenario is seen to improve by front shifting the time zone of garlic cultivation, with early winter plantation of seedlings. Simultaneously the other concerns of land fertility and plant care are also attended.

4 Consistency of Derivative Estimates and Estimation of Higher Derivatives

First we show the consistency of estimates.

4.1 Consistency of Derivative Estimates

Lowess, as least square estimates, inherits properties like almost sure convergence with appropriate rates, see, e.g., Lai et al. (1979) and Krätschmer (2006). We mimic the steps of Dasgupta (2013a) to show consistency of higher order derivative estimates obtained by lowess technique explained in Dasgupta (2013b).

Consider the model $y = g(x) + \epsilon$, where g has continuous second derivative and ϵ denotes error term, and let $(x_i, y_i), i = 1, \cdots, n$ be the growth observations y_i at time x_i.

Growth estimate from lowess curve is weighted least square (ls) regression of a low degree polynomial p usually of degree ≤ 2, to $(x_i, y_i), i = 1, \cdots, n$; fitted locally at the point $x = x_i$, i.e., $\hat{y}_i = \hat{y}(x_i) = p_{ls}^{(i)}(x)|_{x=x_i} = p^{(i)}(x_i)$. With smooth weight function the lowess estimates $p(x)$ are smooth and continuous in the sense that, $|p_{ls}^{(i)}(x) - p_{ls}^{(j)}(x)| \to 0$, as $|x_i - x_j| \to 0$, where $x_i, x_j \in [a, b]$, a finite interval; interpolation is linear in lowess regression for intermediate points $x \in (x_i, x_{i+1})$.

Let $\beta = \beta_n = \max_{1 \leq j < n} |x_j - x_{j+1}| \to 0$, as $n \to \infty$. Then there are sufficiently many observations in a small neighbourhood of x_i, where g' is continuous. Empirical slope estimates $(\hat{y}(x_i) - \hat{y}(x_j))/(x_i - x_j), j \neq i$; in that small neighbourhood are close to the derivative. For small grid spacing,

$$(\hat{y}(x_i) - \hat{y}(x_{i+1}))/(x_i - x_{i+1}) = \frac{d}{dx} p(x)|_{x=x_i} (1 + o(1))$$

as $|x_i - x_{i+1}| \to 0$.

Since the lowess estimate is a.s. consistent, i.e., $\hat{y}_i = g(x_i) + R$, where $R = R_n = o(1)$, a.s. as $n \to \infty$, we have for small $|x_i - x_j|$

$$\begin{aligned} m_i(j) = (\hat{y}_i - \hat{y}_j)/(x_i - x_j) &= g'(x_i) + o(|x_i - x_j|) + \frac{R_n(x_i) - R_n(x_j)}{x_i - x_j} \\ &= g'(x_i) + o(|x_i - x_j|) + o(1), \text{ a.s., } n \to \infty. \\ &\to g'(x_i), \text{ a.s., } n \to \infty \text{ and } |x_i - x_j| \to 0. \quad (4) \end{aligned}$$

The $o()$ terms in (4) are negligible for sufficiently large n and small grid spacing. Weighted average of empirical slopes $m_i(j)$ defined in (4) is then a consistent estimate for $g'(x_i)$. Weights $w(d_{ij}) = w(|x_i - x_j|)$ at x_i for x_j are standardised

with sum of the weights over j as 1, e.g., $w \to w/\bar{w}$. Distant x values from x_i are down-weighted to have negligible contribution to the sum $\tilde{m_i}$. Thus,

$$\tilde{m_i} = \sum_j w(d_{ij}) m_i(j) = g'(x_i) + o(1), \text{ a.s., } n \to \infty; \tag{5}$$

as the weights $w(d_{ij})$, $j \neq i$ are concentrated near x_i.

Usually in practice, exponentially decaying w are considered, this performs well in most of the cases. Estimate $\tilde{m_i}$ of (5) are linear combination of $\hat{y}$, the lowess values, these in turn are least square estimates and hence linear function of observations $y_i, i = 1, \cdots, n$. In most of the applications a robust estimate is taken as $\tilde{m_i}$, e.g., median or trimmed mean of the elements $\{w(d_{ij}) m_i(j), j = 1, \cdots, n; j \neq i\}$, these estimates are insensitive to outliers and choice of weight function.

Lowess smoothing of $(x_i, \tilde{m_i}), i = 1, \cdots, n$; with tricubic weight u, and variable bandwidth based on k-th nearest neighbour is taken as the estimate $\tilde{\tilde{m}}_i$ of $g'(x_i)$. The lowess estimate $\tilde{\tilde{m}}_i$, being a least square estimate, is consistent almost surely. Repeating the above steps with $(x_i, \tilde{\tilde{m}}_i), i = 1, \cdots, n$ lowess estimates of second derivative g'' of g is computed and strong consistency follows. Almost sure limits are considered on the intersection of different sets of probability 1, to obtain convergence results. From earlier proof sketched in Dasgupta (2013a), assumption that R is Lipschitz of order $\alpha > 0$ may be dropped; there is a typo regarding α therein. Since the function g is differentiable, and at each iteration lowess estimate $\hat{y}$ is approximated by polynomials in x, the difference $R = \hat{y} - g$ is then a well-behaved function.

Almost sure consistency of higher order derivatives by lowess follows similarly.

4.2 Growth Curve Derivatives from Calibrated Garlic Yield

We compute first two derivatives and proliferation rates of garlic growth, shown to be consistent as above. In Fig. 8 we plot velocity of calibrated growth that remains nearly constant for garlic weight till 72 days and then it drops down to reach a minimum at plant lifetime of 78 day. Afterwards the velocity rises again with lifetime.

Figure 9 shows the proliferation rate of calibrated garlic weight. This gradually decreases till plant lifetime of 78 days to reach a minimum, and then it rises again with increase of lifetime.

The second derivative curve is shown in Fig. 10. Second derivative reflects curvature of growth. In the beginning the curve of second derivative remains more or less stable, it drops to a minimum at 77 day of plant lifetime, the curve rises again and shows a downward trend near the end.

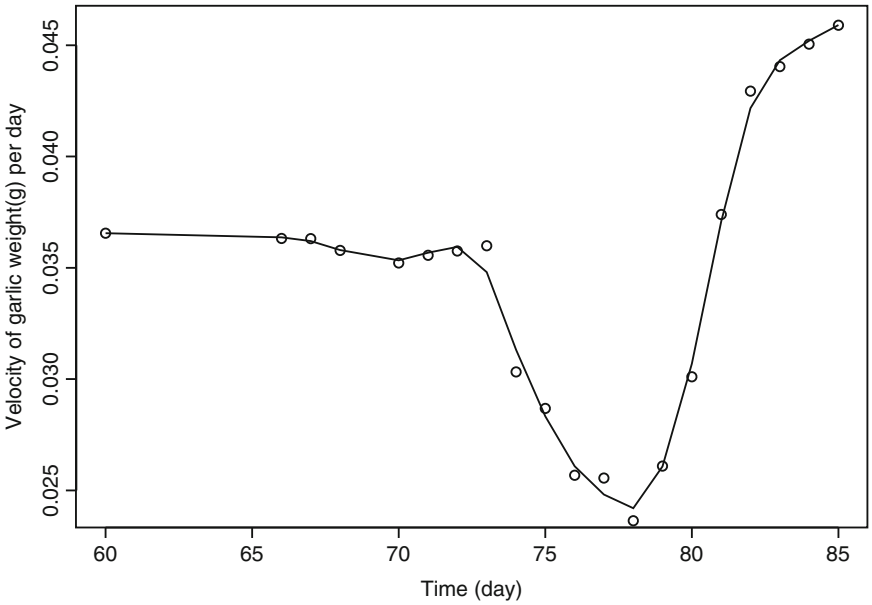


Fig. 8 Velocity of calibrated yield: trimmed mean, wt. exp(-.01 x); spline. Velocity remains nearly constant for garlic weight till 72 days and then it drops down to reach a minimum at plant lifetime of 78 day. Afterwards the velocity rises again with lifetime

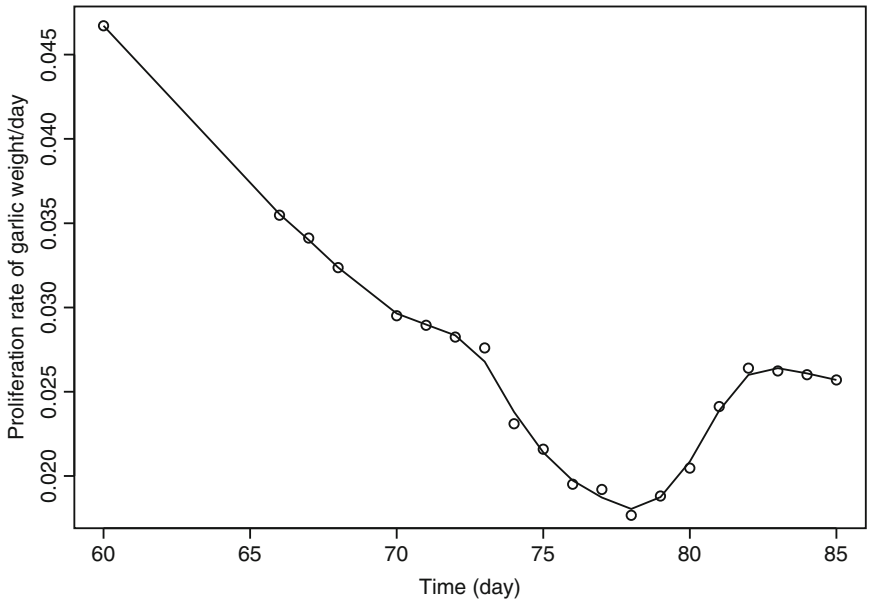


Fig. 9 Proliferation rate of calibrated yield: trimmed mean, wt. exp(-.01 x); spline. Proliferation rate gradually decreases till plant lifetime of 78 days to reach a minimum, and then it rises again with increase of lifetime

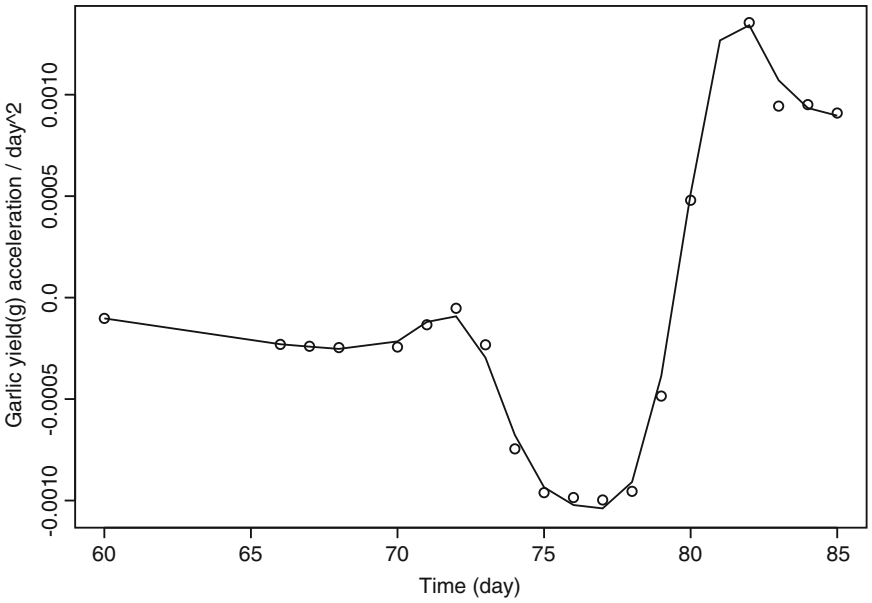


Fig. 10 Calibrated garlic wt second derivative: trimmed mean, wt. exp(-.01 x);spline

5 Estimation of Composite Function and Associated Error

Growth curve may sometimes require non linear calibration for modeling of characteristic growth in ideal environment as explained above. The calibration function may have to be estimated independently. Proliferation rate of composite function and associated errors are of interest, see, e.g., Dasgupta and Pan (2015).

Consider a real valued composite function $fog(x) = f(g(x))$. Let $\hat{f}$ and $\hat{g}$ be estimates of real valued functions f and g, respectively, from available data. Then $\hat{fog}(x) = \hat{f}(\hat{g}(x))$ is a natural estimate of $f(g(x))$. Usually $\hat{f}, \hat{g}$ are based on weighted average type estimates having negligible errors. We would like to study the effect of error in estimation of individual function while estimating composite function.

Write $\hat{fog}(x) = \hat{f}(\hat{g}(x)) = \hat{f}[g(x) + \epsilon]$ and $\hat{f} = f + \delta$, where the errors ϵ and δ in estimating f and g, respectively, are small in large sample. Next, suppose f is thrice differentiable in a neighbourhood of $g(x)$. Then,

$$\begin{aligned}
\hat{fog}(x) &= f([g(x) + \epsilon] + \delta) \\
&= f([g(x) + \epsilon]) + \delta f'([g(x) + \epsilon])(1 + o(1)) \\
&= f(g(x)) + \epsilon f'(g(x))(1 + o(1)) + \delta[f'(g(x)) + \epsilon f''(g(x))](1 + o(1)) \\
&= f(g(x)) + (\epsilon + \delta) f'(g(x))(1 + o(1)) + \epsilon\delta f''(g(x))(1 + o(1)) \qquad (6)
\end{aligned}$$

where $o(1)$ terms are negligible compared to the main terms.

To a first degree of approximation, error in estimation $(\hat{fog}(x) - f(g(x)))$ is symmetric in ϵ and δ, but not in f and g. The magnitude of error depends on first two derivatives of f. The product term $\epsilon\delta f''(g(x))$ is of lower order than the sum $(\epsilon + \delta) f'(g(x))$.

The technique of estimating derivative proposed in Dasgupta (2013a) is strongly consistent, $\hat{\psi}'(x) \to \psi'(x)$, a.s. Results of Dasgupta (2013b), pp. 82–84 indicate the gap is small.

One can heuristically assess error of estimation for derivative of composite function as follows. Considering only those terms of (6) explicitly written in x for differentiation, and pretending others terms to be free from x, write

$$\begin{aligned}\hat{\psi}'(x) &\approx \psi'(x) + [(\epsilon + \delta) f''(g(x)) + \epsilon\delta f'''(g(x))]g'(x)(1 + o(1)) \\ &\approx \psi'(x)\{1 + ((\epsilon + \delta) f''(g(x)) + \epsilon\delta f'''(g(x)))g'(x)/\psi'(x)(1 + o(1))\}\end{aligned}$$

Hence,

$$\frac{\hat{\psi}'(x)}{\hat{\psi}(x)} \approx \frac{\psi'(x)[1+\{(\epsilon+\delta) f''(g(x))+\epsilon\delta f'''(g(x))\}g'(x)(1+o(1))/\psi'(x)]}{\psi(x)[1+\{(\epsilon+\delta) f'(g(x))+\epsilon\delta f''(g(x))\}(1+o(1))/\psi(x)]}$$

$$\approx \frac{\psi'(x)}{\psi(x)}\{1+(\epsilon+\delta)(\frac{f''(g(x))g'(x)}{\psi'(x)}-\frac{f'(g(x))}{\psi(x)})+\epsilon\delta(\frac{f'''(g(x))g'(x)}{\psi'(x)}-\frac{f''(g(x))}{\psi(x)})\}(1+o(1))$$

from which convergence $\hat{\psi}'(x) \to \psi'(x)$ and $\frac{\hat{\psi}'(x)}{\hat{\psi}(x)} \to \frac{\psi'(x)}{\psi(x)}$ are apparent. As already mentioned, proliferation rate $\frac{\psi'(x)}{\psi(x)}$ of $\psi = fog$ are computed from data points by a technique proposed in Dasgupta (2013a) with exponentially decaying normalised weight assigned to individual slope estimates computed at a point, and then considering a robust estimate like trimmed mean or median of these weighted slopes for a particular x. The trimmed mean/median estimates are then smoothed by smooth.spline in SPlus software over the range of x. As a result random jig-jag behaviour in point estimates of proliferation/derivative arising from error components are smoothed out by cubic splines, only the main parts remain. Nonparametric lowess regression may also be alternatively used for smoothing the trimmed mean/median estimates. In large samples, variation of the quantities ϵ, δ and the $o(1)$ terms are then small compared to the main terms written explicitly involving x, and this heuristic works.

References

Cleveland WS (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54

Dasgupta R (2013a) Optimal-time harvest of elephant foot yam and related theoretical issues, Chap 6. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, Berlin, pp 101–130

Dasgupta R (2013b) Non uniform rates of convergence to normality for two sample U-statistics in non IID case with applications, Chap 4. In: Advances in growth curve models: topics from the indian statistical institute. Springer proceedings in mathematics & statistics, vol 46. Springer, Berlin, pp 60–88

Dasgupta R (2015) Growth curve of elephant foot yam, one sided estimation and confidence band, Chap 5. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Dasgupta R, Pan A (2015) Growth curve of phase change in presence of polycystic ovary syndrome, Chap 8. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Diriba-Shiferaw G, Nigussie-Dechassa R, Woldetsadik K, Tabor G, Sharma, JJ (2013) Growth and nutrients content and uptake of garlic (Allium sativum L.) as Influenced by different types of fertilizers and soils. Sci Technol Arts Res J 2(3):35–50

Gilat D, Hill TP (1992) One-sided refinements of the strong law of large numbers and the glivenko-cantelli theorem. Ann Probab 20:1109–1602

Gourdji SM, Mathews Ky L, Reynolds M, Crossa J, Lobell DB (2013) An assessment of wheat yield sensitivity and breeding gains in hot environments. Proc R Soc B 280:20122190

Halweil B (2014) The irony of climate. World Watch Magazine, March/April 2005, vol 18, no 2. http://www.worldwatch.org/node/572

Jelodar GA, Maleki M, Motadayen MH, Sirus S (2005) Effect of fenugreek, onion and garlic on blood glucose and histopathology of pancreas of alloxan-induced diabetic rats. Ind J Med Sci 59(2):64–9

Krätschmer V (2006) Strong consistency of least-squares estimation in linear regression models with vague concepts. J Multivar Anal 97:633–654

Lai TL, Robbins H, Wei CZ (1979) Strong consistency of least squares estimates in multiple regression II. J Multivar Anal 9:343–361

Matthew EE, Bassey EA, Clement A, Giddings EA, Edet EA, Kingsley HE (2007) A comparative assessment of the antimicrobial effects of garlic (Allium sativum) and antibiotics on diarrheagenic organisms. Southeast Asian J Trop Med Public Health 38(2):343–348

Scotton DC, Benedito VA, Molfetta JB, Rodrigues BIFP, Tulmann-Neto A, Figueira A (2013) Response of root explants to in vitro cultivation of marketable garlic cultivars. Hortic Bras 31(1):80–85

Growth Curve of Phase Change in Presence of Polycystic Ovary Syndrome

Ratan Dasgupta and Anwesha Pan

Abstract We study the time gap between two menstrual phases of semi-rural PCOS patients from West Bengal in relation to age at menarche, body mass index (BMI) and age of women. Mean response curve of phase gap based on spline regression starting from the age at menarche 9+ years reaches a maximum at the menarche age 11+ years. Growth curve of phase gap then gradually decreases till the menarche age of 16+. Phase gap of menstrual cycle is positively correlated with age of women. BMI also affects phase gap in an increasing manner. Effect of oral contraceptive pill medication on growth curve of phase gap and on proliferation rate is studied in terms of composite function. Discrete time proliferation rates computed from data, as approximation of proliferation rates are proposed and convergence rate of these estimates is studied in terms of the height of discrete steps in time measurements, and on errors in estimating the individual components functions.

Keywords Menarche • PCOS • Phase gap • BMI • Lowess regression • Spline regression • Proliferation rate

MS subject classification: Primary: 62P10, secondary: 62G08.

1 Introduction

Polycystic ovary syndrome (PCOS) refers to accumulation of multiple cysts in the ovaries, associated with high male hormone levels, chronic absence of ovulation and other metabolic disorders. Symptoms include excess facial and body hair, acne,

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

A. Pan
Ex. Project trainee, BAU, ISI; 50, Khaluibill Math, Burdwan 713101, West Bengal, India
e-mail: anwesha.pan1986@gmail.com

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_8

obesity, irregular menstrual cycles and infertility. Pre-pubertal obesity and early menarche are some of the possible factors for developing PCOS at later stage. Females who reach menarche at an early age expose their reproductive organs to the female hormone estrogens at an earlier age. This combined with a late marriage, results in a long gap between menarche and pregnancy, leading to other health problems. A child, who has irregular menstruation that continues into adulthood, develops imbalance in hormones, which often leads to problems conceiving a child. Exposure to inappropriate content in the media leads to early mental and physical maturity in girls. This also has an effect on a girl's hormones that have a strong feedback mechanism from the brain.

A long time gap between two menstrual phases instead of normal cycle creates confusion and tension. Phase gap is a measure of severity of the problem. Agrawal et al. (2004) studied prevalence of PCOS based on lifestyle. Association between PCOS and metabolic complications is discussed in Kelley et al. (2014). Correlation between biochemical and clinical features of PCOS is studied in Yousouf et al. (2012). Anti-mullerian hormone as a marker of PCOS is studied in Wiweko et al. (2014).

In the present study we investigate the association of phase gap with menarche age, body mass index (BMI) and age of patients having PCOS. Seventy-three females of semirural background were interviewed in a well-known hospital of North Kolkata on these variables in 2014. Some information was not available for recording. We observe that menarche age has an initial increasing and then decreasing effect on time gap between menstrual cycles; the peak on phase gap is seen for women with 11+ years of menarche age. The two variables BMI and age of patients are seen to have an overall increasing effect on phase gap for a long range. We study the multiple regression of phase gap on BMI, menarche age and age of patient. Regression on logarithmic scale provides improved value of R^2 than usual scale, thus the relationship seems to be nonlinear.

A model for assessing efficacy of oral contraceptive pill treatment on PCOS patients to regularise menstrual cycle is studied.

In Sect. 2 we explain materials and the methodology adopted. Results obtained from analysed data are discussed and interpreted. Growth of phase gap over age in PCOS patients without other complications is seen to slow down under oral contraceptive pill treatment, indicating improvement in patient status. Section 3 explores the relationship between continuous and discrete time proliferation rate for growth in composite function. Convergence rate of discrete time proliferation estimate is analytically seen to depend on the height of discrete steps in time measurements, and on errors in estimating the components functions of the composite function from observed data.

2 Materials, Methodology and the Results

Seventy-three patients with PCOS were interviewed during once in a week visit by interviewer, when the patients came for treatment in a government hospital at North Kolkata in the time period March–October 2014. Data recorded on the variables include age of patient, BMI, age at menarche and "phase gap", collected from willing patients. Nonparametric lowess and spline regression are used to obtain the response curves. Proliferation rate $d \log y(t)/dt$ of growth $y(t)$ is a measure of rate of change in $y(t)$ independent of unit of measurement for y. One may estimate this from observed data by a technique proposed in Dasgupta (2013).

In Fig. 1 we observe that the distribution of menarche age, starting from an early age of 9+, is positively skew in the interviewed patients with PCOS. Mode of the distribution is at 12+ years. The distribution of phase gap has a sharp fall towards right, as seen in Fig. 2. Largest value of the phase gap observed in the sample is 1 year, in contrast to normal gap of 28 days. The response curves are obtained by nonparametric lowess (with $f = 0.58$) and spline regression (with shape parameter $1/3$) in Fig. 3. *Menarche age has an initially increasing and then a decreasing effect on time gap between menstrual cycles; the peak of phase gap is seen for women with*

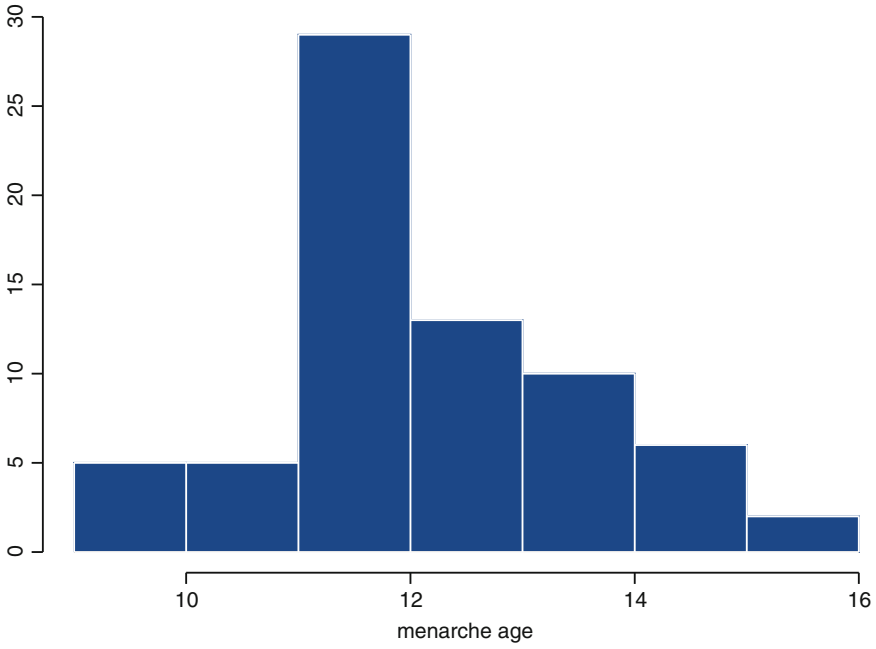


Fig. 1 Histogram of menarche age in PCOS patients

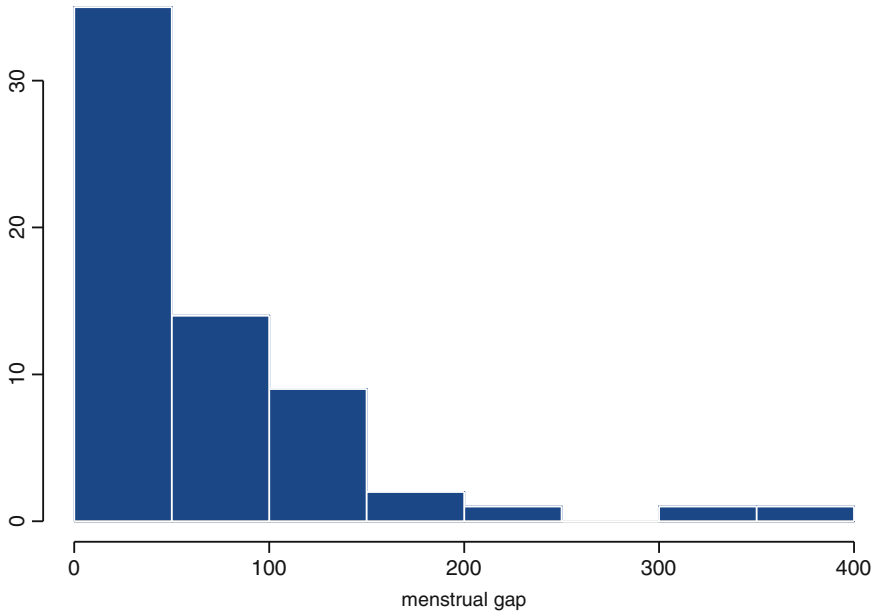


Fig. 2 Histogram of menstrual gap in PCOS patients

11+ years of menarche age in spline regression. The same feature is seen in lowess regression, with peak at age 13+. However, with $f = 0.5$ in lowess regression (not shown in figure) the peak is once again seen at the age 11+, as in spline regression. *There is some evidence that the phase gap is highest for women with menarche age at 11+.* The least square fit for linear regression is shown as a straight line in Fig. 3. The line has a decreasing trend with intercept 135.001 and slope -4.926.

Histogram of BMI in patients with PCOS is shown in Fig. 4, the distribution seems to be negatively skew.

A large percentage (>37 %) *of patients with PCOS is obese with BMI* > 25.

In Fig. 5 we plot lowess regression on data points and group means, *BMI is seen to have an overall increasing effect on phase gap.* Except near end of the graph, for two points, one with BMI 32.23 (for a woman aged 36 years, having androgen excess, performing regular physical exercise and *yoga* exercise to remain fit), and the other point with BMI 32.41 (for a young woman aged 22, under treatment for primary infertility for four years), increasing trend in the curves of Fig. 5 is apparent. These two extreme points have a pull down effect on the curve towards end. Lowess regression on data (with $f = 1/3$) and on group mean (with $f = .38$) are of similar increasing pattern. The least square linear fit has an increasing trend with intercept 71.1503 and slope 0.1736.

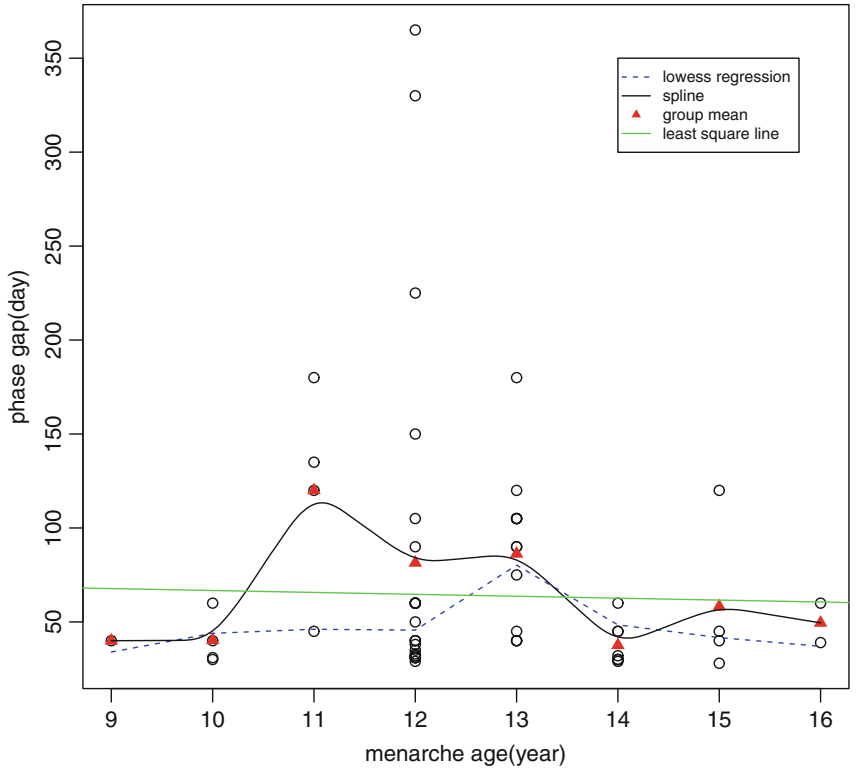


Fig. 3 Growth curve of phase gap on menarche age

Figure 6 indicates a positively skew age distribution of PCOS patients with long tail in the right, PCOS may affect women's health in the long range of reproductive age. The mode of the age distribution when the problem is reported is 21+. Difficulty in conceiving child is one of the main reasons for patients' problem reporting from semi-rural area.

Growth curves of phase gap based on lowess regression (with $f = 0.61$) and smooth.spline (with spar= .01) in Splus software are shown in Fig. 7 and Fig. 8, respectively. Lowess oversmooths the data points in this case. *Figure 8 indicates a general upward trend of phase gap with progress of age, eventually towards menopause, although a little bit of downward tendency of the curve within the age range* $(24, 29)$ *years is observed, indicating temporary relief. This growth pattern matches with the general experience of expert doctors where these patients were treated.*

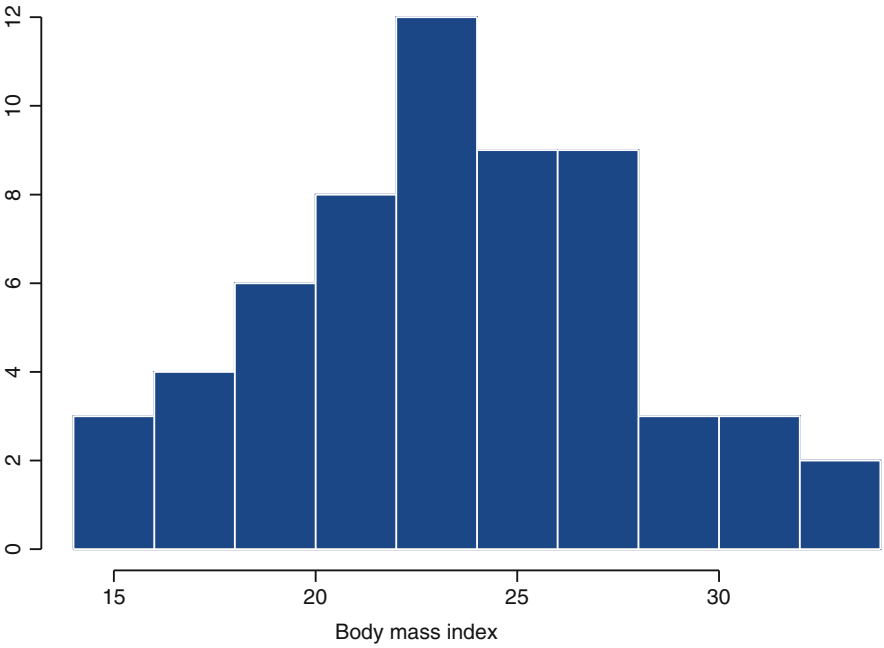


Fig. 4 Histogram of BMI in PCOS patients

Next we consider multiple regression of phase gap (y) in days, on age of the patient (x_1) in years, BMI (x_2) and menarche age (x_3) in years, by the method of least square. The estimated regression line is

$$y = 130.4300 + 1.3823x_1 - 0.5751x_2 - 5.8471x_3$$

The residual standard error is 72.97 on 47 degrees of freedom. Multiple correlation squared is $R^2 = 0.01866$, a low value; and the intercept term and the coefficients of $x_i, i = 1, 2, 3$ are all insignificant. Absolute residual plot of phase gap versus phase gap is shown in Fig. 9a. The residuals show an increasing trend towards upper right corner of the plot and five data points seem to be outliers. Deleting these five outliers we have the multiple regression as

$$y = 56.061 - 2.614x_1 + 1.649x_2 + 1.582x_3$$

This improves the value of multiple correlation squared to $R^2 = 0.1373$, still a low value. Residual standard error is 30.07 on 42 degrees of freedom. Intercept term is insignificant, so are the coefficients of x_2, x_3. Coefficient of x_1 is barely significant with $p = 0.0218$. Absolute residual plot of phase gap versus phase gap after deleting the five outliers is shown in Fig. 9b.

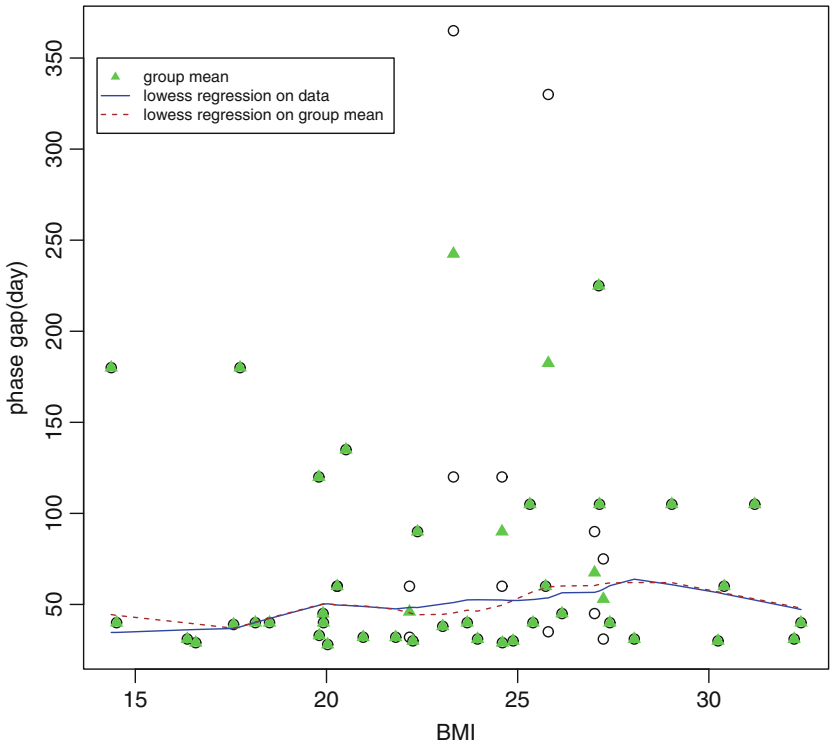


Fig. 5 Growth curve of phase gap on BMI

Next we check the performance of multiple regression in logarithmic scale. The accuracy of regression is slightly lower in logarithmic scale, $R^2 = 0.0178$.

$$\log y = 5.6992 - 0.4195 \log x_1 + 0.1480 \log x_2 - 0.3209 \log x_3$$

The residual s.e. is 0.6932 on 47 degrees of freedom. Absolute residual plot of phase gap versus phase gap in logarithmic scale is shown in Fig. 10a. The residuals show an increasing trend towards upper right corner of the plot as before, and the same five data points those were detected earlier seem to be outliers. *Deleting the outliers the multiple regression in logarithmic scale turns out to be*

$$y = 3.7344 - 1.1185x_1 + 0.7808x_2 + 0.4643x_3$$

This improves the value of multiple correlation squared as $R^2 = 0.194$. Intercept term is insignificant, so are the coefficients of x_2, x_3. Coefficient of x_1 is highly significant with $p = 0.00598$. The residual s.e. is 0.4561 on 42 degrees of freedom.

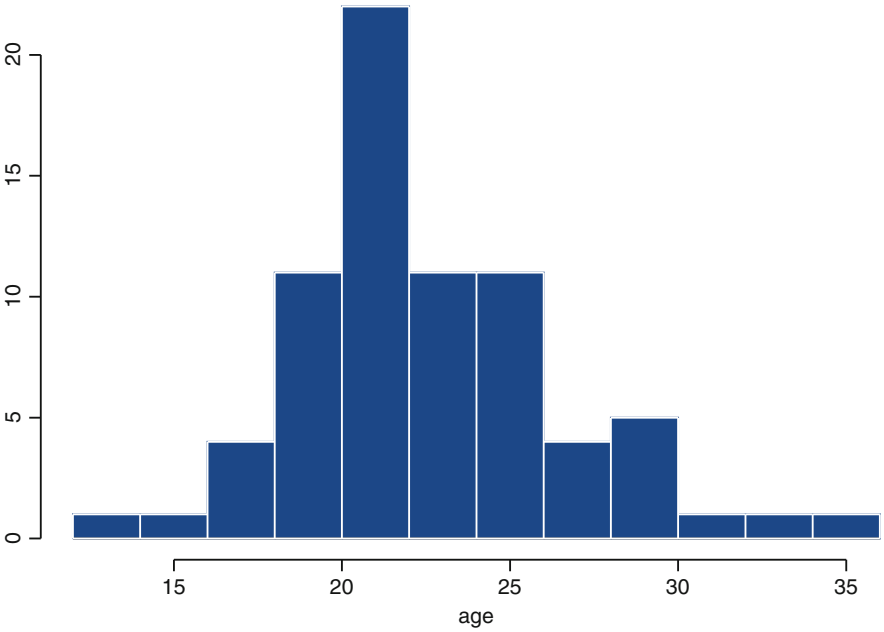


Fig. 6 Histogram of age in PCOS patients

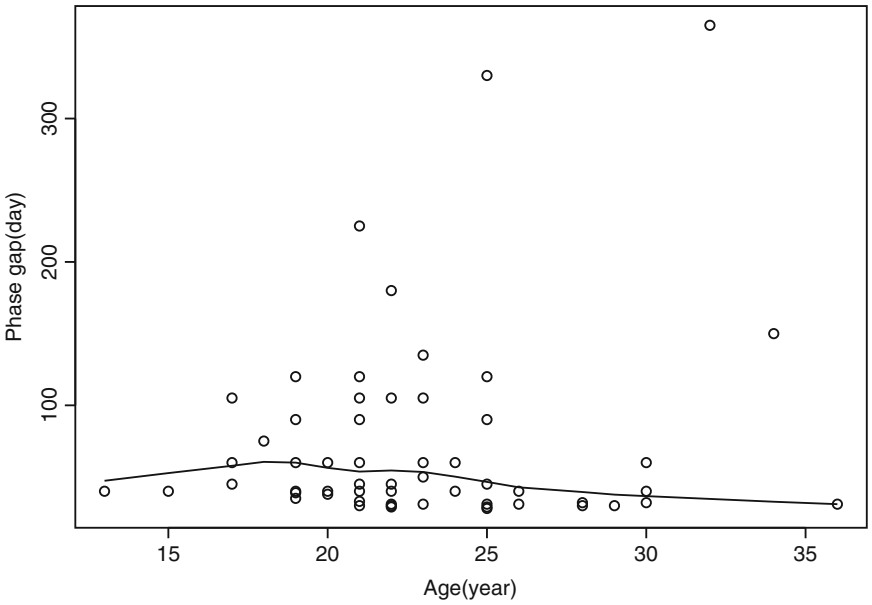


Fig. 7 Growth curve (lowess) of phase gap on age

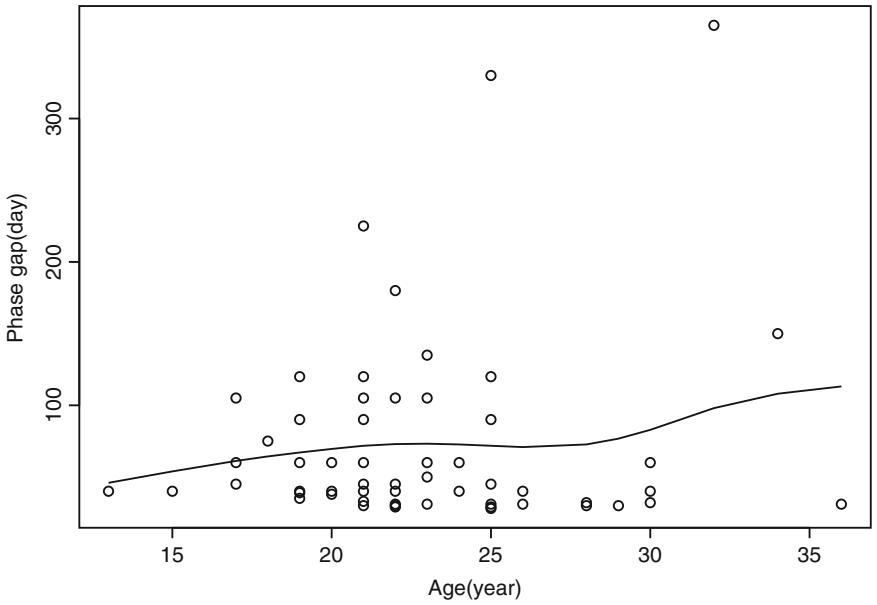


Fig. 8 Growth curve (spline) of phase gap on age

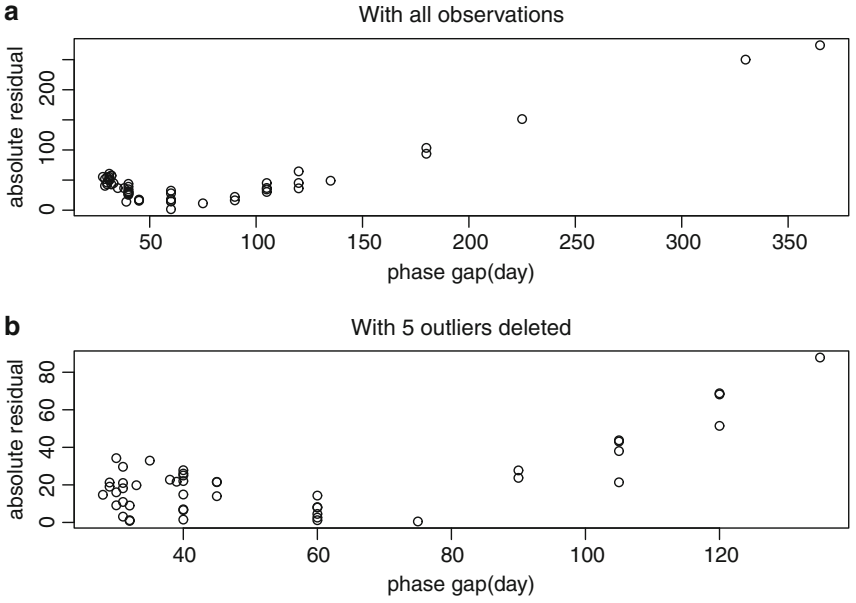


Fig. 9 Absolute residual plot of phase gap in multiple regression. (**a**) With all observations; (**b**) with five outliers deleted

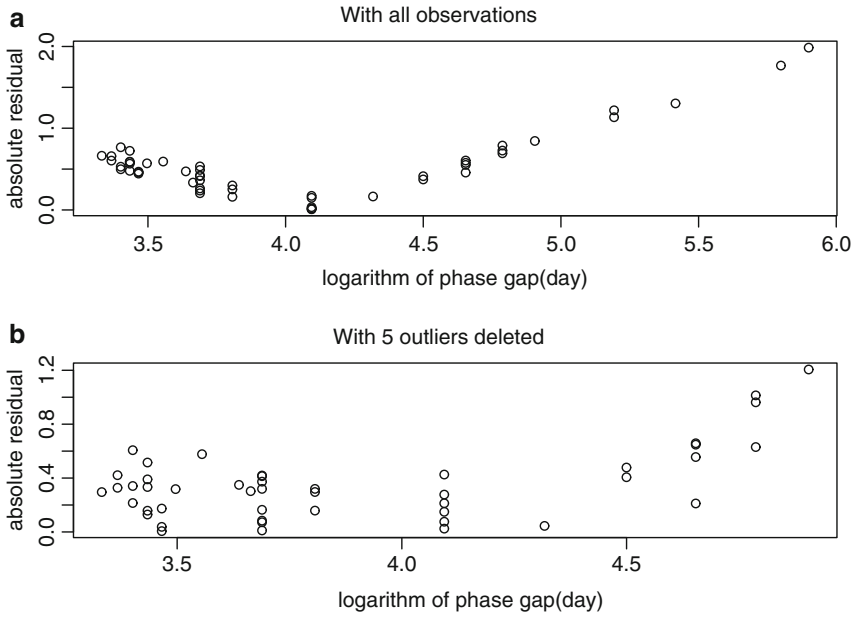


Fig. 10 Absolute residual plot of phase gap in multiple regression(log scale), (**a**) With all observations, (**b**) With five outliers deleted

Absolute residual plot of phase gap versus phase gap in logarithmic scale after deleting the five outliers is shown in Fig. 10b. *No specific pattern in residual scatterplot in Fig. 10b is observed.*

A low value of R^2 is not surprising, PCOS is heterogeneous disorders of uncertain causes. Not all the underlying processes surrounding it are properly understood at present.

Proliferation rate $d \log y(t)/dt$ is a measure of how rapidly phase gap changes over time, the measure is independent of y unit. To compute this, first y values are smoothed by smooth.spline in SPlus with spar$=$ 0.01. By a technique described in Dasgupta (2013), proliferation rate of phase gap is computed and shown in Fig. 11, taking spar$=$ 0.00001 in Splus. *The proliferation rate decreases initially, and then stabilises in the age interval* $[26, 28]$ *years, and then the curve rises again before falling beyond age 32.*

Standard treatment of oligomenorrhoea with oral contraceptive pill may reduce the phase gap in absence of further complicacy in patients. Proportion relief may be defined as proportionate lowering of the excess time gap over a normal period of 28 days when contraceptive pill medication is taken by a patient over a time span of 6/7 months. Proportion relief curve is computed in Fig. 12 by smooth.spline (spar$=$ .001) from data on two patients with age 21 and 26 years, with the assumption that relief is full at age 9 years, the smallest observed menarche age; and relief is almost nil towards the fag end of the growth curve for aged patients nearing menopause.

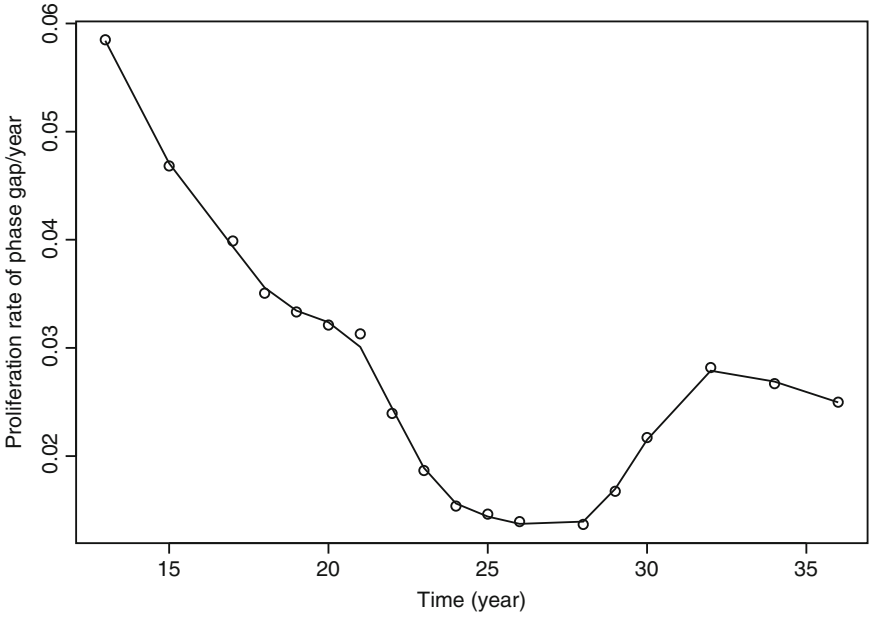


Fig. 11 Proliferation rate of phase gap with trimmed mean, wt. exp(−.01 x); spline

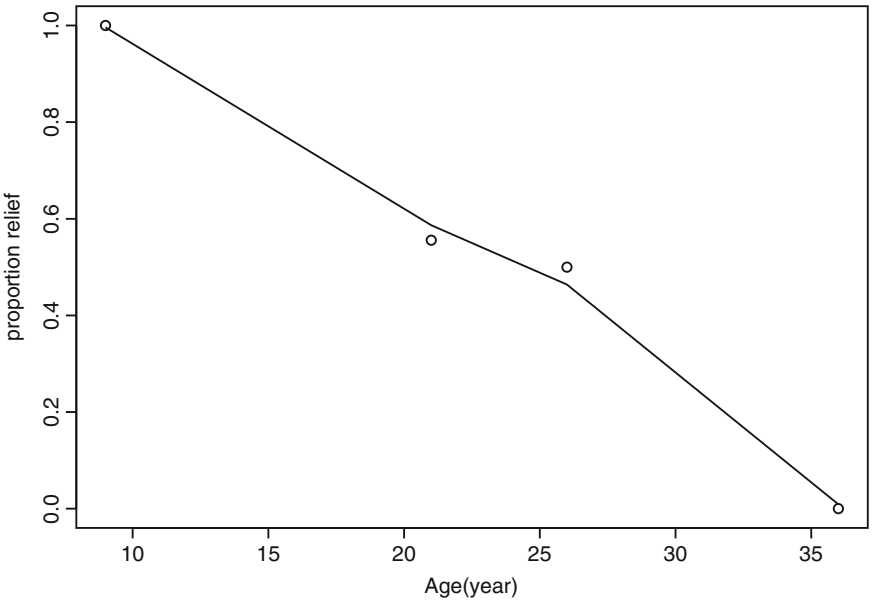


Fig. 12 Proportion relief growth curve of phase gap on age under treatment

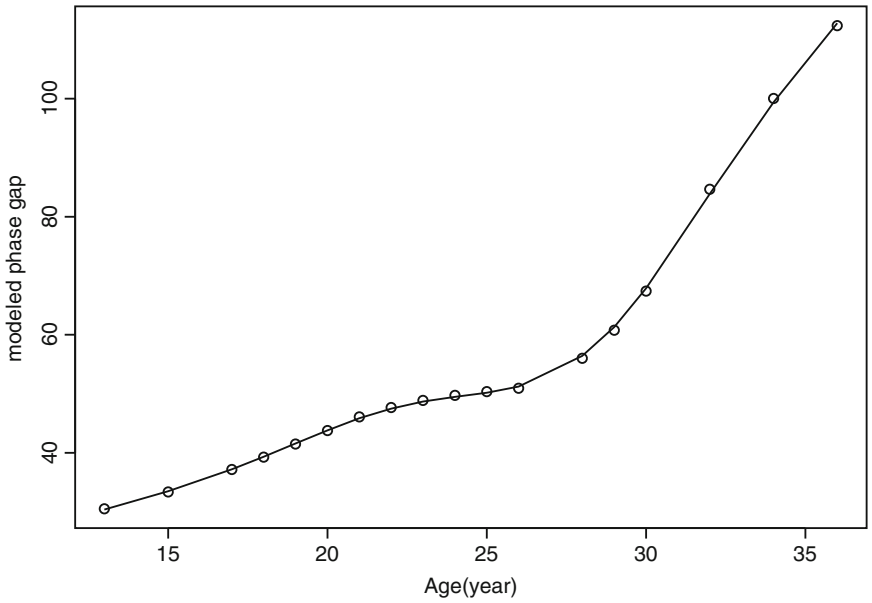


Fig. 13 Growth (spline) of modeled phase gap on age under treatment

In PCOS patients, benefitting from OCP treatment to regularise cycle, there are individuals with full relief having normal cycle after treatment. A common belief is sooner the medical treatment starts, the better the outcome. The mean benefit from OCP treatment for PCOS in the favorable outcome group may then be taken as full at 9+ years, the lowest age of menarche observed in the present data.

Menopause is a natural biological process in women approaching old age. So, benefit from OCP treatment to restore normal menstruation cycle for aged women may be considered almost nil.

Combining Fig. 8 (the function g, say) and Fig. 12 (the function f, say), we obtain slightly dampened modeled growth curve fog of phase gap with age for PCOS patients under treatment with oral contraceptive pill; the modeled curve, indicating efficacy of treatment, is shown in Fig. 13.

Proliferation rate of modeled growth curve is computed in two stages following similar procedures to obtain Fig. 11, modeled data points are smoothed initially by smooth.spline (spar= .00001), and then the rates are obtained by a technique described in Dasgupta (2013) with spar= 0.00001 in SPlus; the proliferation curve is shown. The pattern in Fig. 14 is almost similar to Fig. 11; except that for medicated patients the unique minimum of proliferation rate is attained at the age of 26 years. The age 26 years is slightly lower than the midpoint of the range $(24, 29)$ years, the time period of temporary relief for patients in between with slightly lower phase gap, as seen in the growth curve of Fig. 8.

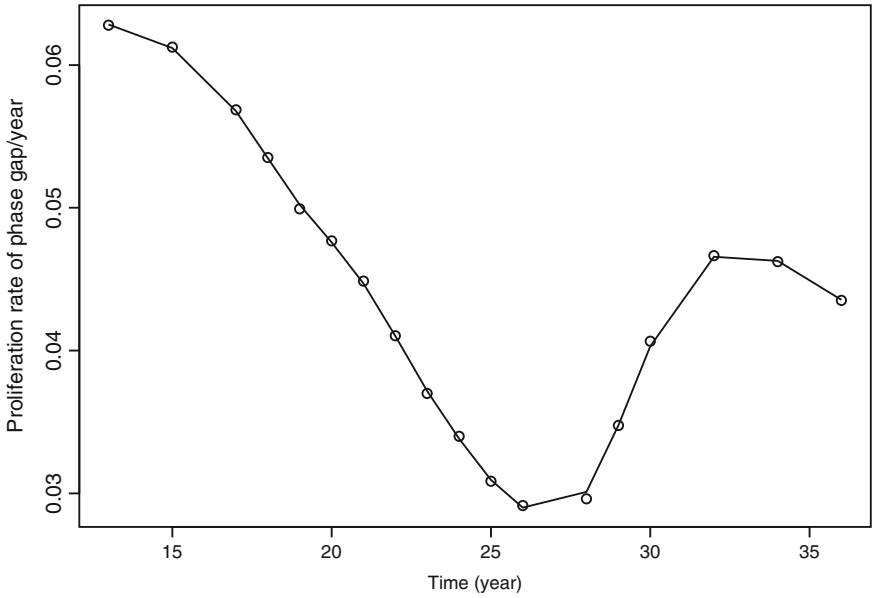


Fig. 14 Proliferation rate of treated phase gap: trimmed mean, wt. exp(−.01 x); spline

Proliferation rate based on composite function increases beyond the stage of 26 years and up to age 32, indicating aggravation of the problem in that time range. Errors associated with proposed estimation from composite function are derived in Dasgupta (2015).

The lowess regression of growth curve of phase gap on age of patients with PCOS (Fig. 7, oversmoothing the data points) leads to negative proliferation rate (not shown in figures) for age higher than 13, implying that on an average from age 13 onward women with PCOS would have a lowering trend in phase gap, which seems unrealistic. Thus an upward trend of phase gap, as shown in Fig. 8, seems plausible.

3 Discrete Time Proliferation Rate of Growth in Composite Function

In the above analysis we considered proliferation rate $d \log y/dt$ for a continuous function $y = y(t) = fog(t)$ of time t. The proliferation rate

$$\xi(t) = \frac{d}{dt} \log y = g'(t) \frac{f'(g(t))}{f(g(t))} \tag{1}$$

is then a calibration i.e., a scale multiplication of $g'(t)$ that may be nonlinear, with proliferation rate of f computed at $g(t)$. However, in many practical situations, data may only be available at discrete time points. Thus, it is of interest to study discrete version of proliferation rate and its relation to the continuous case (1).

Let the time spacing be h for observations taken at discrete time. A discrete version of proliferation rate $\xi(t)$ is then $\xi_h(t) = \frac{y(t+h)-y(t)}{hy(t)} = \frac{f(g(t+h))-f(g(t))}{hf(g(t))}$. In order to obtain a convergence rate of discrete proliferation, assume that for some $\alpha \in (1,2)$ the function $y = fog$ satisfies the following differentiability condition.

$$|y(t+h) - y(t) - hy'(t)| = o(|h|^{\alpha}) \tag{2}$$

Condition (2) is weaker than assuming second derivative of y. In the Young's form of Taylor's expansion this is equivalent in assuming that remainder term in first order expansion for y in a small neighbourhood of t with radius h is $o(|h|^{\alpha})$, $\alpha \in (1,2)$. That is,

$$\left|\frac{f(g(t+h)) - f(g(t))}{hf(g(t))} - g'(t)\frac{f'(g(t))}{f(g(t))}\right| = |\xi_h(t) - \xi(t))| = o(|h|^{\alpha-1}) \tag{3}$$

For $\alpha > 1$, difference between discrete proliferation rate and continuous proliferation rate in (3) decreases polynomially fast with small grid size.

Individual slope estimates $\frac{f(g(t+h))-f(g(t))}{hf(g(t))}$ from data points are smoothed by spline or lowess technique to obtain estimates of $\xi(t) = g'(t)\frac{f'(g(t))}{f(g(t))}$, as described in Dasgupta (2013).

The functions f and g are usually estimated from data. Effect of error in estimation of individual function, while estimating composite function, is studied in Dasgupta (2015). Effect of such errors in discrete time proliferation may be investigated. Write $\hat{fog}(t) = \hat{f}(\hat{g}(t)) = \hat{f}[g(t) + \epsilon]$ and $\hat{f} = f + \delta$, where the errors ϵ and δ in estimating f and g, respectively, are small in large sample. Next, suppose f is twice differentiable in a neighbourhood of $g(t)$. Then, following Dasgupta (2015)

$$\begin{aligned}\hat{fog}(t) &= f(g(t)) + (\epsilon+\delta)f'(g(t))(1+o(1)) + \epsilon\delta f''(g(t))(1+o(1)) \\ &= f(g(t)) + O(\epsilon+\delta)\end{aligned} \tag{4}$$

where $o(1)$ terms are negligible compared to the main terms.

To a first approximation, error in estimation $(\hat{fog}(t) - f(g(t)))$ is symmetric in ϵ and δ, but not in f and g. The magnitude of error depends on first two derivatives of f. The product term $\epsilon\delta f''(g(t))$ is of lower order than the sum $(\epsilon+\delta)f'(g(t))$.

Thus from (3) and (4)

$$|\frac{\hat{f}(\hat{g}(t+h))-\hat{f}(\hat{g}(t))}{h\hat{f}(\hat{g}(t))}-g'(t)\frac{f'(g(t))}{f(g(t))}| = |\hat{\xi}_h(t)-\xi(t))| = o(|h|^{\alpha-1})+O(\epsilon+\delta) \tag{5}$$

Equation (5) provides an order of overall error approximation in estimating the proliferation rate.

References

Agrawal R, Sharma S, Bekir J, Conway G, Bailey J, Balen AH, Prelevic G (2004) Prevalence of polycystic ovaries and polycystic ovary syndrome in lesbian women compared with heterosexual women. Fertil Steril 82:1352–1357

Dasgupta R (2013) Nonuniform rates of convergence to normality for two sample U-statistics in non iid case with applications, Chap 4. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, Berlin, pp 61–88

Dasgupta R (2015) Growth curve in damaged experiment via nonlinear calibration, Chap 7. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Kelley CE, Brown AJ, Diehl AM, Setji TL (2014) Review of nonalcoholic fatty liver disease in women with polycystic ovary syndrome. World J Gastroenterol 20:14172–14184

Wiweko B, Maidarti M, Priangga MD, Shafira N, Fernando D, Sumapraja K, Natadisastra M, Hestiantoro A (2014) Anti-mullerian hormone as a diagnostic and prognostic tool for PCOS patients. J Assist Reprod Genet 31:1311–1316

Yousouf R, Khan M, Kounsar Z, Ahangar S, Lone W (2012) Polycystic ovarian syndrome: clinical correlation with biochemical status. Surg Sci 3:245–248

Declining Patterns of Average Height of Adult Indians Between 20 and 49 Years: State Wise Trends and Influence of Socioeconomic Factors

Susmita Bharati, Manoranjan Pal, and Premananda Bharati

Abstract In the present study, changes in the average height over ages among women and men have been studied through third round National Family Health Survey data. It is also aimed to study the extent of influence of the different socioeconomic variables on such changes. The sample sizes for female and male are 94,417 and 52,460, respectively. For this study, only adult male and female data and the age ranges 20–49 years have been considered. During the 30 years span, the data set has been divided into three consecutive time periods with 10 years span for each period like (20–29), (30–39) and (40–49) years. Height has been considered as the dependent variable. The background explanatory variables are type of places, educational attainment, religion, ethnicity, occupational categories and wealth index of the families. The study shows that negative changes occur in the heights over the successive age-groups for men and women separately. The changes are found to be negative in all the zones and most of the states in India though it varies in its intensities. It is also an interesting feature to note that the maximum of absolute growth occurs among the men and women in urban areas, among the richest families, higher educated persons and professionals, while it is not so pronounced among the manual labourers, and scheduled tribes. Is it because of the changing lifestyles of most of the urban families and some of the rural families?

Keywords Height • Decadal changes • Socio-economic condition • Regression • India

S. Bharati
Sociological Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata, West Bengal 700108, India

M. Pal
Economic Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata, West Bengal 700108, India

P. Bharati (✉)
Biological Anthropology Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata, West Bengal 700108, India
e-mail: bharati@isical.ac.in; pbharati@gmail.com

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_9

1 Introduction

There is a considerable variation in trend in adult stature with changes in age and this trend is generally negative with the advancement of age. The negative effect in stature in human body becomes conspicuous in post adulthood phase, i.e., after when one attains around 40 years of age. For some adults it may be visible even in late thirties. There is considerable variation of it among different populations in the world (Harvey 1974; Roche et al. 1981; Malina et al. 1982) and in India (Sidhu et al. 1975; Singh 1978; Bagga 2010, 2013). In India, this type of studies has been carried out for adults who have already attained 60 years and the adults in their twenties (Sharma et al. 1975). It has been found that magnitude of differences between young adults (around 20 years) and late adults (around 70 years) is generally 7–10 cm, though it varies among different communities widely. It may seem to be quite a considerable difference. But Miall et al. (1967) found the decline in stature about 6–7.2 cm among the Welsh women. Among Indian population, very recently, Bagga (2013) studied on Maratha women of 30–70 years and found that it declines up to 3.6 cm.

The decline in the stature with the advancement of age is mainly associated with the changes in the vertebral column, i.e. mainly compression of inter-vertebral discs and hypnosis. This decrement is related with the advancement of age. So, to study the decadal changes, along with the total changes, may reveal features including the intensity of decline in the stature over ages. This change in the length of vertebrae is also associated with osteoporosis and vertebral diseases which cause degenerative changes in vertebral column. Besides this, a few studies in India also stated that socio-economic status has effect on the intensity of degeneration in the stature of human irrespective of gender.

Though there have been studies on the decline in the stature of adults in India, most of the studies, except Bagga's study, dealt with very old data. Even Bagga's study consists of very small sample size and no conclusion can be drawn from such a small sample. In this context, our study provides an opportunity to investigate among adult Indian population through national level data. The objective of the study is to find (1) the decadal changes of the height of women and men of 20–49 years of age and (2) the extent of influence of the different socio-economic variables on such changes.

2 Materials and Methods

For this study, we have used the National Family Health Survey (NFHS-III) data conducted by the International Institute for Population Sciences (IIPS), Mumbai, in 2005–2006 (IIPS 2007). IIPS collected unit level data on reproductive aged men of age (15–54) years and women of age (15–49) years from 29 states in India. However, to maintain parity we have taken age range of (15–49) years for both

males and females. The sample sizes thus consist of 94,417 women and 52,460 men in the age-group 20–49 years. It may be noted that each round of NFHS is a cross-section data. The background explanatory variables are (1) type of place—rural and urban areas, (2) educational attainment of women grouped into four categories—illiterate (those who can neither read nor write), primary (literate up to class IV standard), middle (Class V to Class X standard) and high school & above (Class XI and above), (3) religion which is classified into four categories, namely Hindu, Muslim, Christian and Others, (4) ethnicity having four categories such as Scheduled Castes (SC), Scheduled Tribes (ST), Other Backward categories (OBC) and Others, (5) Occupations of the women are clubbed into five major groups like not working; professionals, managers, technicians; engaged in service or sales; engaged in agriculture related works and skilled, unskilled or manual labourers, and (6) wealth index of the families. Wealth index represents the economic status of the households. It is an indicator of the level of the wealth, which is consistent with expenditure and income measure (Rutstein 1999). It is based on 33 household assets and housing characteristics like type of windows, sources of drinking water, types of toilet facility, flooring, roofing, ownership of a mattress, a pressure cooker, a chair, a cot/bed, a table, an electric fan, a radio/transistor, television, telephone, a computer, a car, etc. Each household was assigned a score for each asset and the scores were summed for each household and individuals were ranked according to the score of the household and the scores were divided into five quintile groups starting from lower strata to higher strata like poorest, poorer, medium, richer and richest.

Since the ages of males and females span 30 years, the data set has been divided into three consecutive time periods with 10 years span for each period taking age ranges (20–29), (30–39) and (40–49) years. Height has been considered as the dependent variable. To measure the decadal changes of mean height, the mean height of youngest group (20–29 years) has been subtracted from elder groups (30–39 and 40–49 years) and also the mean height of the middle group (30–39 years) has been subtracted from the mean height of the eldest group (40–49 years), so that the differences between the two consecutive age-groups as well as between the two extreme groups can be compared. Besides correlation between height with age, education and wealth index, we have carried out a regression analysis to see how the socio-economic variables influence the height or rather changes in the height. It was done for each decadal age-group separately for males and females. Thus six regression equations have been found. Here height is the dependent variable and place of residence, education, religion, ethnicity, occupation and wealth index have been considered as independent or explanatory variables. Symbolically we can write

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \varepsilon_i. \quad (1)$$

where y is the dependent variable, i.e., height and the independent variables are $x_1 =$ Place of residence, $x_2 =$ Education, $x_3 =$ Religion, $x_4 =$ Caste/tribe, $x_5 =$ Respondent's occupation, $x_6 =$ Wealth Index and $x_7 =$ Age in years. α is the intercept term and the regression coefficients are β_1, β_2, β_3, β_4, β_5, β_6 and β_7, corresponding to the variables x_1, x_2, x_3, x_4, x_5, x_6 and x_7.

We have taken binary data for all explanatory variables but age. The binary variables take only 0 for base and 1 for the other category. The base categories are 'Rural' for Place of Residence, 'Primary educated or less' for Education, 'Hindus or Muslims' for Religion, 'SC, ST or OBC' for Castes, 'Other than not working, professionals, technicians or managers' for Occupation and 'Poor or middle income persons' for Economic level. The details of variables, sample sizes, etc. used in the analysis are given in the Appendix. The statistical package for the social sciences (SPSS, version 16.0) has been used for all the analysis.

3 Results

Table 1 and Fig. 1a, b give a vivid picture of the mean heights of women of ages (20–29) years, (30–39) years and (40–49) years, along with decadal changes of the mean heights by zones and states in India. The adult decadal growth in height is found to be negative with a reduction of 0.12 cm from 20–29 years to 30–39 years aged women and 0.31 cm from 30–39 years to 40–49 years, the total reduction being 0.43 cm. Small positive changes have occurred only in eight states out of 29 states in India. As many as 21 states witnessed negative changes. The growth is negative in all the zones. The highest total change occurs in South zone (−1.32 cm) and the lowest total change occurs in North zone (–0.10 cm). Out of total eight states in India, where positive changes have been observed, four states, namely Haryana, New Delhi, Punjab and Rajasthan belong to the north zone of India. The other positive growths are seen in Nagaland, Orissa, Uttar Pradesh and Goa. Almost same trend is seen for both the decadal changes. In India, it is also seen that magnitude of reduction in height due to decadal change from 30–39 to 40–49 years is more than 20–29 to 30–39 years. Since there are six zones in India and for each zone two changes are observed, we have altogether 12 changes for the zones. Out of these 12 changes, only 1 case shows positive growth from 20–29 years to 30–39 years in the central zone and the growth is only 0.11 cm.

Table 2 gives almost similar picture for men so far as positive and negative trends in the height, but here positive changes are found to be lesser in number. Also, the amounts of changes are seen to be more than those of women. The total difference is 1 cm, i.e., the change from 20–29 years to 40–49 years is less by 1 cm on the average for all men taken together. When seen zone-wise, the highest difference is −1.79 cm in west zone. The lowest difference is observed in North-east zone. Out of 29 states, the averages in the heights increased for 6 states, namely Arunachal Pradesh, Meghalaya, Jharkhand, Orissa, New Delhi and Punjab, and for the other 23 states the changes are either negative or remain more or less same. The magnitude of difference of this change for men is a bit more than that of women.

Table 3 describes the total difference and decadal changes in the mean height of women in respect of different socio-economic variables. It is seen that total difference is negative in older aged women than in the younger aged women and the magnitude of difference is more or less double in urban areas (−0.59 cm)

Table 1 Zone and state wise changes in the mean height of adult women between (20–29) years and (40–49) years in India

	Mean height (cm)						Total change (20–29) years and (40–49) years (cm)	Decadal changes	
	(20–29) years (x)		(30–39) years		(40–49) years (y)				
Zones/states	N	Mean	N	Mean	N	Mean	(y−x)	(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
North-east	*7,730*	*151.43*	*5,672*	*151.35*	*3,189*	*151.02*	*−0.41*	−0.33	−0.08
Arunachal Pradesh	582	150.89	423	151.15	190	150.69	−0.20	−0.46	0.26
Assam	1,356	150.61	1,057	150.79	563	150.34	−0.27	−0.45	0.18
Manipur	1,580	152.19	1,227	152.07	726	151.52	−0.67	−0.55	−0.12
Meghalaya	728	149.47	476	149.34	304	149.43	−0.04	0.09	−0.13
Mizoram	654	151.92	499	151.43	287	151.85	−0.07	0.42	−0.49
Nagaland	1,502	152.78	968	152.83	470	153.25	0.47	0.42	0.05
Sikkim	759	151.59	535	151.48	305	150.50	−1.09	−0.98	−0.11
Tripura	569	150.02	487	149.67	344	149.41	−0.61	−0.26	−0.35
East	*6,102*	*150.92*	*4,671*	*150.89*	*3,022*	*150.72*	*−0.20*	−0.17	−0.03
Bihar	1,286	150.77	884	150.68	593	150.48	−0.29	−0.20	−0.09
Jharkhand	1,014	150.19	736	150.25	396	149.78	−0.41	−0.47	0.06
Orissa	1,559	150.97	1,228	151.09	739	151.09	0.12	0.00	0.12
West Bengal	2,243	151.29	1,823	151.14	1,294	150.89	−0.4	−0.25	−0.15
Central	*7,196*	*151.76*	*5,618*	*151.87*	*3,652*	*151.58*	*−0.18*	−0.29	0.11
Madhya Pradesh	2,184	152.95	1,689	153.02	1,195	152.73	−0.22	−0.29	0.07
Chhattisgarh	1,257	151.77	1,013	151.55	667	150.72	−1.05	−0.83	−0.22
Uttar Pradesh	3,755	151.07	2,976	151.31	1,790	151.13	0.06	−0.18	0.24

(continued)

Table 1 (continued)

Zones/states	Mean height (cm)						Total change (20–29) years and (40–49) years (cm)	Decadal changes	
	(20–29) years (x)		(30–39) years		(40–49) years (y)				
	N	Mean	N	Mean	N	Mean	(y−x)	(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
West	*5,290*	*152.41*	*4,354*	*152.17*	*2,808*	*151.79*	*−0.62*	−0.38	−0.24
Goa	1,117	152.36	1,040	152.39	625	152.53	0.17	0.14	0.03
Gujarat	1,299	152.70	1,005	152.68	699	152.18	−0.52	−0.50	−0.02
Maharashtra	2,874	152.30	2,309	151.84	1,484	151.29	−1.01	−0.55	−0.46
North	*7,549*	*154.16*	*5,935*	*154.13*	*3,955*	*154.06*	*−0.1*	−0.07	−0.03
Haryana	931	154.81	757	155.02	483	155.12	0.31	0.10	0.21
Himachal Pradesh	1,059	153.95	924	153.80	569	153.41	−0.54	−0.39	−0.15
Jammu & Kashmir	1,072	159.59	776	154.44	520	153.85	−5.74	−0.59	−5.15
New Delhi	901	153.26	723	153.72	480	154.06	0.8	0.34	0.46
Punjab	1,291	154.75	960	154.47	707	154.78	0.03	0.31	−0.28
Rajasthan	1,311	154.54	1,001	154.52	737	154.56	0.02	0.04	−0.02
Uttaranchal	984	152.88	794	152.83	459	152.15	−0.73	−0.68	−0.05
South	*7,474*	*152.95*	*6,059*	*152.38*	*4,141*	*151.63*	*−1.32*	−0.75	−0.57
Andhra Pradesh	2,443	152.65	1,776	152.08	1,235	151.31	−1.34	−0.77	−0.57
Karnataka	1,962	152.69	1,497	152.73	939	152.13	−0.56	−0.60	0.04
Kerala	1,067	154.05	1,091	152.77	797	151.58	−2.47	−1.19	−1.28
Tamil Nadu	2,002	152.99	1,695	152.16	1,170	151.62	−1.37	−0.54	−0.83
India	*41,341*	*152.31*	*32,309*	*152.19*	*20,767*	*151.88*	*−0.43*	−0.31	−0.12

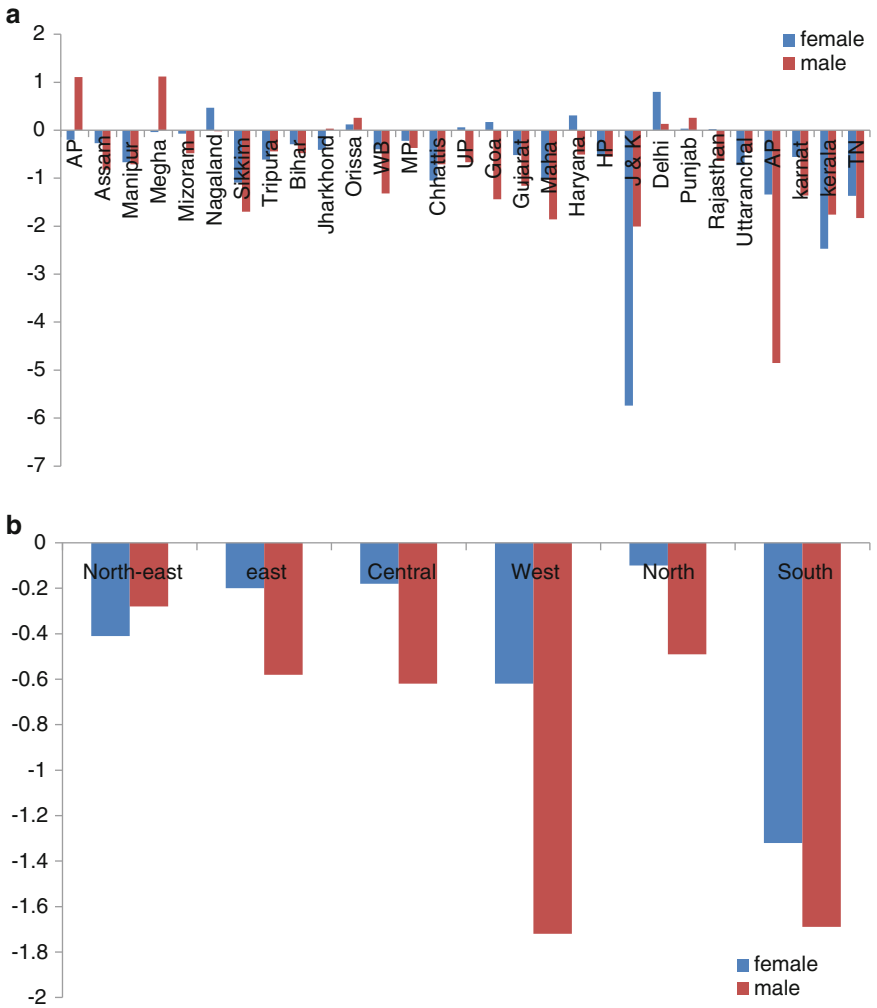


Fig. 1 (**a**) State wise changes in the mean height of adult female and male between the age-groups (20–29) years and (40–49) years in India. (**b**) Zone-wise changes in the mean height of adult female and male between the age-groups (20–29) years and (40–49) years in India

than in rural areas (–0.31 cm). In case of relationship with the status of education, illiteracy and primary educated women show positive increments at younger ages and negative growth in older age. The highest educated women have substantially higher heights than other women and do not show much negative trend. Among the Christian and other ethnic groups of women, positive changes are observed, while among Hindus and Muslims, always negative changes occur for all the age-groups. In case of ethnicity, highest negative changes have been found among other backward classes followed by scheduled castes and the lowest differenceis

Table 2 Zone and state wise changes in the mean height of adult men between (20–29) years and (40–49) years in India

Zones/states	Mean height (cm)						Total changes (20–29) years and (40–49) years (cm)	Decadal changes	
	(20–29) years		(30–39) years		(40–49) years			(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
	N	Mean	N	Mean	N	Mean			
North-east	*3,966*	*162.64*	*3,048*	*162.59*	*2,198*	*162.36*	*−0.28*	−0.23	−0.05
Arunachal Pradesh	228	161.07	163	161.96	125	162.18	1.11	0.22	0.89
Assam	376	163.42	329	163.46	249	162.60	−0.82	−0.86	0.04
Manipur	1,186	163.70	946	163.21	690	162.99	−0.71	−0.22	−0.49
Meghalaya	203	157.87	143	157.84	98	158.99	1.12	1.15	−0.03
Mizoram	226	162.67	159	161.71	120	162.19	−0.48	0.48	−0.96
Nagaland	1,247	163.13	959	163.21	645	163.11	−0.02	−0.10	0.08
Sikkim	291	160.46	181	160.46	122	158.76	−1.7	−1.70	0.00
Tripura	209	161.63	168	161.68	149	161.19	−0.44	−0.49	0.05
East	*1,830*	*163.54*	*1,593*	*163.52*	*1,259*	*162.96*	*−0.58*	−0.56	−0.02
Bihar	340	163.81	292	163.72	230	163.33	−0.48	−0.39	−0.09
Jharkhand	281	162.81	240	162.98	171	162.84	0.03	−0.14	0.17
Orissa	438	162.98	405	163.06	314	163.24	0.26	0.18	0.08
West Bengal	771	164.00	656	163.91	544	162.68	−1.32	−1.23	−0.09
Central	*4,553*	*165.01*	*3,594*	*164.61*	*2,660*	*164.39*	*−0.62*	−0.22	−0.40
Madhya Pradesh	861	165.81	674	165.54	536	165.44	−0.37	−0.10	−0.27
Chhattisgarh	385	163.72	386	163.84	266	163.01	−0.71	−0.83	0.12
Uttar Pradesh	3,307	164.96	2,534	164.48	1,858	164.29	−0.67	−0.19	−0.48
West	*3,317*	*166.00*	*2,710*	*165.36*	*1,961*	*164.28*	*−1.72*	−1.08	−0.64
Goa	312	165.19	318	164.92	219	163.75	−1.44	−1.17	−0.27

Gujarat	415	166.19	359	165.52	294	165.03	−1.16	−0.49	−0.67
Maharashtra	2, 590	166.07	2, 033	165.40	1, 448	164.21	−1.86	−1.19	−0.67
North	*2, 428*	*167.08*	*1, 838*	*166.71*	*1, 377*	*166.59*	*−0.49*	−0.12	−0.37
Haryana	333	168.38	252	168.40	194	167.87	−0.51	−0.53	0.02
Himachal Pradesh	271	166.09	264	164.96	192	165.54	−0.55	0.58	−1.13
Jammu & Kashmir	319	167.70	244	166.55	143	165.69	−2.01	−0.86	−1.15
New Delhi	371	165.64	189	165.17	155	165.77	0.13	0.60	−0.47
Punjab	421	168.37	288	168.97	248	168.63	0.26	−0.34	0.60
Rajasthan	457	167.29	346	167.38	273	166.65	−0.64	−0.73	0.09
Uttaranchal	256	165.27	255	164.67	172	164.82	−0.45	0.15	−0.60
South	*5, 747*	*165.92*	*4, 741*	*165.11*	*3, 640*	*164.23*	*−1.69*	−0.88	−0.81
Andhra Pradesh	2, 243	166.04	1, 664	165.33	1, 244	161.19	−4.85	−4.14	−0.71
Karnataka	1, 517	165.81	1, 320	165.19	958	164.45	−1.36	−0.74	−0.62
Kerala	281	167.55	290	166.15	243	165.79	−1.76	−0.36	−1.40
Tamil Nadu	1, 706	165.61	1, 467	164.58	1, 195	163.78	−1.83	−0.80	−1.03
India	*21, 841*	*165.08*	*17, 524*	*164.63*	*13, 095*	*164.08*	*−1.00*	−0.55	−0.45

Table 3 Decadal and total changes in the mean height of women of age (20–49) years in relation to different socio-economic variables

Independent variables	Time period						Total change (20–49) years (cm)	(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
	(20–29) years		(30–39) years		(40–49) years				
	N	Mean	N	Mean	N	Mean			
Place of residence									
Rural	22,698	152.01	17,655	151.90	11,195	151.70	−0.31	−0.20	−0.11
Urban	18,643	152.68	14,654	152.54	9,572	152.09	−0.59	−0.45	−0.14
Education									
Illiterate	11,275	151.11	13,173	151.48	9,746	151.29	0.18	−0.19	0.37
Primary	5,372	151.24	4,806	151.56	3,564	151.60	0.36	0.04	0.32
Secondary	18,622	152.61	11,082	152.64	5,868	152.49	−0.12	−0.15	0.03
Higher	6,068	154.59	3,246	154.49	1,586	153.85	−0.74	−0.64	−0.10
Religion									
Hindu	29,760	152.20	23,775	152.09	15,551	151.70	−0.50	−0.39	−0.11
Muslim	5,617	152.60	3,906	152.41	2,270	152.15	−0.45	−0.26	−0.19
Christian	3,783	152.11	2,850	152.19	1,737	152.18	0.07	−0.01	0.08
Others	2,181	153.23	1,778	152.99	1,209	153.21	−0.02	0.22	−0.24
Caste\-tribe									
SC	6,840	151.21	5,193	151.49	3,290	150.69	−0.52	−0.80	0.28
ST	5,847	151.63	4,172	151.59	2,419	151.47	−0.16	−0.12	−0.04
OBC	12,941	152.22	10,226	152.02	6,370	151.59	−0.63	−0.43	−0.20
Others	13,800	153.18	11,208	152.97	7,735	152.69	−0.49	−0.28	−0.21
DK	208	151.88	199	151.58	143	151.61	−0.27	0.03	−0.30
Occupation									
Not working	25,629	152.44	16,837	152.39	11,045	152.16	−0.28	−0.23	−0.05
Prof/tech/man	1,988	153.93	1,454	154.07	925	153.12	−0.81	−0.95	0.14

Clerk/sales/service	2,642	152.30	3,170	151.79	1,999	151.40	−0.90	−0.39	−0.51
Agriculture	7,204	151.71	7,399	151.79	4,887	151.47	−0.24	−0.32	0.08
Skilled/unskilled lab	3,849	151.75	3,418	151.59	1,888	151.41	−0.34	−0.18	−0.16
Wealth index									
Poorer	4,713	150.67	3,927	150.65	2,083	150.64	−0.03	−0.01	−0.02
Poorest	5,803	150.87	4,684	151.11	2,778	150.70	−0.17	−0.41	0.24
Middle	7,943	151.61	6,094	151.61	3,727	151.28	−0.33	−0.33	0.00
Richer	10,446	152.36	7,646	152.07	4,786	151.66	−0.70	−0.41	−0.29
Richest	12,436	154.02	9,958	153.74	7,393	153.12	−0.90	−0.62	−0.28

observed among scheduled tribes. Occupation-wise the highest total difference is found among service holders (–0.90 cm) and it is followed by professionals (–0.81 cm) while lower magnitudes of differences are observed among non-working women (–0.28 cm) and women engaged in agriculture (–0.24 cm) as well as for skilled/unskilled women labourers (–0.34 cm).The trend is same in both the decades but magnitude is higher in later period than younger ages. Regarding wealth Index, magnitude of negative changes is the highest among the richest women and the lowest among the poorest women.

Table 4 also describes the relationship between decadal changes in stature with different socio-economic variables among Indian men of aged (20–49 years).Secular total change is negative irrespective of all socio-economic variables. High magnitude of total negative changes has been observed in case of urban areas, Hindu religious group, not working and professional occupation holders and richest wealth index families. The same trend is more or less observed in case of decadal changes also.

Table 5 and Fig. 2 show the correlation between adult height with age, wealth index and educational level of men and women. The result shows that height is significantly positively correlated with these three socio-economic variables either negatively or positively. It is also seen that adult height is significantly negatively correlated with the age. Thus it proves that there is a negative trend in the heights with the advancement of adult age.

Table 6 contains the decadal age-group wise results of the linear regression of height with different socio-economic variables separately for female and male data in India. The six fitted regression equations are as follows:

Female height (40–49 years)

$$\underset{(0.000)}{\widehat{y} = 154.6} \underset{(0.000)}{- 0.668\, x_1} \underset{(0.000)}{+ 0.615\, x_2} \underset{(0.000)}{+ 0.758\, x_3} \underset{(0.000)}{+ 0.983\, x_4} \underset{(0.188)}{- 0.122\, x_5} \underset{(0.000)}{+ 1.309\, x_6} \underset{(0.000)}{- 0.088 x_7}. \quad (2)$$

Male height (40–49 years)

$$\underset{(0.000)}{\widehat{y} = 165.0} \underset{(0.000)}{- 0.518\, x_1} \underset{(0.000)}{+ 1.155\, x_2} \underset{(0.000)}{- 0.462\, x_3} \underset{(0.000)}{+ 1.451\, x_4} \underset{(0.072)}{+ 0.329\, x_5} \underset{(0.000)}{+ 1.784\, x_6} \underset{(0.001)}{- 0.063\, x_7}. \quad (3)$$

Female height (30–39 years)

$$\underset{(0.000)}{\widehat{y} = 151.6} \underset{(0.000)}{- 0.445\, x_1} \underset{(0.000)}{+ 0.719\, x_2} \underset{(0.065)}{+ 0.178\, x_3} \underset{(0.000)}{+ 0.763\, x_4} \underset{(0.440)}{- 0.055\, x_5} \underset{(0.000)}{+ 1.441\, x_6} \underset{(0.096)}{- 0.002 x_7}. \quad (4)$$

Table 4 Decadal and total changes in the mean height of men of age (20–49) years in relation to different socio-economic variables

Independent variables	Time period						Total change (20–49) years	Decadal changes	
	(20–29) years		(30–39) years		(40–49) years			(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
	N	Mean	N	Mean	N	Mean			
Place of residence									
Rural	10,444	164.38	8,899	164.08	6,722	163.60	−0.78	−0.48	−0.30
Urban	11,397	165.71	8,625	165.19	6,373	164.59	−1.12	−0.60	−0.52
Education									
Illiterate	2,378	162.86	3,223	162.92	2,822	162.67	−0.19	−0.25	0.06
Primary	2,940	163.11	2,900	163.42	2,643	163.02	−0.09	−0.40	0.31
Secondary	12,172	165.11	8,373	164.94	5,681	164.51	−0.6	−0.43	−0.17
Higher	3,023	166.76	3,023	166.76	1,944	166.32	−0.44	−0.56	0.00
Religion									
Hindu	15,978	165.19	13,137	164.72	9,866	164.05	−1.14	−0.67	−0.47
Muslim	2,872	165.53	2,026	164.93	1,416	164.70	−0.83	−0.23	−0.60
Christian	1,975	163.24	1,570	163.28	1,187	162.97	−0.27	−0.31	0.04
Others	1,016	165.62	791	165.08	626	165.43	−0.19	0.35	−0.54
Caste\-tribe									
SC	3,883	164.10	3,045	163.34	2,070	162.61	−1.49	−0.74	−0.76
ST	2,737	162.69	2,157	162.65	1,558	162.39	−0.3	−0.26	−0.04
OBC	7,892	165.20	6,346	164.77	4,883	164.03	−1.17	−0.74	−0.43
Others	6,625	166.54	5,377	165.99	4,183	165.52	−1.02	−0.47	−0.55
DK	3,883	164.10	3,045	163.34	2,070	162.61	−1.49	−0.73	−0.76

(continued)

Table 4 (continued)

Independent variables	Time period						Total change (20–49) years	Decadal changes	
	(20–29) years		(30–39) years		(40–49) years			(40–49) and (30–39) years (cm)	(30–39) and (20–29) years (cm)
	N	Mean	N	Mean	N	Mean			
Occupation									
Not working	2,593	166.63	235	164.42	216	164.39	−2.24	−0.03	−2.21
Prof/tech/Man	1,570	166.93	1,695	166.15	1,340	165.38	−1.55	−0.77	−0.78
Clerk/sales/service	5,127	165.51	4,608	165.32	3,400	164.66	−0.85	−0.66	−0.19
Agriculture	4,763	164.27	4,798	163.99	4,062	163.66	−0.61	−0.33	−0.28
Skilled/unskilled lab	7,760	164.39	6,163	164.21	4,054	163.57	−0.82	−0.64	−0.18
Wealth index									
Poorer	1,836	162.46	1,955	162.27	1,391	161.97	−0.49	−0.30	−0.19
Poorest	2,917	163.44	2,533	163.23	1,905	162.48	−0.96	−0.75	−0.21
Middle	4,527	163.88	3,505	163.82	2,540	163.39	−0.49	−0.43	−0.06
Richer	6,107	165.15	4,433	164.77	3,219	164.12	−1.03	−0.65	−0.38
Richest	6,454	167.34	5,098	166.67	4,040	165.97	−1.37	−0.70	−0.67

Table 5 Correlation between adult heights with age, wealth index and educational level of men and women in India

	Variables		
Height	Highest educational level	Wealth index	Age
Men	0.198** (52,442)	0.223** (52,460)	−0.066** (52,460)
Women	0.158** (94,408)	0.183** (94,417)	−0.030** (94,417)

**The correlation is significant at 0.01 level (figures in parentheses show the number of individuals for which data are available for both the variables of the correlation coefficient)

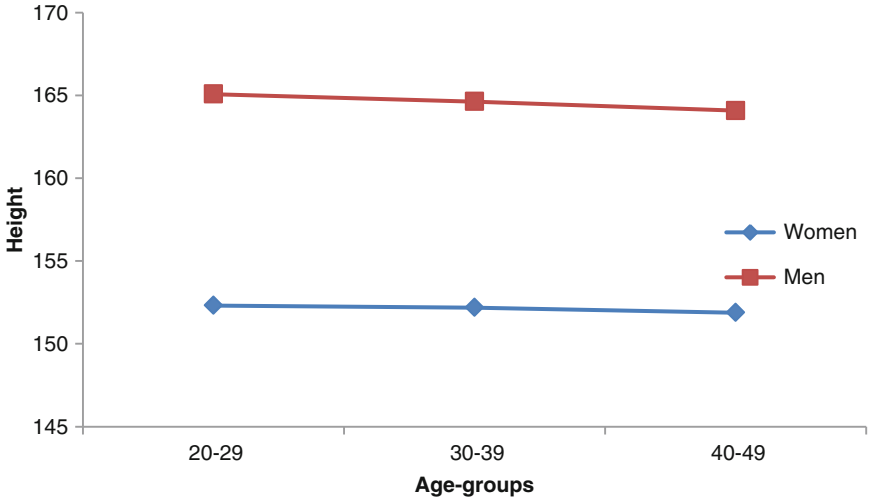


Fig. 2 Trend in the mean height of men and women over decadal age-groups

Male height (30–39 years)

$$\widehat{y} = \underset{(0.000)}{163.7} - \underset{(0.038)}{0.239}\, x_1 + \underset{(0.000)}{1.229}\, x_2 - \underset{(0.000)}{0.959}\, x_3 + \underset{(0.000)}{1.302}\, x_4 + \underset{(0.112)}{0.260}\, x_5 + \underset{(0.000)}{1.775}\, x_6 - \underset{(0.081)}{0.031}\, x_7. \quad (5)$$

Female height (20–29 years)

$$\widehat{y} = \underset{(0.000)}{150.7} - \underset{(0.000)}{0.514}\, x_1 + \underset{(0.000)}{1.099}\, x_2 + \underset{(0.237)}{0.101}\, x_3 + \underset{(0.000)}{0.846}\, x_4 - \underset{(0.330)}{0.065}\, x_5 + \underset{(0.000)}{1.623}\, x_6 - \underset{(0.860)}{0.002} x_7. \quad (6)$$

Male height (20–29 years)

$$\widehat{y} = \underset{(0.000)}{163.4} - \underset{(0.011)}{0.275}\, x_1 + \underset{(0.000)}{1.635}\, x_2 - \underset{(0.000)}{1.230}\, x_3 + \underset{(0.000)}{1.366}\, x_4 + \underset{(0.112)}{0.971}\, x_5 + \underset{(0.000)}{1.984}\, x_6 - \underset{(0.021)}{0.039}\, x_7. \quad (7)$$

(Figures in parentheses represent level of significance)

Table 6 Linear regression of height with different socio-economic variables for each decadal group of ages among adult females and males in India

		Dependent variables			
		Female height		Male height	
Decadal age-group	Explanatory variables	Coeff.	Sig*	Coeff.	Sig*
40–49 years	Intercept	154.6	0.000	165.0	0.000
	Place of residence	−0.668	0.000	−0.518	0.000
	Education	0.615	0.000	1.155	0.000
	Religion	0.758	0.000	−0.462	0.006
	Caste/tribe	0.983	0.000	1.451	0.000
	Respondent's occupation	−0.122	0.188	0.329	0.072
	Wealth Index	1.309	0.000	1.784	0.000
	Age	−0.088	0.000	−0.063	0.001
30–39 years	Intercept	151.6	0.000	163.7	0.000
	Place of residence	−0.445	0.000	−0.239	0.038
	Education	0.719	0.000	1.229	0.000
	Religion	0.178	0.065	−0.959	0.000
	Caste/tribe	0.763	0.000	1.302	0.000
	Respondent's occupation	−0.055	0.440	0.260	0.112
	Wealth index	1.441	0.000	1.775	0.000
	Age	−0.019	.096	−0.031	0.081
20–29 years	Intercept	150.7	0.000	163.4	0.000
	Place of residence	−0.514	0.000	−0.275	0.011
	Education	1.099	0.000	1.635	0.000
	Religion	0.101	0.237	−1.230	0.000
	Caste/tribe	0.846	0.000	1.366	0.000
	Respondent's occupation	0.065	0.330	0.971	0.000
	Wealth index	1.623	0.000	1.984	0.000
	Age	−0.002	0.860	−0.039	0.021

*The p-values are shown in this column

The results of the linear regressions can easily be understood if the values of the regressors are known. When we look at the regression results we see that some relations give different or opposite results than those obtained from taking the simple group means. The place of residence is consistently negatively related with height in the regression equation and the coefficient is significant in all the cases. Observe that we have taken the value 1 for urban and 0 for rural and the negative coefficient of place of residence clearly indicates that rural adults have more height if the effect of other variables is eliminated. The mean values of height in the rural and urban cases give the opposite results. The mean height of urban adults is always more than the mean height of rural adults for each combination of age-group and gender. Other regression coefficients, except religion, more or less give expected results. Age is seen to have a negative relation with height both for the regressions and for group averages. Wealthier or more educated persons have higher heights on the

average. Caste is also positively related with height. This means that General Caste Hindus, Christians, etc. have higher heights than SC, ST and OBC people. However, occupation is not significantly related with height for most cases. Religion needs special mention here, because it is positively related with height for females, but negatively related with height for males. This result conforms to the result of the corresponding group means.

4 Discussion

The paper investigates the changes in height vis-à-vis changes in age-groups of adult men and women in India taking three age-groups, namely (20–29), (30–39) and (40–49) years. The reduction in the average height is 0.12 cm from 20–29 years to 30–39 years aged women and 0.31 cm from 30–39 years to 40–49 years, the total reduction being 0.43 cm. So the study shows negative changes in the heights over the successive age-groups. The decadal changes are found to be negative in all the zones of India though it varies zone-wise greatly. Among the women, south zone shows the highest (1.32 cm) and north zone shows the lowest (0.10 cm) change. In most of the states, negative growth occurs but in a few states, positive growth occurs for both the gender groups. In case of men, the highest (1.79 cm) and the lowest (0.28 cm) changes are observed in west and north-east zones, respectively. When male and female heights are compared, the magnitude in the total change is found to be more in males more than in females. The changes in the heights have also been seen among the different socio-economic groups. It is seen in almost all cases that negative changes occur regardless whether it is found for men and women separately or found for all adults in India. It is also an interesting feature that in urban areas, among the richest and richer families, maximum negative increments occur, while among the manual labourers, and scheduled tribes, low magnitude of negative increment occurs. But it is firmly confirmed that height reduces with the advancement of age. Thus it propagates the idea that in human body, post adulthood changes do occur in height. It is supported by many findings (Miall et al. 1967; Roche et al. 1981; Malina et al. 1982; Kirchengast 1994). The most supporting relevant work staking Indian data are (Bagga 1998, 2013; Bagga and Sakurkar 2013).This type of study was mainly done in India or around the world during 1980s and 1990s and in that period, the difference was 5–7 cm (approximately) from younger to older generation, but in our study, the difference is found to be only 0.43 cm. It may be due to the fact that we have taken a smaller span of total years (20–49 years) compared to the time span (20–70 years) taken by them. But, even then, the change in height found by us is too less compared to the changes found by them. It is true that the degeneration starts after 40 years (Roche et al. 1981; Noppa et al. 1980; Sussame 1977; Cline et al. 1989). To understand the changes in the height over age, the span of age must be from 30 to 70 or 80 years. But here, the terminal point of age is 49 years only. As the data is from secondary sources, the male data is available up to 54 years and female data is available up to 49 years. So

we have taken all data from 20 years to 49 years to maintain the parity between male and female data. It needs further investigations to see when the degeneration starts and how much degeneration occurs. The effect of the socio-economic variables also needs to be further explored. Is it true that more changing lifestyle in a broad sense which includes changing food habits also result into more degeneration of height? We have in fact seen that more negative changes occur among the urban people, and richer and richest families as age increases.

Appendix

Data Type Unit level data as obtained from the third National Family Health Survey (NFHS – III) conducted by the International Institute for Population Sciences (IIPS), Mumbai, in 2005–2006.

Sample Size The sample sizes consist of 94,417 women and 52,460 men in the age-group 20–49 years. IIPS collected unit level data on reproductive aged men of age (15–54) years and women of age (15–49) years from 29 states in India. However, to maintain parity we have taken age range of (15–49) years for both males and females

Time span for total and consecutive period for Decadal changes: 20–49 years with three consecutive time span like (20–29), (30–39) and (40–49) years.

The Variables Considered in the Paper All the variables, except height, are grouped into categories. (For regression analysis the variables are treated in a different manner.)

Height: The height is measured in centimetres.
Age: (1) 20–29 years, (2) 30–39 years and (3) 40–49 years.
Place of residence: (1) Rural and (2) Urban areas.
Educational level: (1) Illiterate (those who can neither read nor write), (2) Primary level (literate up to class IV standard), (3) Middle level (Class V to class X standard) and (iv) High school & above (class XI and above).
Religion: (1) Hindu (2) Muslim (3) Christian and (4) Others,
Ethnicity: (1) Scheduled Castes (SC) (2) Scheduled Tribes (ST), (3) Other Backward Categories (OBC) and (4) Others.
Occupations: (1) Not working; (2) Professionals, managers and technicians, (3) Service or sales (4) Agriculture related works and (5) Skilled, unskilled or manual labourers, and
Wealth index of the families: (1) Poorest (2) poorer (3) Middle (4) Higher and (5) Highest. The details of how wealth index is classified into these categories are given in the main text.

The Variables Taken in the Linear Regression Analysis

The dependent variable is Height. All the independent variables, except age, are taken as binary variables where '0' is the base category and the rest of the categories are grouped and given the value '1'. Age is taken in years. We shall mention only the base categories below:

Place of residence: Rural;

Educational level: Primary level or less, i.e., Illiterate or literate up to class IV standard;

Religion: Hindu or Muslim;

Ethnicity: SC, ST or OBC;

Occupations: 'Service or sales', 'Agriculture related works or Skilled', 'Unskilled or manual laborers', i.e., Other than not working, professionals, technicians or managers; and

Wealth index of the families: Poorest, poorer or Middle income persons.

Age: Age is taken in years. It should be mentioned here that the regression analyses were performed separately for each group of (1) 20–29 years, (2) 30–39 years and (3) 40–49 years. Thus for the group 20–29 years, say, the age as an explanatory variable takes values from 20 to 29 years.

References

Bagga A (1998) Normality of ageing - a cross cultural perspective. J Hum Ecol 9:35–46

Bagga A (2010) Anthropological studies in the new millennium: biological anthropological research in gerontology. Indian Anthropol 40:1–24

Bagga A (2013) Age changes in some linear measurements and secular trend in height in adult Indian women. Acta Biol Szeged 57:51–58

Bagga A, Sakurkar A (2013) Women ageing and mental health. Mittal Publications, New Delhi

Cline MG, Meredith KE, Boyer JT, Burrows B (1989) Decline of height with age in adults in a general population sample: estimating maximum height and distinguishing birth cohort effects from actual loss of stature with aging. Hum Biol 61: 415–425

Harvey RG (1974) An anthropometric survey of growth and physique of the populations of KarkarLisland and LufaSubdistricts, NewBuinee. Phil Trns R Soc Lord B 268:279–292

International Institute for Population Sciences (IIPS) and ORC Macro (2007) National Family Health Survey (NFHS-3), 2005–2006, vol 1. IIPS, Mumbai

Kirchengast S (1994) Body dimensions and thyroid hormone levels in pre-menopausal and post-menopausal women from Austria. Am J Phys Anthropol 94:487–497

Malina RH, Buschang PH, Aronson WL, Selby HA (1982) Ageing in selected anthropometric dimensions in a rural Zapotec speaking community in the valley of Oaxaco Mexico. Soc Sci Med 16:217–222

Noppa H, Anderson M, Bengtsson C, Ake B, Isaksson B (1980) Longitudinal studies of anthropometric data and body composition, The population study of women in Goteberg, Sweden. Am J clin nutr 33:155–162

Miall WE, Ascheroft MT, Lovel HG, Moore F (1967) A longitudinal study of decline adult height with age in two welsh communities. Hum Biol 39:445–454

Roche AF, Garn SM, Reynold EL, Robinew M, Sontag LW (1981) The first seriatim study of human growth and middle ageing. Am J Phys Anthropol 54:23–24

Rutstein S (1999) Wealth versus expenditure: comparison between the DHS wealth index and household expenditures in your departments of Guatemala. ORC Macro, Calverton
Sharma A, Sapra P, Saran AB (1975) Effects of age-changes in some segmental measurements in Mundas/Oraons of Chotanagpur. Ind J Phys Anthropol Hum Genet 1:9–16
Sidhu LS, Sodhi NS, Bhatnagar DP (1975) Anthropometric changes from adulthood to old age. Ind J Phys Anthropol Hum Genet 1:119–127
Singh AP (1978) Effects of age changes in some somatic measurements in the adult Bhoska males of Nainital. Ind J Phys Anthropol 9:311–324
Susanne CF, Orbach HL (1977) Individual age changes of the morphological characteristics. J Hum Evol 6:181–189

Growth Model of Some Vernacular Word Usage During Political Transition

Ratan Dasgupta

Abstract We consider frequencies of some vernacular and other words having relevance in democratic elections. In periods of intense political activities, usages of certain words are frequent. Relative cumulative frequencies of some words and inflections appearing in a vernacular daily in West Bengal are modeled by a modification of Gompertz curve. Estimates of the relevant parameters are obtained from observed data over the period 2001–2010, covering several elections at state and the national level in India. Proliferation rates of the words indicate intensity of use and possibility of further appearance in subsequent news reporting, having impact on public opinion and poll results. The rates are calculated from observed data and compared with theoretical proliferation rates. The proposed growth model explains the data satisfactorily. Discrete versions of Gompertz and related models are considered in limiting form of the model parameters. Under certain assumptions Gompertz growth model is derived.

Keywords Relative cumulative frequency • Gompertz curve • Vernacular daily • Lowess regression • Spline regression • Proliferation rate

MS subject classification: Primary: 62G05, secondary: 62P25

1 Introduction

Words used in newspapers have applications e.g., in vocabulary learning and analysis of socio-economic-political scenarios. Frequencies of words used in specialised vocabularies are of interest in psycholinguistic research. Subtitle-based word frequency list is studied by Cuetos et al. (2011).

Transitions of political scenarios occurred in last several elections in both state and national levels in India. At that time some specific vernacular words together

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_10

with inflections were in frequent use in West Bengal and *Bengali* speaking areas. Appearances of some of these words are transitory in nature. With sudden and sharp rise in frequency of use mainly for a limited period, these reflect mass opinion with an impact on election results. We model frequencies of some such words along with other associated words of common use in election scenario, appearing during the period 2001–2010, in a vernacular daily of high circulation published from West Bengal.

Gompertz curve may sometimes be more appropriate model for phenomenon of sharp rise compared to other models like logistic. The model is generally used in bacterial growth, tumour-immune system, etc. See, e.g., d'Onofrio (2005). In treatment of malignant tumours following Gompertzian growth, O'Rourke et al. (2009) observed that nature of repopulation resulted in a poorer prognosis for the patient due to higher potential to repopulate the tumour. Berger (1981) noted that Gompertz model has an edge over its competitors in modeling plant diseases.

However, SSgompertz package in software R grossly overestimated the relative cumulative word frequencies in all the cases under present study, prompting for a modification of the Gompertz growth curve to be fitted only in a bounded time zone. Relative cumulative frequencies of each word over time are computed with respect to the total number ($n = 142985088$) of *all* words appearing in that particular daily newspaper during the years 2001–2010. Relative cumulative frequencies of words in log (-log) scale are shown in Fig. 1a–c.

We fit a modification of Gompertz growth curve to relative cumulative frequencies $y(t)$ of words over a segment of time. Specifically, we fit the Gompertz curve $y = y(t) = ae^{-b\exp(-ct)}, a > 0, b > 0, c > 0;\ t \in [1, 10]$. We estimate the parameters by the method of least squares to the transformed growth data $y \to y/a$, and approximate the asymptote of the curve by the largest value observed in data. The fit seems satisfactory in all the cases. Linearisation being achieved in $\log(-\log y)$ scale for transformed $y \in [0, 1]$, we do not attempt to fit other lower rate growth models to the observed data. See Fig. 2a–c.

Proliferation rate of the words used over time indicates how aggressively a word is used and how likely it is to appear in subsequent news reporting. This has application in impact analysis of public opinion affecting election results. Proliferation rates are calculated from observed data and compared with the model.

For growth observations recorded at discrete time, discrete equivalent of proliferation rates are relevant, e.g., see Novile et al. (1982). We consider discrete versions of growth curves relevant to Gompertz model while studying models in terms of limiting behaviour of parameters. Gompertz growth curve is derived from some basic considerations.

The paper is arranged as follows. In Sect. 2 we observe some features of data on word frequencies and search for a model. In Sect. 3 we discuss Gompertz curve fitting in a bounded time zone. Behaviour of proliferation rates in related models along with discrete versions of these is also discussed in the section. Gompertz growth curve is derived in Sect. 4.

2 Some Features of Data and Search for a Model

As already mentioned, data were collected on 12 words together with inflections used in a widely circulated daily vernacular newspaper of West Bengal, frequencies are obtained from electronic version of the newspaper.

The words are *Manmohan* (name of ex. Prime Minister of India), *Modi* (present PM of India), *Congress* (name of a political party), *Trinamul* (grass-root, name of a political party), *Harmad* (goons, antisocial elements, armed cadres, etc.), *BJP* (name of a political party), *Sonia* (Congress leader), *CPM* (name of a political party), *Paribartan* (change), *Singur* (name of a place, where land acquisition from farmers for industry led to agitation), *Andolan* (agitation), *Ma-Mati-Manush* (mother-land-people; a term coined during political campaigns).

Relative cumulative frequencies of each word over time are computed with respect to the total number of *all* words appearing in that daily, during the years 2001–2010. Each word with successive appearances over time may have cumulative effect. Division by the total number of words ($n = 142985088$) in that period helps to analyse the frequencies on equal footing as bounded variables. Relative cumulative frequency of 12 words over 10 years is given in Table 1. First 4 years' entries of the word *Ma-Mati-Manush* are zero. Initially not much in use, the word made a sharp increase in frequency during 2008–2009 and then remained steady; see the uppermost curve in Fig. 1a. All the curves of relative cumulative frequencies in Fig. 1a are plotted in $\log(-\log)$ scale to check appropriateness for a Gompertz growth, the curves lie within the band of *Ma-Mati-Manush* and that for *Congress,* one of the oldest political parties. Linear trend in data points is quite prominent for some words, e.g., *CPM,* and *Trinamul*. Presence of approximate linearity indicates sharp growth, as in a Gompertz model.

During the year 2005–2006, no change in cumulative frequency is seen for the word *Harmad*. This word and the word *Ma-Mati-Manush* seem transient in nature.

3 Fitting the Model and Some Theoretical Issues

Since the rise of cumulative frequencies are steep, Gompertz growth curve is a candidate model. Discrete versions of the growth curve become relevant when observations are taken in discrete time. Proliferation rates and its analogue in discrete case for these models are of interest.

We assume that cumulative frequencies of words do not blow up in the limit $t \to \infty$, for which no normalisation is possible. For the time being, we assume that the rise of cumulative frequencies is not that high so that we are able to deal with bounded quantilies like relative cumulative frequencies $y(t)$.

Table 1 Relative cumulative frequency of 12 vernacular words over 10 years

	Relative cumulative frequency of vernacular words											
Year	Manmohan	Modi	Congress	Trinamul	Harmad	BJP	Sonia	CPM	Paribartan	Singur	Andolon	Ma-mati-manush
2001	0.0000006710	0.0000010000	0.0000639000	0.0000000979	0.0000000280	0.0000033200	0.0000102000	0.0000041500	0.0000109802	0.0000002588	0.0000156590	0.0000000000
2002	0.0000018500	0.0000183000	0.0001290000	0.0000001890	0.0000000280	0.0000198000	0.0000210000	0.0000092300	0.0000239955	0.0000006644	0.0000403609	0.0000000000
2003	0.0000026100	0.0000223000	0.0002390000	0.0000002940	0.0000000350	0.0000441000	0.0000376000	0.0000196000	0.0000382907	0.0000016156	0.0000703780	0.0000000000
2004	0.0000251000	0.0000305000	0.0004100000	0.0000006220	0.0000000420	0.0001000000	0.0000699000	0.0000395000	0.0000547190	0.0000021890	0.0001076826	0.0000000000
2005	0.0000487000	0.0000377000	0.0005570000	0.0000009230	0.0000000490	0.0001350000	0.0000854000	0.0000975000	0.0000710214	0.0000030143	0.0001514633	0.0000000140
2006	0.0000685000	0.0000417000	0.0006800000	0.0000011100	0.0000000490	0.0001700000	0.0001030000	0.0001840000	0.0000870231	0.0000174424	0.0001939853	0.0000000210
2007	0.0000822000	0.0000529000	0.0007760000	0.0000013600	0.0000001260	0.0002040000	0.0001160000	0.0002990000	0.0001022274	0.0000417036	0.0002378150	0.0000000280
2008	0.0001040000	0.0000619000	0.0009630000	0.0000019100	0.0000002170	0.0002490000	0.0001310000	0.0005420000	0.0001264118	0.0000801412	0.0003062487	0.0000000629
2009	0.0001310000	0.0000732000	0.0012500000	0.0000028100	0.0000005460	0.0003180000	0.0001520000	0.0008700000	0.0001525194	0.0000943945	0.0003727801	0.0000005735
2010	0.0001560000	0.0000947000	0.0015300000	0.0000038300	0.0000015200	0.0003820000	0.0001710000	0.0011700000	0.0001874601	0.0001049130	0.0004516695	0.0000005735

3.1 Fitting the Model

In Fig. 1a behaviour of relative cumulative frequencies $y(t)$ in a transformed scale is explained. This indicates approximate linear relationship of the transformed variables with time, as expected in a Gompertz curve; the figure becomes clumsy as cluster of lines appear towards bottom of the graph. To understand the pattern better, we plot curves for six words in Fig. 1b, and plot the remaining six curves in Fig. 1c; the pictures of linear relationship become clearer.

The word *Ma-Mati-Manush* made its appearance in the year 2005. As such there are six points only in the corresponding curve with equal values in last two years, indicating transient appearance of the word.

Usual fit for Gompertz curve $y(t) = ae^{-b\exp(-ct)}, a > 0, b > 0, c > 0$ over the entire range $t \geq 0$ do not perform well for the word *Paribartan*; the package SSgompertz in R provides the following values of parameters $a = 0.1002742, b = 0.00001875003, c = 0.00000001152055$. The fit is bad as seen from Fig. 2d, the model overestimates the observed values. The same holds true for other words as

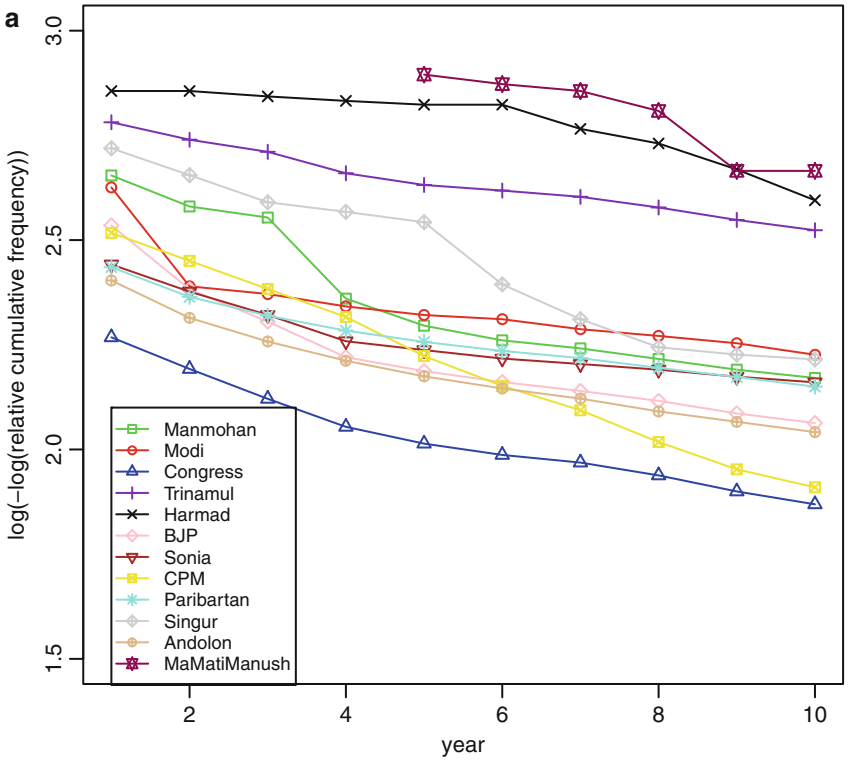


Fig. 1 (**a**) Gompertz model for 12 words appearing in a vernacular daily. (**b**, **c**) Gompertz model for six words appearing in a vernacular daily

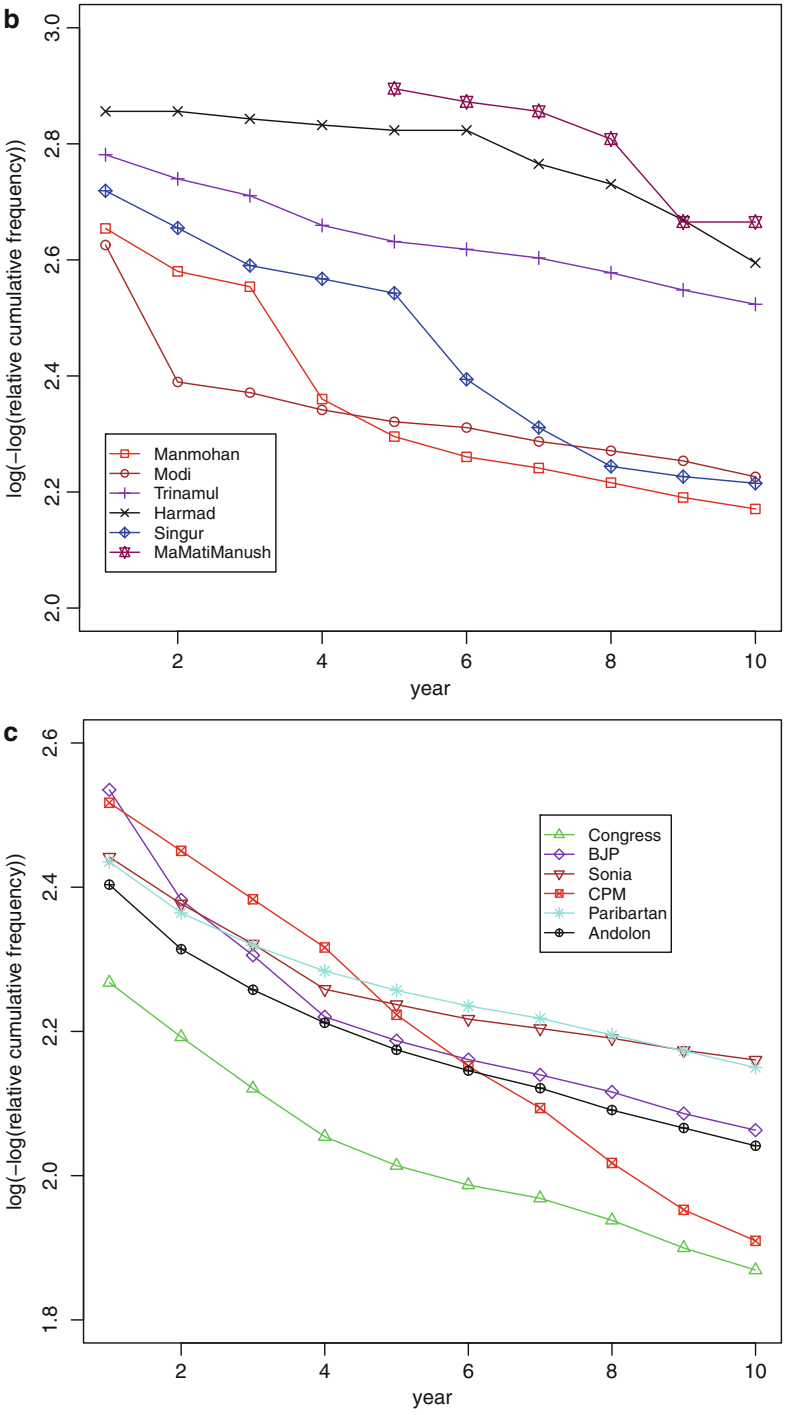


Fig. 1 (continued)

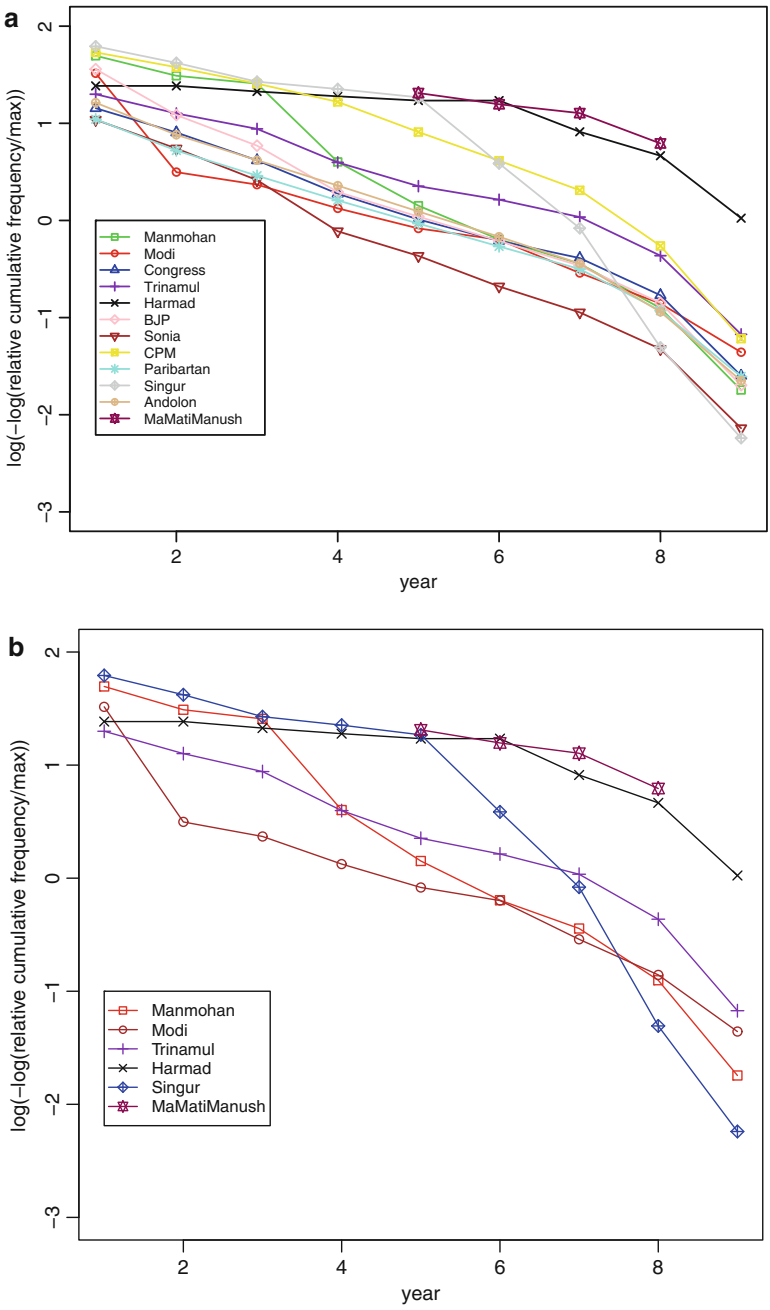


Fig. 2 (**a**) Gompertz model for 12 words: asymptote estimated as max. (**b**, **c**) Gompertz model for six words: asymptote estimated as max. (**d**) Gompertz and modified model for the word "Paribartan"

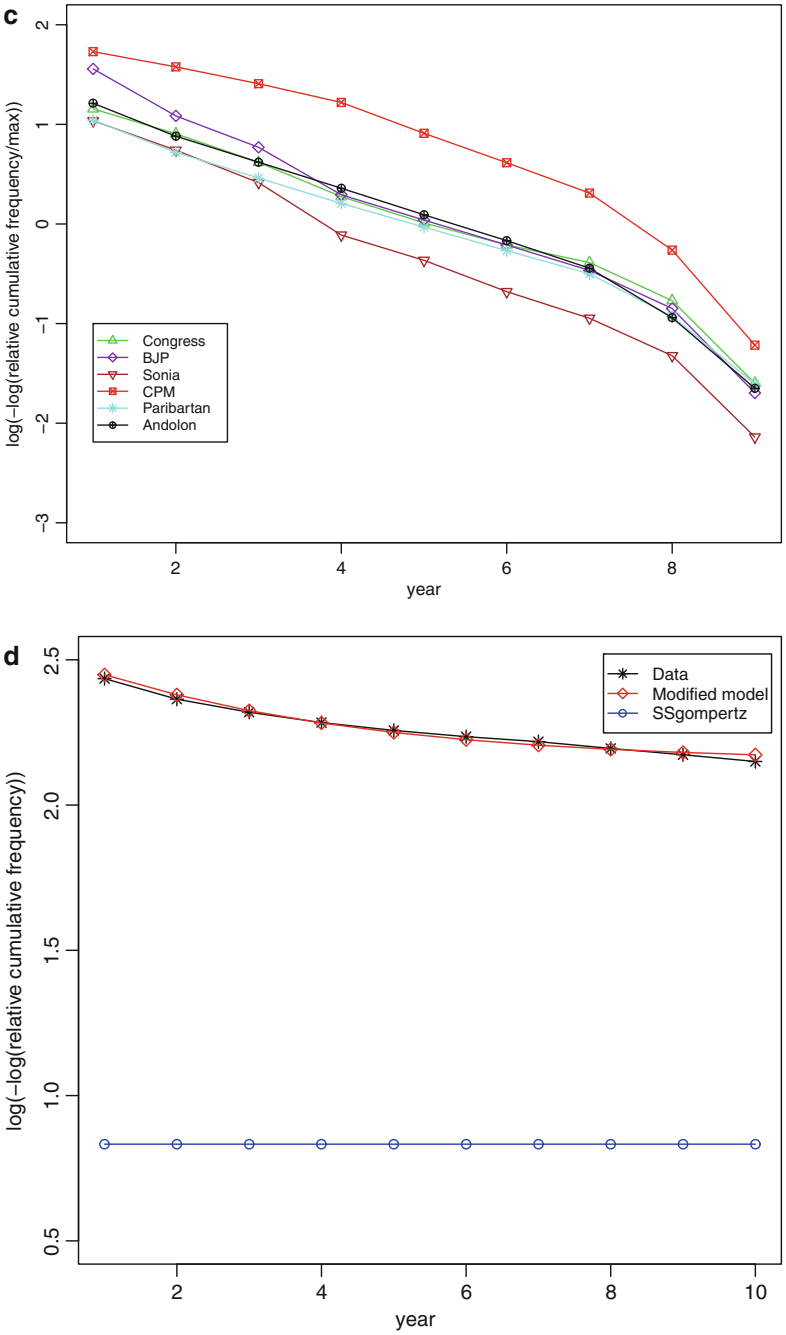


Fig. 2 (continued)

well. The package provides a high value of the asymptote $a = 0.1002742$, whereas the estimate of asymptote taken as the highest observed value from Table 1 for the word *Paribartan* is $a = 0.0000109802$.

An alternative method is to explore fitting the model on observed time range with estimate of asymptote taken as the highest observed value from Table 1 for the concerned word. The parameter $a(> 0)$ in Gompertz model represents the limiting value of the increasing curve. One of the reasons SSgompertz fails to model the data on time range $t > 0$ with overestimated model values could be that the limit a is approached for large t, and transient nature of some of the words may cause this assumption to remain unfulfilled.

To overcome the problem, we estimate the asymptote a by the observed highest value i.e., the value at the largest time $t = 10$ and then estimate the remaining two parameters of the (modified) model from least squared linear regression of $\log(-\log(y/a))$ values on time t. The intercept and slope of regression line provide estimates for b and c.

The red curve in Fig. 2d represents the modified Gompertz model fitted to the word *Paribartan*. The model fit seems good to the observed data represented by black curve. The red curve and black curve are almost indistinguishable, compared to the distant blue curve derived from SSgompertz at the bottom of Fig. 2d.

Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14 refer to modified Gompertz model and show least squared regression fit for 12 words. The estimated values of the parameters of the fitted model $y(t) = ae^{-b\exp(-ct)}, a > 0, b > 0, c > 0; t \in [1, 10]$

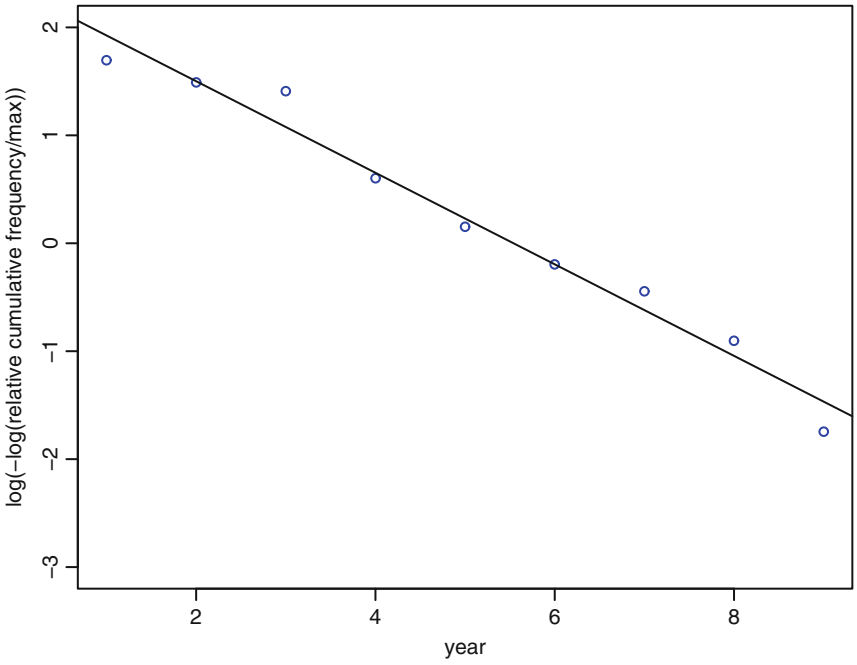


Fig. 3 Gompertz fit for relative cumulative frequency of the word "Manmohan"

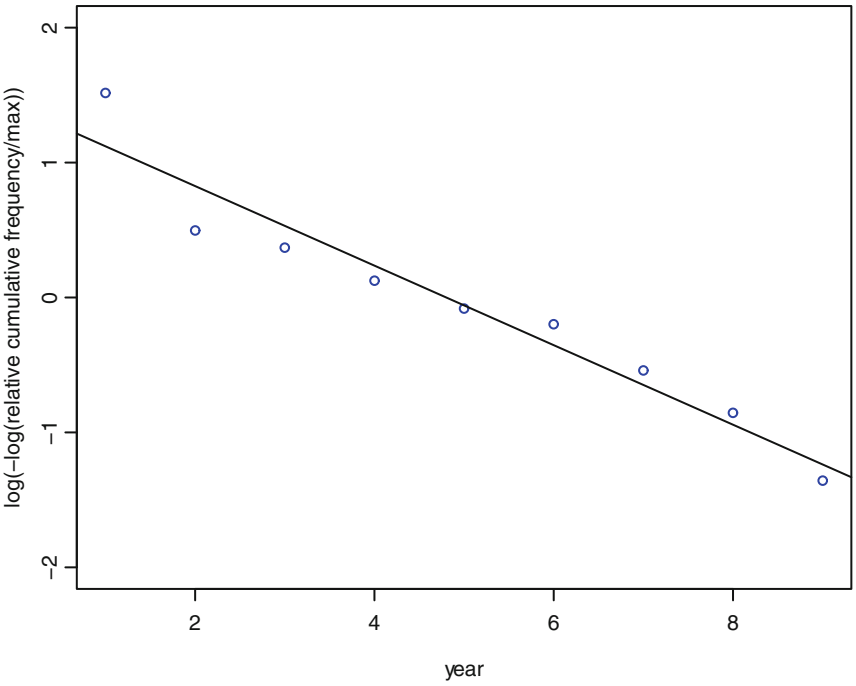


Fig. 4 Gompertz fit for relative cumulative frequency of the word "Modi"

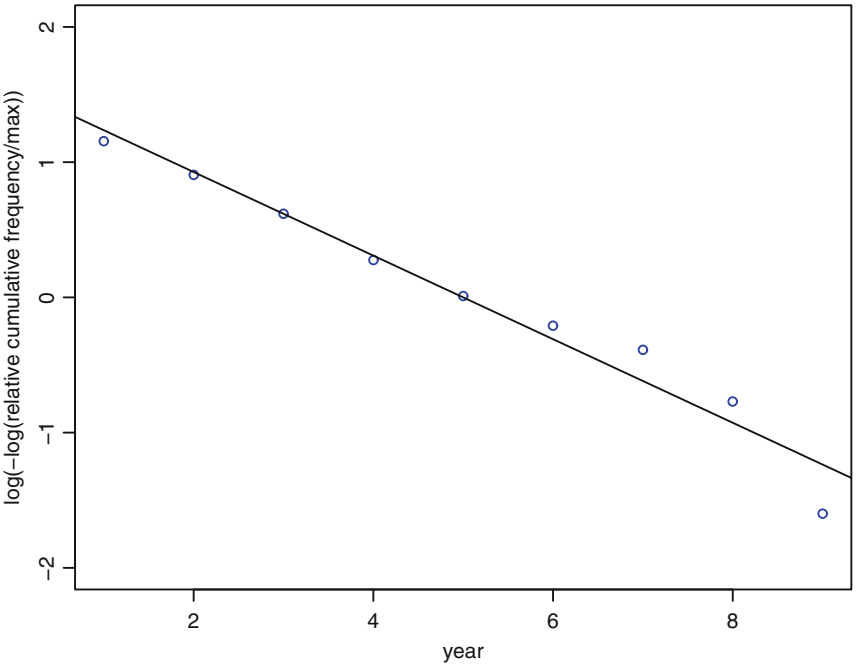


Fig. 5 Gompertz fit for relative cumulative frequency of the word "Congress"

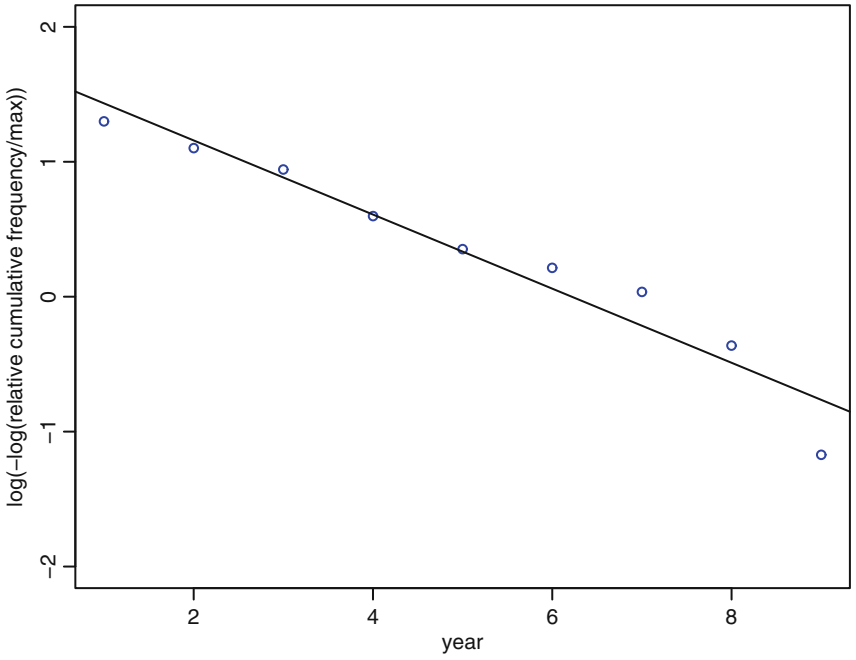


Fig. 6 Gompertz fit for relative cumulative frequency of the word "Trinamul"

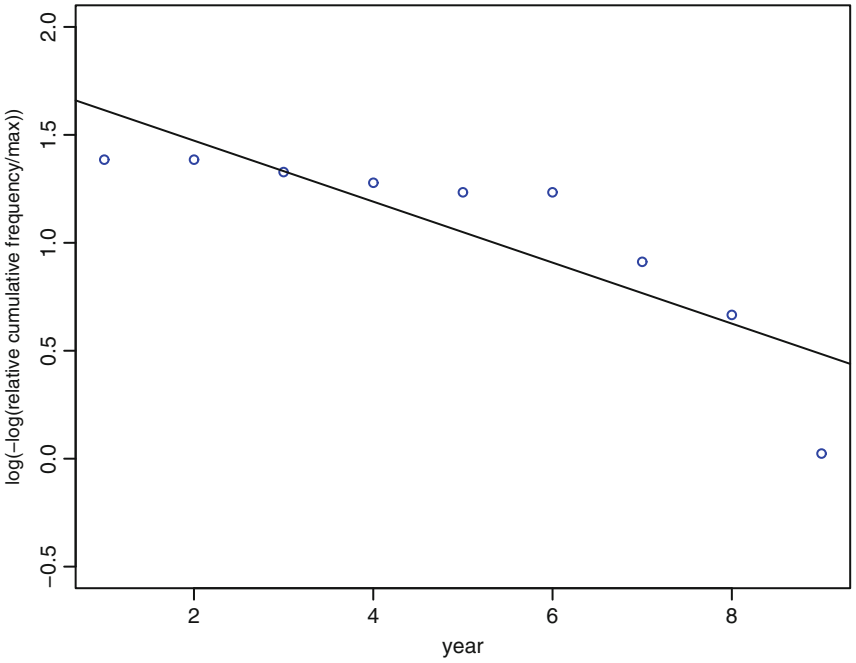


Fig. 7 Gompertz fit for relative cumulative frequency of the word "Harmad"

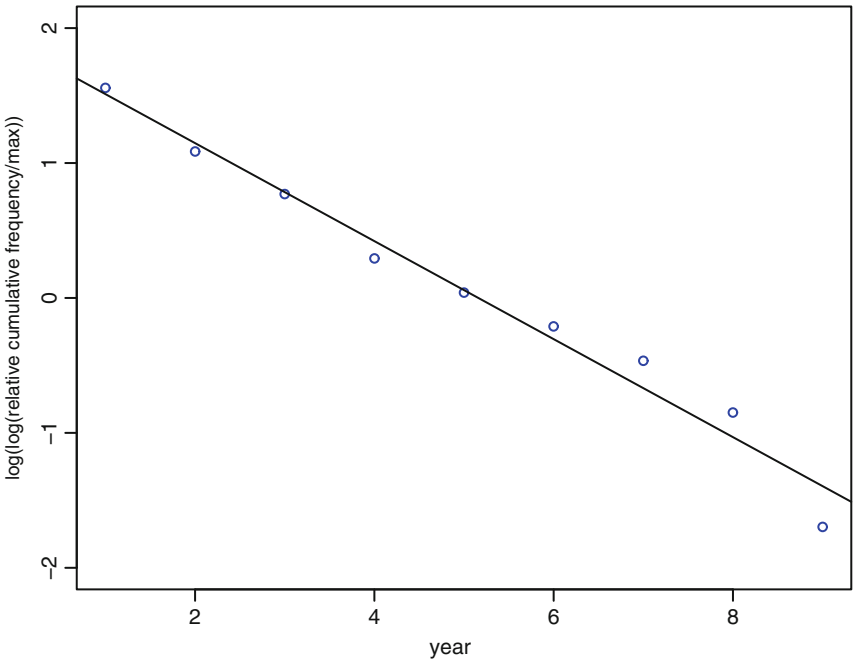


Fig. 8 Gompertz fit for relative cumulative frequency of the word "BJP"

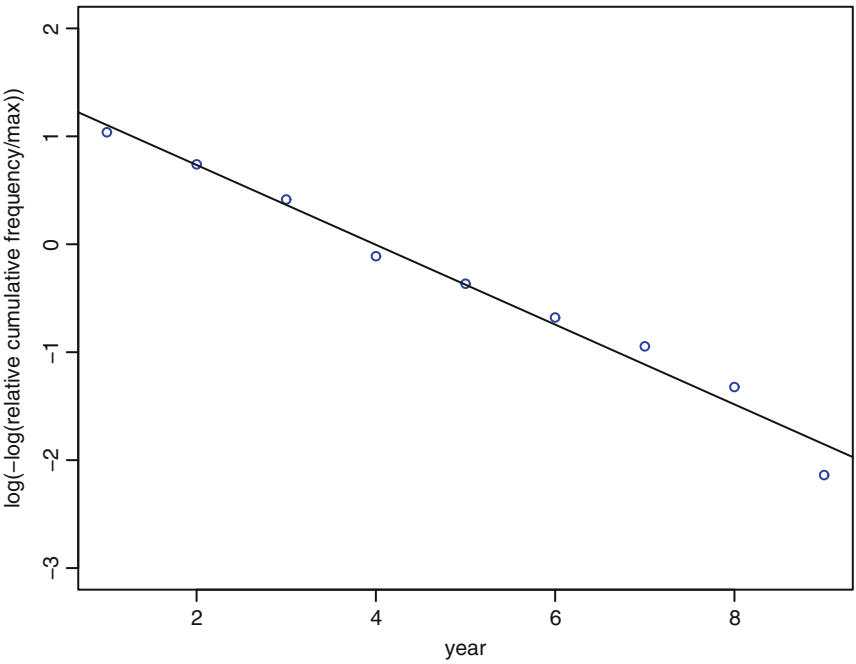


Fig. 9 Gompertz fit for relative cumulative frequency of the word "Sonia"

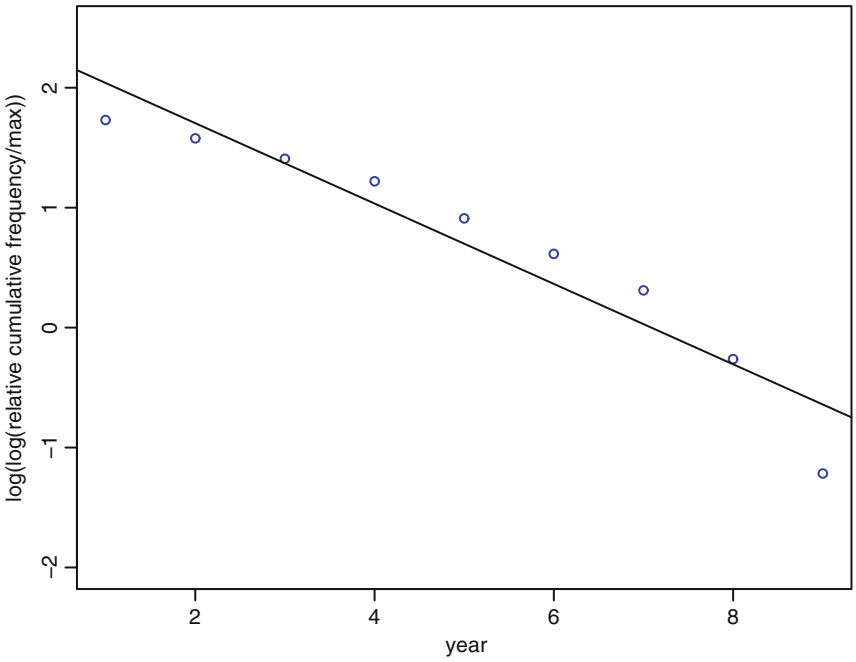


Fig. 10 Gompertz fit for relative cumulative frequency of the word "CPM"

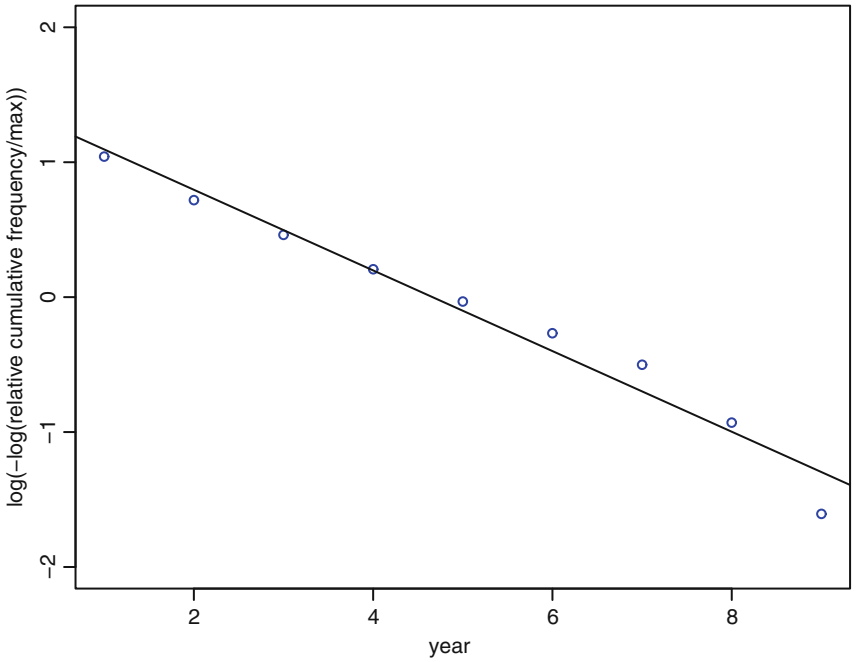


Fig. 11 Gompertz fit for relative cumulative frequency of the word "Paribartan"

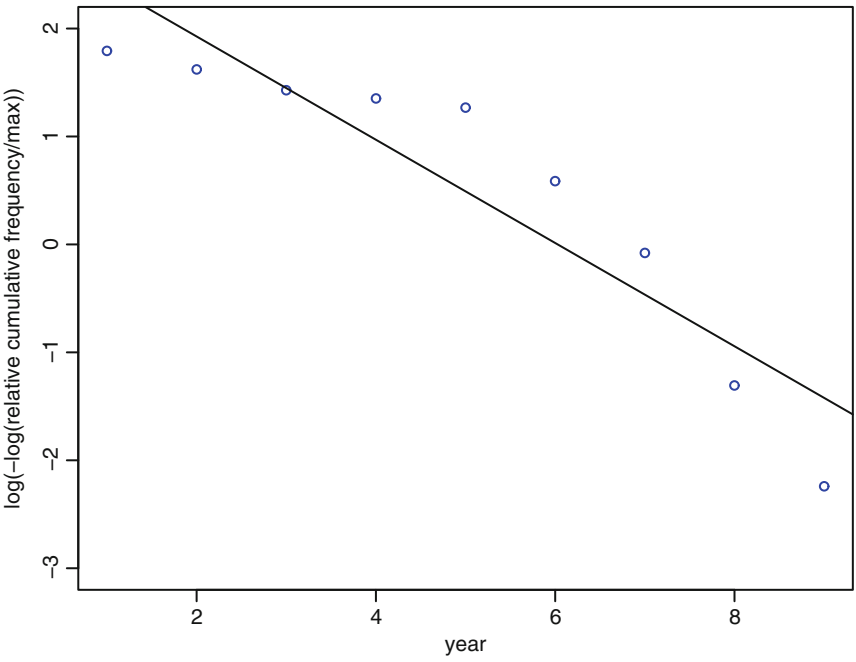


Fig. 12 Gompertz fit for relative cumulative frequency of the word "Singur"

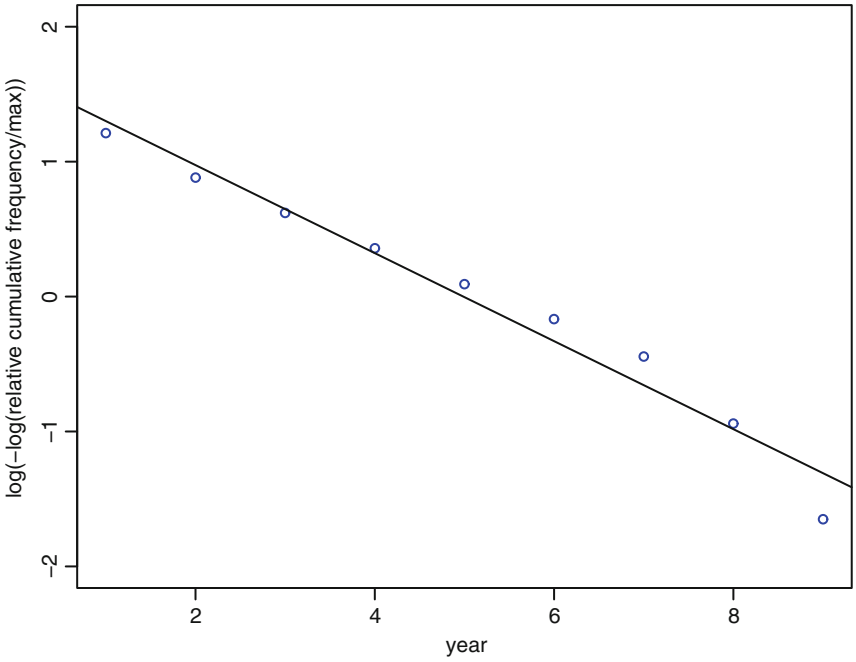


Fig. 13 Gompertz fit for relative cumulative frequency of the word "Andolon"

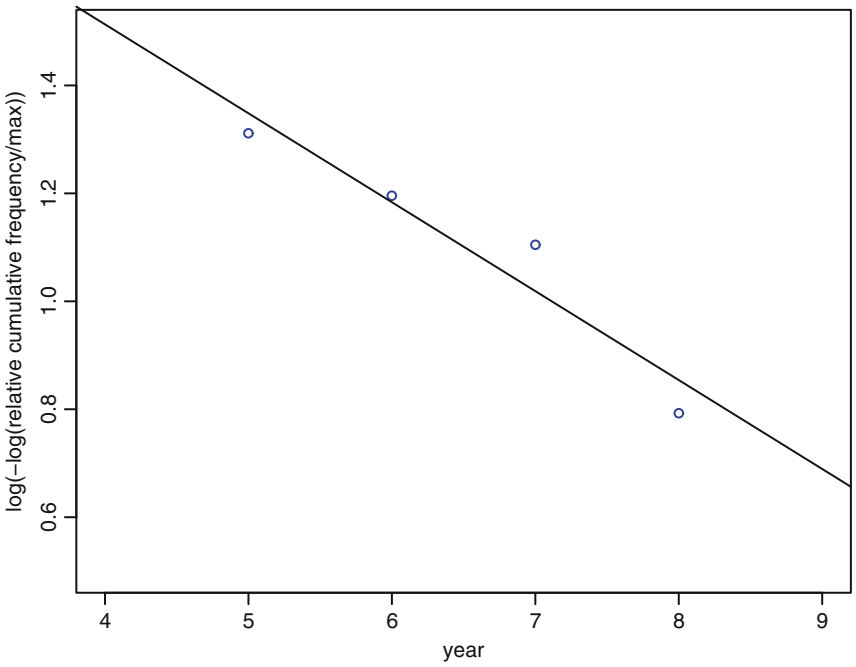


Fig. 14 Gompertz fit for relative cumulative frequency of the word "MaMatiManush"

are given in Table 2. The least squared fit is remarkably good in Figs. 3, 4, 5, and 6. The fit is moderate in Fig. 7 that refers to the word *Harmad*, a word of transient nature. Least squared fit is remarkably good in Figs. 8 and 9. For the word *CPM* the fit is good as seen in Fig. 10, although there is a tendency of sharp rise in frequency towards tail. The same can be said about the word *Singur*, having a sharp increase of frequency towards tail as shown in Fig. 12. For the word *Paribartan*, the least squared fit shown in Fig. 11 is very good.

The coefficient of determination r^2 given in the last row of Table 2 is high for all the words, indicating that the modified Gompertz model fits the data well.

Observed values of relative cumulative frequency are given in Table 1 for 12 words, and Table 3 provides relative cumulative frequency of these words over 10 years under fitted Gompertz model. The values obtained from model provide a glimpse of political atmosphere prevailing at that period and may be used in prediction purposes.

We also consider the problem of estimating the proliferation rate of some of the words. To this end, for the word *Manmohan*, observed values of $y(t)$ from Table 1 are lowess smoothed with parameter $f = 1/5$. Then with exponentially decaying normalised weights, ten crude individual slope estimates at a time point are obtained and divided by y value at that time point; and the resultant values are ordered from lowest to largest. Consider these as point estimates on proliferation at a time-point. The median of these ten values (trimmed mean of 5-th and 6-th order statistics) for each time point are then smoothed by SPlus package smooth.spline

Table 2 Estimated parameters of modified Gompertz curve $y = a\, e^{-b \exp(-ct)}$, $0 \leq t \leq 10$; and model accuracy r^2

	Word											
Parameter	Manmohan	Modi	Congress	Trinamul	Harmad	BJP	Sonia	CPM	Paribartan	Singur	Andolon	Mamatimanush
a	0.0001560000	0.0000947000	0.0015300000	0.0000038300	0.0000015200	0.0003820000	0.0001710000	0.0011700000	0.0001874601	0.0001049130	0.0004516695	0.0000005735
b	10.47718	4.117722	4.6883	5.514605	5.787498	6.513585	4.366405	10.74865	4.029289	17.87137	5.086907	8.776784
c	0.42405	0.2948	0.30902	0.27464	0.14126	0.36319	0.36971	0.33509	0.29889	0.4783	0.32618	0.16475
r^2	0.9731	0.9353	0.9618	0.9391	0.7297	0.9757	0.9811	0.9098	0.9689	0.8461	0.9671	0.9146

Table 3 Relative cumulative frequency of 12 words over 10 years under Gompertz model

	Relative cumulative frequency of words											
	Word											
Year	Manmohan	Modi	Congress	Trinamul	Harmad	BJP	Sonia	CPM	Paribartan	Singur	Andolon	Mamatimanush
2001	0.0000001643	0.0000044120	0.0000489614	0.0000000580	0.0000000100	0.0000041186	0.0000083709	0.0000005361	0.0000094431	0.0000000016	0.0000114944	0.0000000003
2002	0.0000017563	0.0000096527	0.0001222453	0.0000001586	0.0000000194	0.0000163640	0.0000212676	0.0000047854	0.0000204367	0.0000001094	0.0000319321	0.0000000010
2003	0.0000082801	0.0000172921	0.0002393173	0.0000003408	0.0000000344	0.0000427130	0.0000405049	0.0000229046	0.0000362310	0.0000014877	0.0000667523	0.0000000027
2004	0.0000228411	0.0000266933	0.0003918873	0.0000006093	0.0000000567	0.0000832409	0.0000632156	0.0000701955	0.0000553993	0.0000075021	0.0001136513	0.0000000061
2005	0.0000443706	0.0000368820	0.0005628724	0.0000009475	0.0000000874	0.0001323923	0.0000859793	0.0001563903	0.0000759068	0.0000204511	0.0001668628	0.0000000122
2006	0.0000685185	0.0000469221	0.0007342716	0.0000013251	0.0000001273	0.0001828162	0.0001063363	0.0002773663	0.0000958782	0.0000380782	0.0002201528	0.0000000219
2007	0.0000910542	0.0000561362	0.0008925114	0.0000017098	0.0000001765	0.0002288148	0.0001231533	0.0004178740	0.0001140121	0.0000559766	0.0002688985	0.0000000359
2008	0.0001096760	0.0000641546	0.0010300055	0.0000020752	0.0000002344	0.0002674667	0.0001363024	0.0005602192	0.0001296420	0.0000710761	0.0003106534	0.0000000547
2009	0.0001238780	0.0000708611	0.0011442571	0.0000024042	0.0000002999	0.0002981340	0.0001461989	0.0006909098	0.0001426019	0.0000824160	0.0003447595	0.0000000782
2010	0.0001341529	0.0000763068	0.0012361276	0.0000026887	0.0000003714	0.0003215114	0.0001534534	0.0008027062	0.0001530432	0.0000903357	0.0003716764	0.0000001058

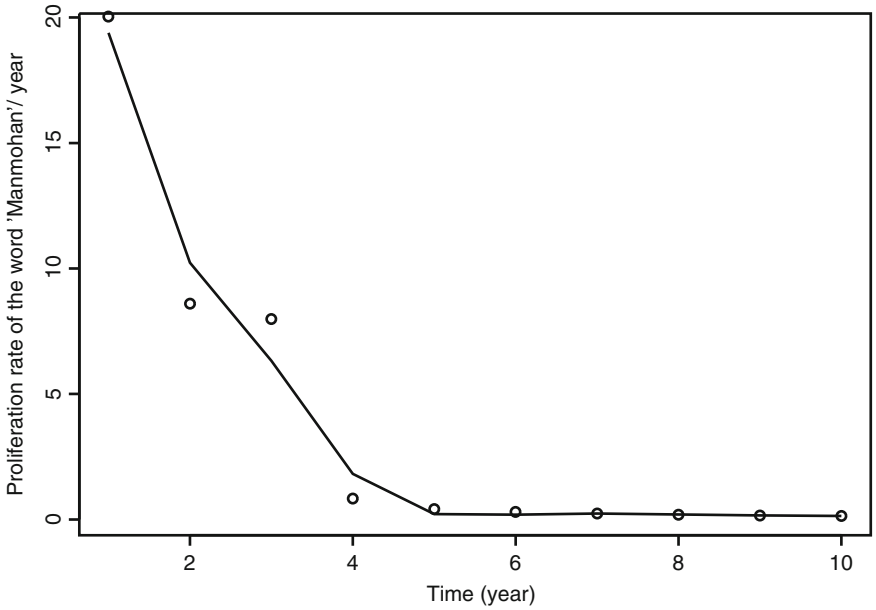


Fig. 15 Proliferation rate of "Manmohan" with trimmed mean, wt. e^(−.01 x); spline

with parameter spar = 0.0001. The resultant proliferation curve thus estimated for the word *Manmohan* is shown in Fig. 15. See Dasgupta (2013) for details of the adopted technique.

In a similar manner proliferation rates of the words *Paribartan, Harmad* and *Trinamul* are computed, these are shown in Figs. 16, 17 and 18, respectively. The transitory nature of the word *Harmad* is reflected in haphazard behaviour of the curve shown in Fig. 17. The curve may be rectified if the modeled values of $y(t)$ for the word *Harmad* are considered from Table 3, for computing proliferation. The resultant curve from modeled values is obtained in Fig. 19, following a similar procedure described above for observed data sets.

Figures 16 and 18 of proliferation rates computed from data for the words *Paribartan* and *Trinamul*, respectively, mimic the theoretical proliferation rate of a Gompertz model. The modeled proliferation rate the word *Ma-Mati-Manush* are computed from reconstructed y values given in the last column of Table 3. The smooth curve is shown in Fig. 20.

Instead of considering model fitting in a finite range, unrestricted $t \in (0, \infty)$ may be considered via the above proposed method. The parameter a is then underestimated. In Fig. 2d the curve corresponding to fitted model lies above that of data towards the end, indicating that the theoretical values by Gompertz model before transformation are smaller than the data points for large t. The Gompertz model on unrestricted time zone provides a bad fit (see the blue curve, and the distant black curve corresponding to data), in contrast to a limited range Gompertz fit, which seems appropriate for observed data.

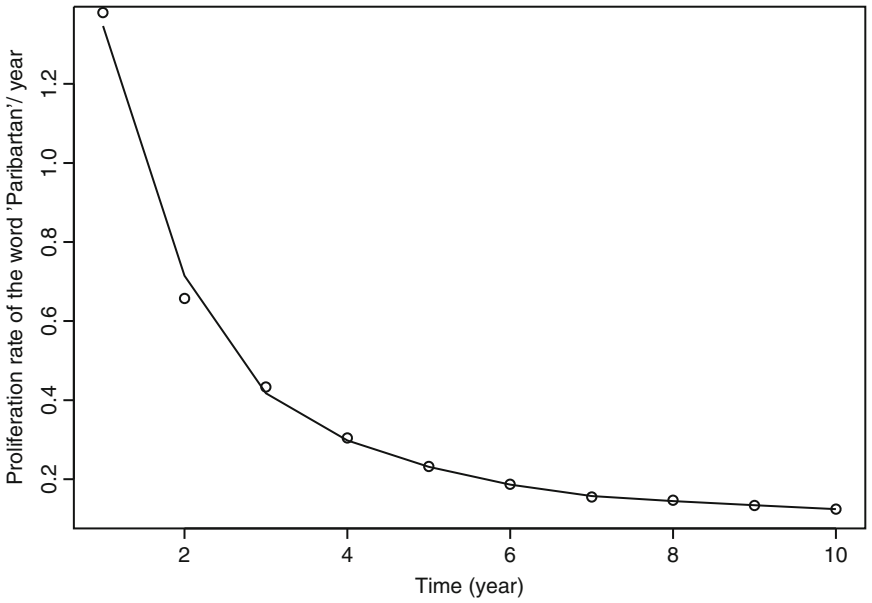


Fig. 16 Proliferation rate of "Paribartan" with trimmed mean, wt. e^(−.01 x); spline

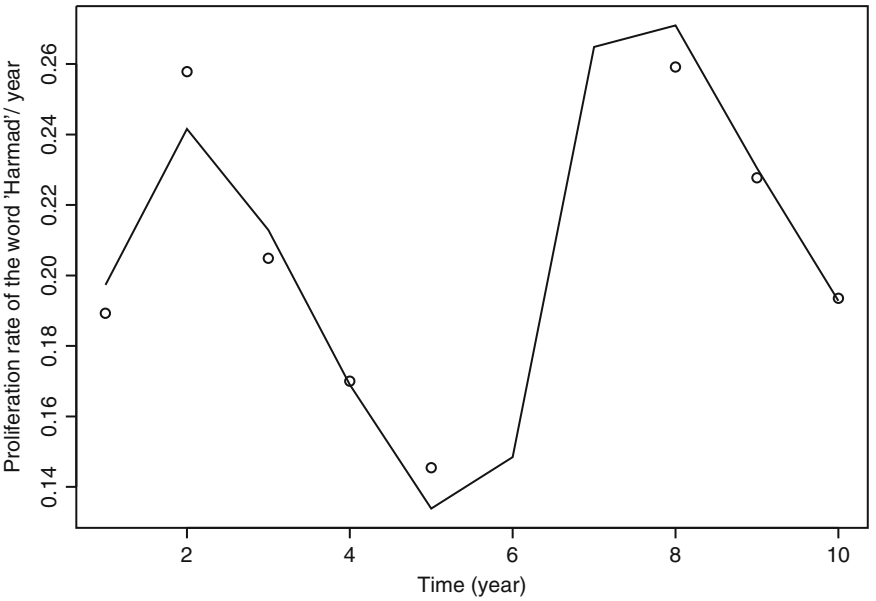


Fig. 17 Proliferation rate of "Harmad" with trimmed mean, wt. e^(−.01 x); spline

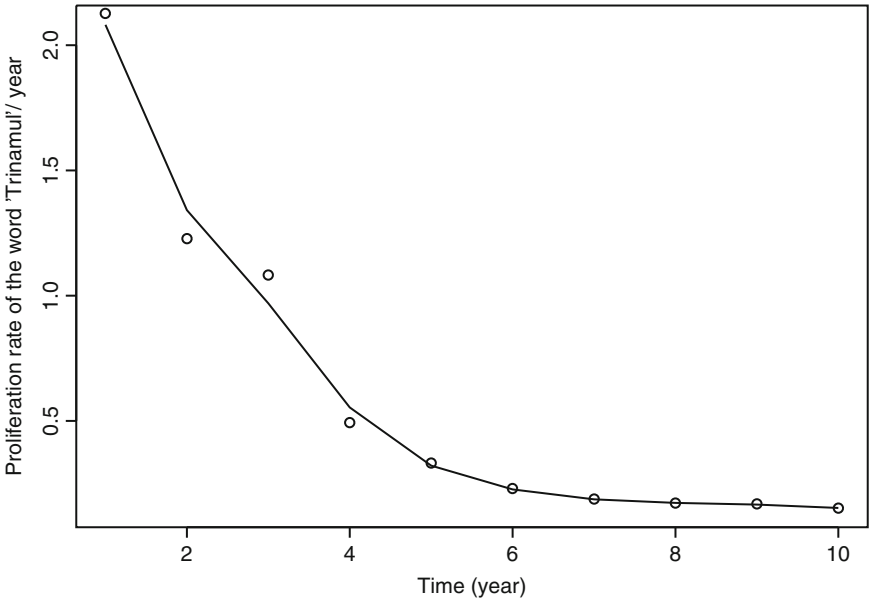


Fig. 18 Proliferation rate of "Trinamul" with trimmed mean, wt. e^(−.01 x); spline

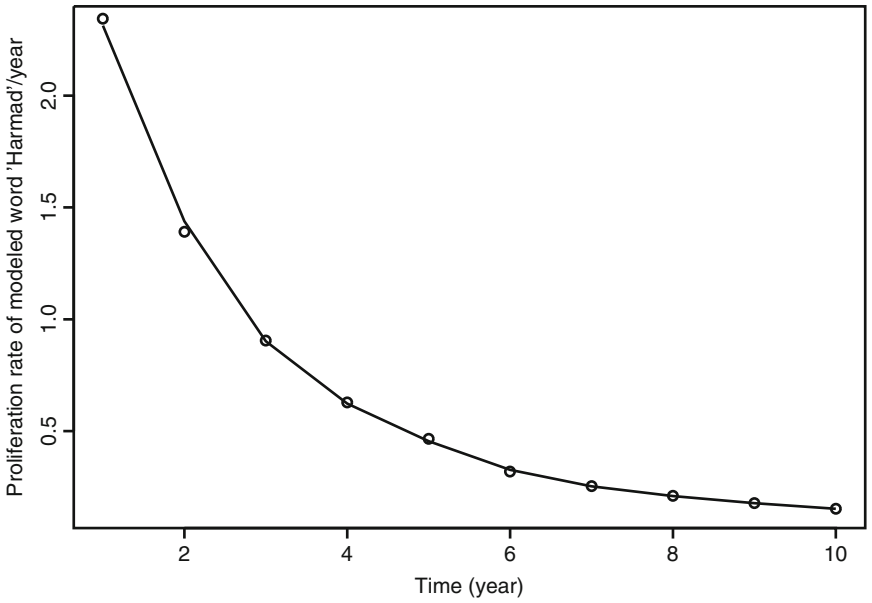


Fig. 19 Proliferation rate of modeled word "Harmad", ref. Fig. 17

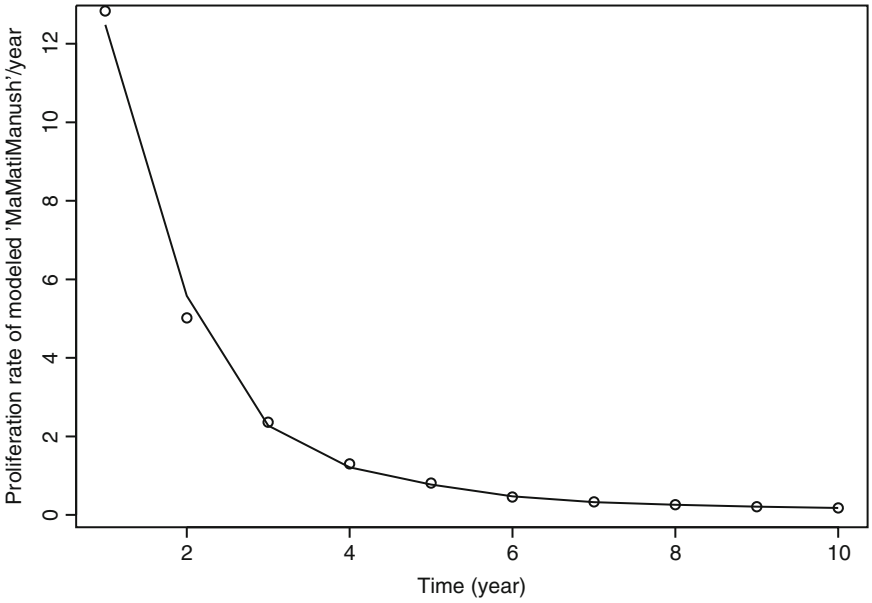


Fig. 20 Proliferation rate of modeled "MaMatiManush"

3.2 Some Discrete and Continuous Models

Proliferation rate of Gompertz curve may be considered as a limit of proliferation rate in logistic model. These distributions along with discrete versions of these are interconnected by model parameters. Gompertz and other models in discrete time are studied by difference equations on growth.

3.2.1 Gompertz Curve, Discrete Version and Other Related Growth Models

Proliferation rate $c \log(a/y(t))$ for Gompertz curve is relatively slow in decay than that of generalised logistic function having proliferation rate $c\nu[1 - \{\frac{y(t)}{a}\}^{1/\nu}]$. The former is logarithmically decaying with growth $y(t)$, whereas the latter is polynomially decaying.

An exponentially decaying proliferation rate may reduce to polynomial decay in limiting form of the model parameters $a > 0, c > 0, \beta > 0, \nu > 0$, see, e.g., Dasgupta (2013).

$$c\nu\beta[1 - e^{\{(y(t)/a)^{1/\nu}-1\}/\beta}] \to c\nu[1 - \{\frac{y(t)}{a}\}^{1/\nu}],\ \beta \to \infty,$$
$$\to c \log(a/y(t)),\ \nu \to \infty,\ t > 0 \qquad (1)$$

One may then consider discrete version $\Delta y(t)/\{y(t)\Delta t\}$ of the three proliferation rates given in (1) for continuous curves, which gives rise to the discrete growth curves for $t \in N = \{1, 2, 3, \ldots\}$ satisfying the following similar *difference equations* in (2).

$$\frac{y(t+1)}{y(t)} - 1 = c\nu\beta[1 - e^{\{(y(t)/a)^{1/\nu}-1\}/\beta}] \to c\nu[1 - \{\frac{y(t)}{a}\}^{1/\nu}],\ \beta \to \infty,$$
$$\to c\log(a/y(t)),\ \nu \to \infty \quad (2)$$

That is, for $t \in N_0 = \{0, 1, 2, 3, \ldots\}$

$$y(t+1) = [1 + c\nu\beta\{1 - e^{\{(y(t)/a)^{1/\nu}-1\}/\beta}\}]y(t) \quad (3)$$

$$y(t+1) = [1 + c\nu\{1 - (\frac{y(t)}{a})^{1/\nu}\}]y(t) \quad (4)$$

$$y(t+1) = [1 + c\log(a/y(t))]y(t) \quad (5)$$

For growth observations recorded at discrete time, the above family of models (3)–(5) with some assigned initial value for $y(0)$ may be appropriate. This family covers a broad spectrum of discrete proliferation rates, starting from exponential to logarithmic order of decay, with no "gap" in between, as the next class in the series (3)–(5) is a limit of the former.

4 Derivation of Gompertz Model

Gompertz model may be derived from basic considerations on rate behaviour. Assume that the growth $y = y(t)$ has a limiting value a and the proliferation rate, being independent of unit of measurements, is a decreasing function of $y^* = \frac{y}{a} \in [0, 1]$, the ratio between growth at t with optimal growth. In other words, for an increasing function f, let

$$-\frac{1}{y}\frac{dy}{dt} = f(y^*) = f(\frac{y}{a}) = f(1 - \frac{a-y}{a})$$
$$= f(1-d), d = \frac{a-y}{a} \in [0, 1]$$
$$= w_0 + w_1 d + w_2 d^2 + w_3 d^3 + \cdots \quad (6)$$

by Taylor's series expansion. Here $d = d(y)$ represents relative difference between optimal growth a with present growth. Note that $d \to 0 \Leftrightarrow y \to a$, with proliferation rate $\frac{1}{y}\frac{dy}{dt} \to 0$, as time $t \to \infty$; and therefore $w_0 = 0$.

Let all the coordinates in $\mathbf{w} = (w_1, w_2, w_3, \ldots)$ be nonzero, implying that the proliferation cannot be expressed in terms of a polynomial of finite degree.

Selecting a general harmonic sequence of monotonic constants $w_i = -c/i$, $i = 1, 2, 3, \ldots, c > 0$ in (6), one has Gompertz model $\frac{1}{y}\frac{dy}{dt} = c\log(\frac{a}{y})$.

The series convergences as $-\log(1-x) = \sum_{n=1}^{\infty} \frac{x^n}{n}, x \in [-1, 1)$. For $x = 1$, the series diverges and so does the proliferation rate of Gompertz curve for $y \to 0,\ d \to 1$.

Acknowledgements Thanks are due to Dr. Utpal Garain and Sharod Roy Choudhury for some interesting discussions and providing data on word frequencies.

References

Berger RD (1981) Comparison of Gompertz and logistic equations to describe plant disease progress. Phytopathology 71:716–719

Cuetos F, Glez-Nosti M, Barbon A, Brysbaert M (2011) SUBTLEX-ESP: Spanish word frequencies based on film subtitles. Psicologica 32:133–143

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

d'Onofrio A (2005) A general framework for modeling tumor-immune system competition and immunotherapy. Mathematical analysis and biomedical inferences. Phys D 208:220–235

Novile AG, Ricciardi LM, Sacerdote L (1982) On Gompertz growth model and related difference equations. Biol Cybern 42:221–229

O'Rourke SFC, McAneney H, Hillen T (2009) Linear quadratic and tumour control probability modelling in external beam radiotherapy. J Math Biol 58:799–817

Some Further Results on Nonuniform Rates of Convergence to Normality in Finite Population with Applications

Ratan Dasgupta

Abstract Rates of convergence in CLT are studied while sampling from a finite population under suitable moment assumptions on super population. We assume that all the moments for variate values exist in super population, having specific types of moment bound; but variate values are not necessarily bounded. Consequently probabilities of deviations, nonuniform L_p version of the Berry–Esseen theorem and moment type convergences are proved for standardised sample sum from finite population. In cross-sectional growth data, for each value of time t, the growth observations $y_i = y_i(t)$ may be considered as sample arising from a finite population. Average of observations falling in a small window of time may then be considered as an estimate of growth to be assigned at the average of time points in that interval. Convergence rates in CLT for sample mean in a finite population are compared with optimal rates in iid set-up, in order to assess performance of growth estimates. Growth data of a bulb crop onion is analysed. Derivative and proliferation rate of growth curve of the bulb crop are estimated to find appropriate time for harvesting the crop.

Keywords Nonuniform L_p version of the Berry–Esseen theorem • Probabilities of deviations • Smoothing spline • Bulb crop • Entire function

MS subject classification: Primary: 60F99, secondary: 62P10

1 Introduction and Some Preliminaries

We consider a finite population of N units. Suppose n units are selected by simple random sampling without replacement from this population. In the classical approach with non-random norming, we would like to study the limiting behaviour of the sum of variate values in selected sample as the sample size n increases.

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_11

The quantities n and N may increase independently, $n < N$. A standard assumption usually made is that in the limit $n/N \to \lambda \in (0,1)$. From the convergence results on CLT for independent random variables X_{ni} in a triangular array under the condition $\inf_{n>1} n^{-1} \sum_{i=1}^{n} V(X_{ni}) > 0$, e.g., see Dasgupta (1989); it is seen that the above assumption on n/N may be relaxed to obtain convergence rates in finite population. The essential requirement turns out to be $\inf_n n^{-1} V(\eta^*) > 0$, for computing the normal approximation zone of tail probability in terms of n, where η^* is the Hájek's projection of centred sample sum from a finite population; see also (2.18) of Dasgupta (1994).

The asymptotic normality of the standardised sample sum was proved by Erdös and Rényi (1959) and Hájek (1960). The uniform Berry–Esseen type bounds are obtained by Bikelis (1969) and Höglund (1978). Bloznelis and Götze (2000) studied Edgeworth expansion for finite-population U-statistics. Dasgupta (1994) studied the nonuniform rates of convergence to normality under the assumption of existence of moments of order ≥ 2 of variate values under a super population model and obtained moderate deviation and allied results on moment convergence, L_p version of Berry–Esseen theorem. Robinson (1977) obtained Chernoff type large deviation results, assumptions made therein were relaxed by Hu et al. (2007) in a set-up where the variables are self-normalised to obtain Cramer type large deviation results. The approximation zone in terms of $0 < x < (1/A) w_N \sigma / \max_k |a_k - \mu|$, computed therein is in a different direction and cannot be readily compared with best normal approximation zone $o(n^{1/6})$ for tail probability available for iid set-up in the general case under classical norming, see, e.g., Dasgupta (1989). Assumptions of Hu et al. (2007) are different from moment bounds in the super population model used in the present paper for comparison with standard moment assumptions usually made in the iid set-up.

In the present paper we develop large deviation results in traditional set-up extending the earlier results of Dasgupta (1994) on moderate deviation to higher order deviations. One of the goals is to show that it is possible to obtain best possible normal approximation zone as that for iid r.v's, in the case of finite population sampling as well, in a comparable set-up. We use the same notations of Dasgupta (1994).

A moment bound of Dasgupta (1993) for general stochastic processes that includes martingales as a special case is required to estimate remainder in Hájek's lemma. Slightly modified version of the result on moment bound is given below for completeness.

Theorem A. *Let* $\{X_i, i \geq 1\}$ *be a stochastic process with* $E[sgn(S_{i-1})X_i| \ |S_{i-1}|] \leq 0$, $E(\sum_{i=1}^{n} \pm X_i)^2 \leq n\beta_{2,n}^*$, *where* $S_i = \sum_{j=1}^{i} X_j$, $\gamma_{\nu,n} = E|X_n|^\nu$, $\beta_{\nu,n}^* = \max_{1 \leq j \leq n} \gamma_{\nu,j}$. *If the l.h.s. of* (1) *is finite, then for* $\nu \geq 2$

$$E|S_n|^\nu \leq c_\nu n^{\nu/2} \beta_{\nu,n}^*, \text{where } c_\nu = [2(\nu-1)\delta]^{\nu/2} \tag{1}$$

and for large n, $\delta \approx (1 + \frac{\nu}{2n})$.

An extra term 2 in c_ν above appears due to the fact that expectation of maximum of the terms in (2.2) of Dasgupta (1993) is bounded above by sum of the expectations in (2.3) therein, leading to an extra factor 2; i.e., the correct expression is $E\max(|S_n|^{\nu-2}X_n^2, |S_n^*|^{\nu-2}X_n^2) < E(|S_n|^{\nu-2}X_n^2 + |S_n^*|^{\nu-2}X_n^2)$. This modification does not affect the results in Dasgupta (1994), as L therein is a generic positive constant.

Let y represent a random variable taking the values $y_1, y_2, \cdots, y_N$; the attributes of the N units of the finite population A with equal probabilities $1/N$ and $\bar{Y} = N^{-1}\sum_{i=1}^N y_i$, be the population mean.

Although in a finite population of N elements all the characteristics are finite valued, it may not remain so with increase in N, the size of the population from which the sample is drawn. Consequently, some regularity conditions are assumed on the super population from which the present population A of N elements is considered to be a subset. In a super population model we shall assume that

$$\sup_{N\geq 1} Eg(y-\bar{Y}) = \sup_{N\geq 1} E_A g(y-\bar{Y}) = \sup_{N\geq 1} N^{-1}\sum_{i=1}^N g(y_i-\bar{Y}) < \infty \qquad (2)$$

where $g(x)$ is an even, nondecreasing function on $[0,\infty)$. The case of finite moments $g(x) = |x|^{2+c}u(x), c \geq 2$, and $u(x)$ having growth less than any power of $|x|$, was considered in Dasgupta (1994). Here we consider a higher spectrum of g that ensures existence of all the moments of $\tilde{y} = y - \bar{Y}$.

Let

$$\xi = \sum_{i\in a_n} y_i \qquad (3)$$

be the sum of attributes in a simple random sample a_n of size n. Hájek (1960) explored the fact that Poisson sampling may be interpreted as simple random sampling of size K, where K is a binomial $(N, n/N)$ variable, to split $\eta = (\xi - E\xi) = \eta^* + (\eta - \eta^*)$ into a main part $\eta^* = \sum_{i=1}^n \zeta_i$, $\zeta_i = (y_i - \bar{Y})I(a_K \ni i)$ consisting sum of iid random variables, plus a negligible remainder $(\eta - \eta^*)$.

A standardised version of $\xi = \sum_{i\in a_n} y_i$ may then be written in the form.

$$T_n = \sum_{i\in a_n}(y_i-\bar{Y})/\sqrt{\mathrm{var}(\eta^*)} = \eta/\sqrt{\mathrm{var}(\eta^*)} = [\mathrm{var}(\eta^*)]^{-1/2}\sum_{i=1}^n \zeta_i + R_n \qquad (4)$$

where $R_n = (\eta - \eta^*)/\sqrt{\mathrm{var}(\eta^*)}$ is the standardised remainder. We shall assume that

$$\inf_N \frac{1}{N}\sum_{i=1}^N (y_i-\bar{Y})^2 > 0, \ \lim_{N\to\infty}\frac{n}{N} = \lambda \in (0,1) \qquad (5)$$

R_n has the following moment bound.

$$E\{(\eta-\eta^*)^{2m}|K=k\}=E[\{\sum_{i=1}^{l}(y_i-\bar{Y})\}^{2m}|K=k] \tag{6}$$

where

$$l=|k-n| \tag{7}$$

With an application of Theorem A, the following moment bounds for R_n hold.

$$\begin{aligned} ER_n^{2m} &\le n^{-m/2}L^m e^{(3m/2)\log m}E(y-\bar{Y})^{2m} \text{ for } m=O(n) \\ &\le n^{-m/2}L^m e^{(5m/2)\log m}E(y-\bar{Y})^{2m} \text{ for unrestricted } m \end{aligned} \tag{8}$$

where

$$E(y-\bar{Y})^{2m}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})\}^{2m} \tag{9}$$

see (2.19)–(2.20) of Dasgupta (1994); the case of finite order moment, i.e., m is finite was considered therein to obtain rates of convergence and allied results. Here we consider finiteness of all the moments. Specifically we shall assume the following types of moment bounds for $\tilde{y}=y-\bar{Y}$.
Type 1:

$$\sup_{N\ge 1}E\tilde{y}^{2m}=\sup_{N\ge 1}E(y-\bar{Y})^{2m}\le Le^{w_0 m^{\nu}} \tag{10}$$

$\forall m>1$, and for some $L>1$, where $w_0>0$, $\nu>1$. The above condition is equivalent to

$$\sup_{N\ge 1}E\exp[s\{\log_e(1+|\tilde{y}|)\}^{\nu/(\nu-1)}]<\infty \tag{11}$$

where $s=w_0^{-1/(\nu-1)}$, see (4.15)–(4.16) of Dasgupta (2013) for similar assumptions. Assumption (10) is equivalent to finiteness of m.g.f. in a neighbourhood of zero for the transformed random variable $\{\log_e(1+|\tilde{y}|)\}^{\nu/(\nu-1)}$. The original variable $\tilde{y}$ has moment bounds of high magnitude. After logarithmic transformation the variables are tamed, and some power of transformed random variable $\tilde{y}$ possess m.g.f.

We shall also consider the following type of moment bounds.
Type 2:

$$\sup_{N\geq 1} E\tilde{y}^{2m} = \sup_{N\geq 1} E(y-\bar{Y})^{2m} \leq L^m e^{\nu m \log m} \tag{12}$$

$\forall m > 1$, where $L > 0,\ \nu > 1$. The above condition is implied by

$$\sup_{N\geq 1} E\exp(s|\tilde{y}|^{1/\nu}) < \infty \tag{13}$$

where $0 < s < s_0 = \nu e^{-1}L^{-1/\nu}$, see (4.34)–(4.35) of Dasgupta (2013) for similar assumptions. Assumption (12) is equivalent to finiteness of m.g.f. in a neighbourhood of zero for the transformed random variable $|\tilde{y}|^{1/\nu}$. The assumption is weaker than existence of m.g.f. as $\nu > 1$.

We shall further consider the following type of bound.
Type 3: Bound of *Type 2* with a different parametric zone, where $\nu \in (0, 1]$. That is,

$$\sup_{N\geq 1} E\exp(s|\tilde{y}|^{1/\nu}) < \infty,\ \nu \in (0, 1] \tag{14}$$

This ensures m.g.f. of $\tilde{y}$ exists, but $\tilde{y}$ may not be bounded over supremum of $N \geq 1$. Nonuniform Berry–Esseen bound in such cases were considered in Dasgupta (2006) for independent random variables in a triangular array.

Rates of convergence in CLT for finite population model are quite sharp and comparable with iid set-up, indicating that the former model may be used for analysing crop yield data.

In Sect. 2 we obtain nonuniform rates of convergence from general results of independent random variables see, e.g., Dasgupta (1989, 1994, 2006), by treating the remainder in Hájek's projection for sample sum from finite population to be negligible. Optimal normal approximation zone for probabilities of deviations, as in the iid case is shown to be attainable. Nonuniform L_p version of the Berry–Esseen theorem and moment type convergences are proved for standardised sample sum from finite population under different moment bounds in Sect. 2. The results obtained are compared with those for iid random variables to ascertain precision of finite population techniques applied in growth curve estimation.

From Indian Statistical Institute Giridih farm, yield data are collected on a bulb crop onion that takes about 3–4 months of lifetime from sprouting stage to mature.

We estimate crop growth curve in Sect. 3 from clustered yield observations in small time window and the average of observations in a window is treated like sample sum arising from finite population, for which the theory applies. We estimate derivative and proliferation rate of the growth curve. These have applications to decide harvest time. The appropriate time to harvest the crop turns out to be 90 days from plantation in that region.

2 Nonuniform CLT Bounds in Finite Population

First we consider the *Type 1* moment bound given in (10). The next theorem states the normal approximation zone for tail probability of the standardised sample sum T_n.

Theorem 1. *Under the assumptions* (5) *and* (10)*, for the standardised sample sum* T_n *from a finite population defined in* (4)*, one has*
$1 - P(T_n \le t_n) \sim \Phi(-t_n) \sim P(T_n \le -t_n)$ *for*
$t_n^2 \le \alpha(\log n)^{\nu/(\nu-1)} + M$, $M > 0$, $t_n \to \infty$, *where* $\alpha = (2-\epsilon)w_o^{-1/(\nu-1)}$ $(\nu-1)\nu^{-\nu/(\nu-1)}$, $s = w_o^{-1/(\nu-1)}$, $\epsilon > 0$ *is arbirtary small,* w_o *and* ν *are defined in* (10) *and* $M > 0$ *may be arbitrary large.*

Proof. This is similar to the proof of Theorem 1, given in Dasgupta (2013). As noted therein the order of normal approximation zone is optimal.

Remark 1. The leading term for t_n^2 in Theorem 1 is of order $(\log n)^{\nu/(\nu-1)}$. The normal approximation zone for standardised sample sum of iid random variables is also of same order, see also Theorem 2.3 of Dasgupta (1989). These zones are larger than moderate deviation zone, as $\nu > 1$. Moderate deviation results hold when some finite moment (>2) exists. In (10)/(11) we assumed existence of *all* the moments of $\tilde{y} = y - \bar{Y}$, and hence the resulting zone gets extended beyond moderate deviation.

Denote $G_n(t) = P(T_n \le t)$. The next theorem provides an overall nonuniform bound in the CLT for standardised sample sum from finite population. The theorem and remark below follow along the lines of Theorem 2 and Remark 2 of Dasgupta (2013).

Theorem 2. *Under the assumptions of Theorem 1, there exists a constant* $b > 0$*, depending on* $w > w_0$ *and* ν *such that the following holds.*

$$|G_n(t) - \Phi(t)| \le b\, n^{-\frac{1}{2}+\epsilon_n} e^{-c(\log(1+|t|))^{\nu/(\nu-1)}},\ \nu > 1,\ -\infty < t < \infty \qquad (15)$$

where $c = w(\nu-1)\{2/(w\nu)\}^{\nu/(\nu-1)} > 0$, *and* $\epsilon_n = (2c)^{-(\nu-1)/\nu}(\log n)^{-1/\nu} = O((\log n)^{-1/\nu}) \to 0$, *as* $n \to \infty$.

Remark 2. Observe that in the nonuniform bound (15), the part depending on $|t|$ decreases at a faster rate than any polynomial power of $|t|$. Uniform bound of the rate approaches to the optimal bound $O(n^{-1/2})$, as the excess $\epsilon_n = O((\log n)^{-1/\nu}) \to 0$, $n \to \infty$.

As a consequence of Theorem 2, the following two theorems on nonuniform L_p version of Berry–Esseen theorem and moment type convergence are immediate, see Dasgupta (2013).

Theorem 3. *Under the assumptions of Theorem 2,*

$$||e^{c(\log(1+|t|))^{\nu/(\nu-1)}}(1+|t|)^{-q/p}(G_n(t)-\Phi(t))||_p = O(n^{-\frac{1}{2}+\epsilon_n})$$

for $p \geq 1$ *and any* $q > 1$.

Theorem 4. *Under the assumptions of Theorem 2 and for a non-negative even function g with*

$$\frac{d}{dx}[x^2 g(x)] = O((1+x)^{-q} e^{c(\log(1+|x|))^{\nu/(\nu-1)}}),\ \forall x > 0, \text{and } q > 1$$

the following holds for standardised sample sum T_n *from finite population and a* $N(0,1)$ *variable T.*

$$|E(T_n^2 g(T_n)) - E(T^2 g(T))| = O(n^{-\frac{1}{2}+\epsilon_n})$$

Next we consider the *Type 2* moment bound given in (12). The next theorem states the normal approximation zone for tail probability of the standardised sample sum T_n.

Theorem 5. *Under the assumptions* (5) *and* (12)*, for the standardised sample sum* T_n *from a finite population defined in* (4)*, one has*

$$1 - P(T_n \leq t_n) \sim \Phi(-t_n) \sim P(T_n \leq -t_n),\ t_n \to \infty, \text{for } t_n = o(n^{1/(2\nu+5)}).$$

Proof. This is similar to the proof of Theorem 5, given in Dasgupta (2013) with a value of $\nu = 3/2$, where one may use the moment bound (8) for $m = O(n)$. Note that the order of q in (4.39)–(4.40) of Dasgupta (2013) is $q = O(n^{(1-2\gamma)/(\nu+\nu)})$. In the present context $\nu = 3/2, \nu > 1$. So $m = O(n^{2(1-2\gamma)/5})$ satisfies $m = O(n)$ of (8), as m here takes the role of q in Dasgupta (2013).

Remark 3. In the above theorem we considered $\nu > 1$. The proof goes through for a wider zone $1/2 \leq \nu < \infty$. For $\nu = 1/2$ the normal approximation zone is $t = o(n^{1/(2\nu+5)}) = o(n^{1/6})$. This zone, in general, is the best possible zone even for iid random variables, see Theorem 2.3 of Dasgupta (1989). The condition $\nu = 1/2$ implies that the characteristic function of $\tilde{y} = y - \bar{Y}$ is an entire function of order ≤ 2, possibly having zeroes. For rates of convergence in CLT in independent set-up with this condition, see Dasgupta (1992).

The next theorem provides an overall nonuniform bound for $|G_n(t) - \Phi(t)|$.

Theorem 6. *Under the assumptions of Theorem 5, there exists a constant* $b > 0$*, depending on* β, ν *and* δ *such that the following holds.*

$$|G_n(t) - \Phi(t)| \leq b\, n^{-\frac{1}{2}} (\log n)^{\delta} e^{-\beta |t|^{(1/\nu)\wedge(4/(2\nu+5))}},\ -\infty < t < \infty$$

where $\beta > 0$*, may be arbitrary large and* $\delta > (2\nu+5)/4$*, may be arbitrary near to* $(2\nu+5)/4$.

Proof. Proof of the above follows the lines similar to that for Theorem 6 of Dasgupta (2013), with $v = 5/2$, where one uses the moment bound (8) for unrestricted m; see (4.44) of Dasgupta (2013), where t is unbounded in $a_n(t)$ requiring the unrestricted bound of ER^{2m}.

As a consequence of Theorem 6, following two theorems are immediate; see Dasgupta (2013).

Theorem 7. *Under the assumptions of Theorem 5, for any* $p > 1$

$$||e^{\beta|t|^{(1/v)\wedge(4/(2v+5))}}(G_n(t) - \Phi(t))||_p = O(n^{-\frac{1}{2}}(\log n)^{\delta})$$

where $\beta(> 0)$ *is a fixed constant that may be made arbitrary large; and* $\delta > (2v + 5)/4$, *may be taken arbitrary near to* $(2v + 5)/4$.

Theorem 8. *Under the assumptions of Theorem 5, and for a non-negative even function g with*

$$\frac{d}{dx}[x^2 g(x)] = O(e^{\beta|x|^{(1/v)\wedge(4/(2v+5))}}),\ \forall x > 0$$

where $\beta(> 0)$ *may be arbitrary large, the following holds for the standardised sample sum* T_n *from a finite population defined in* (4), *and an* $N(0, 1)$ *random variable* T,

$$|E(T_n^2 g(T_n)) - E(T^2 g(T))| = O(n^{-\frac{1}{2}}(\log n)^{\delta}).$$

Next consider the moment bound of *Type 3*, i.e., take $v \in (0, 1]$ in (13). This assumption ensures the existence of m.g.f for $\tilde{y} = y - \bar{Y}$ in the super population. However, $\tilde{y}$ may not be bounded as $n \to \infty$. We then have

$$\sup_{N\geq 1} E\tilde{y}^{2m} = \sup_{N\geq 1} E(y - \bar{Y})^{2m} \leq L^m e^{vm\log m},\ \forall m > 1, L > 0,\ v \in (0, 1] \quad (16)$$

The above condition is implied by

$$\sup_{N\geq 1} E\exp(s|\tilde{y}|^{1/v}) < \infty \quad (17)$$

where $0 < s < s_0 = ve^{-1}L^{-1/v}$, $v \in (0, 1]$.

Observe that from (8), the remainder has the bound

$$\begin{aligned} ER_n^{2m} &\leq n^{-m/2}L^m e^{(5m/2)\log m}E(y - \bar{Y})^{2m} \text{ for unrestricted } m \\ &\leq n^{-m/2}L^m e^{(v+\frac{5}{2})m\log m} \text{ under (16), for unrestricted } m \end{aligned} \quad (18)$$

Then proceeding like Theorem 4.1 in Dasgupta (2006), one may obtain

Theorem 9. *Under the assumption* (5) *and* (16)/(17)*, there exist constants* $b(>0)$*, and* $k \in (0, 1/2)$ *such that the following holds for the distribution function* $G_n(t) = P(T_n \leq t)$*, where* T_n *is standardised sample sum* (4) *from a finite population.*

$$| G_n(t) - \Phi(t) | \leq b\, n^{-1/2} (\log\ n)^{\nu+\frac{5}{2}} \exp\,(-k \mid t \mid^{2\wedge 1/(\nu+\frac{5}{2})})$$

Consequently, following two theorems are immediate.

Theorem 10. *Under the assumptions of Theorem 9, for any* $p > 1$

$$||e^{k|t|^{2\wedge 1/(\nu+\frac{5}{2})}}(G_n(t) - \Phi(t))||_p = O(n^{-1/2}(\log\ n)^{\nu+\frac{5}{2}})$$

where $k \in (0, 1/2)$ *may be taken arbitrary close to* $1/2$.

Theorem 11. *Under the assumptions of Theorem 5, and for a non-negative even function g with*

$$\frac{d}{dx}[x^2 g(x)] = O(e^{k|x|^{2\wedge 1/(\nu+\frac{5}{2})}}),\ \forall x > 0$$

and $k \in (0, 1/2)$ *may be taken arbitrary close to* $1/2$*; the following holds for the standardised sample sum* T_n *defined in* (4)*.*

$$|E(T_n^2 g(T_n)) - E(T^2 g(T))| = O(n^{-\frac{1}{2}}(\log n)^{\delta})$$

where T *is a* $N(0, 1)$ *random variable.*

3 An Application

Onion is a bulb crop that takes about 3–4 months time from sprouting stage to mature for harvest. The following yield versus lifetime data of onion is obtained from Indian Statistical Institute Giridih farm experiments. In total 100 sprouting seeds were planted underground on 3 February 2014 in plots of barren land having sandy soil composition mixed with "dhoincha" (*Sesbania bispinosa*) plant compost manure. Out of 100 seedlings planted, 7 did not germinate.

The followings are the grouped frequency distribution of onion plant lifetime x in day (Table 1).

Table 1 Frequency distribution of onion plant lifetime

x	75	82	83	87	88	89	90	91	92	93	94	95	96	97	98	99	100	106	117
f_x	1	1	1	4	14	5	9	11	3	11	3	4	3	9	3	2	7	1	1

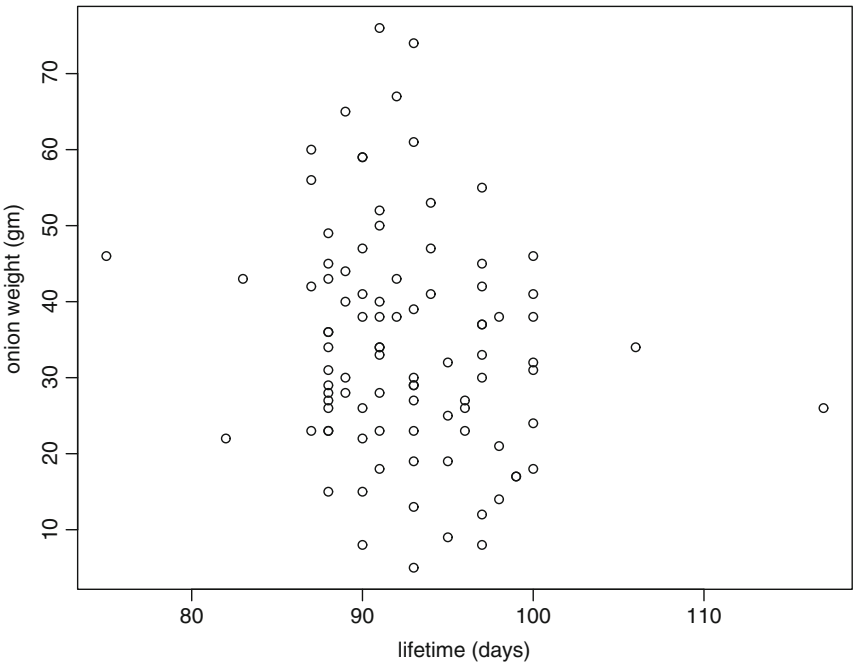


Fig. 1 Scatter diagram of bulb crop lifetime and yield. Scatter diagram of 93 onion lifetime versus yield data shows wide fluctuation

Scatter diagram of lifetime versus yield with 93 observations is shown in Fig. 1. Results proved in previous section indicate that the data analysis techniques under finite population model can be considered. Wide fluctuations among the yield data points over narrow time zones are observed, indicating moment bounds of large magnitude for yield as considered in earlier section. One more data point is added at lifetime zero for mean of 93 initial weights of seed-onion used in planting, thus making 94 points in the scatter diagram for estimating the growth curve. Next, the lines joining the group means, shown in green colour summarise the fluctuations of points at fixed lifetime. Blue line corresponds to smoothed spline curve with shape parameter 1 drawn in R software. The lowess curve with $f = 2/3$ in red is seen to be almost overlapping with the spline curve in blue (Fig. 2). These curves still show some fluctuations that require being straightened out further, to have a smooth estimate of growth curve. To this end, the cross-sectional data was grouped with at least three observations in a cluster. The group mean of yield was assigned to the weighted average of the corresponding time points. These simple averages of observations inherit the property of sample mean from a finite population; the population is of Jharkhand cultivar of onion to be modeled

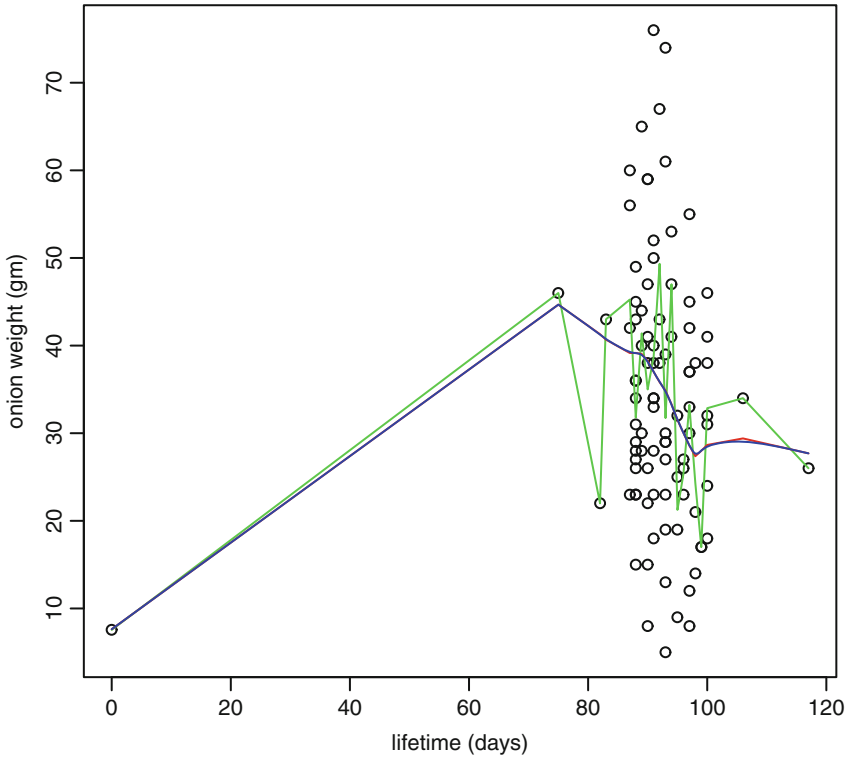


Fig. 2 Group means, lowess and spline curve of a bulb crop yield

from a super population of global hypothetical onion production in a season with this soil structure. According to the results proved in previous section, estimated growth curve under super population model has optimal properties comparable to that computed from iid observations in traditional set-up. To obtain a smooth response curve, lowess regression was used to the above-mentioned group mean of yield in clusters. A three-point moving average of the lowess points so obtained, along with a smoothing spline with shape parameter 0.48 fitted over the points is shown in Fig. 3 as an estimate of smooth growth curve for the bulb crop onion. The growth curve is sharply increasing in the beginning. Transformed leaf structures in onion plants constitute the food storage part at the bottom of the crop at mature stage. Harsh summer of April in Giridih, Jharkhand made the storage suffer due to evaporation of water content, thus causing a fall of the growth curve of onion towards end beyond 20 April 2014; farmers in Jharkhand Giridih farm should have harvested the bulb crop by that time. The plants that could withstand hard summer beyond 20 April 2014, have contributed to the rise of curve again towards end.

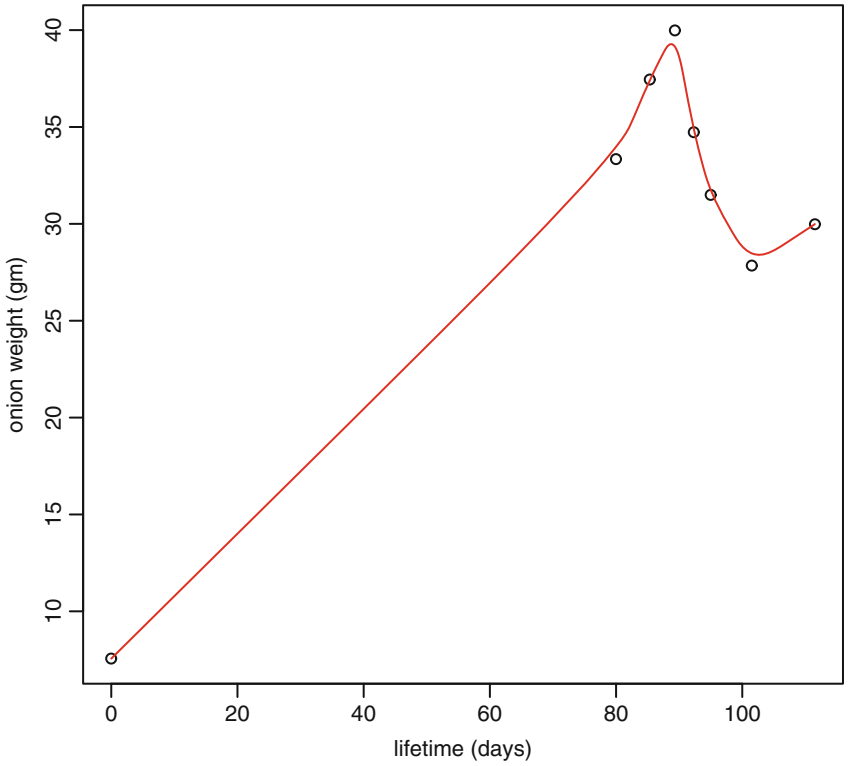


Fig. 3 Smoothed spline growth curve of a bulb crop

In Fig. 4 we compute derivative of the onion growth curve $y = y(t)$ using a technique of Dasgupta (2013). The derivative remains positive slightly beyond 80 days of lifetime, after which the curve falls down below zero, and then again moves upward about 100 days onward.

Proliferation rate $d \log y(t)/dt$ is computed by a similar technique and shown in Fig. 5, this exhibits a downward tendency till about 100 days before rising up again. Proliferation rate is independent of unit of weight measurements. The peak of onion growth curve is seen around 90 days, which seems appropriate time for harvesting the crop.

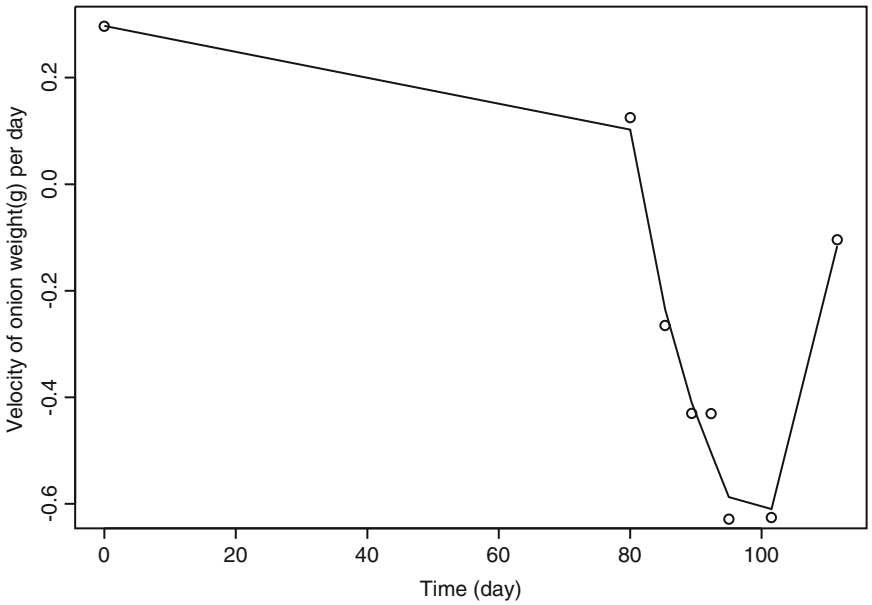


Fig. 4 Velocity of bulb crop yield: trimmed mean, wt. exp(-.01 x); spline

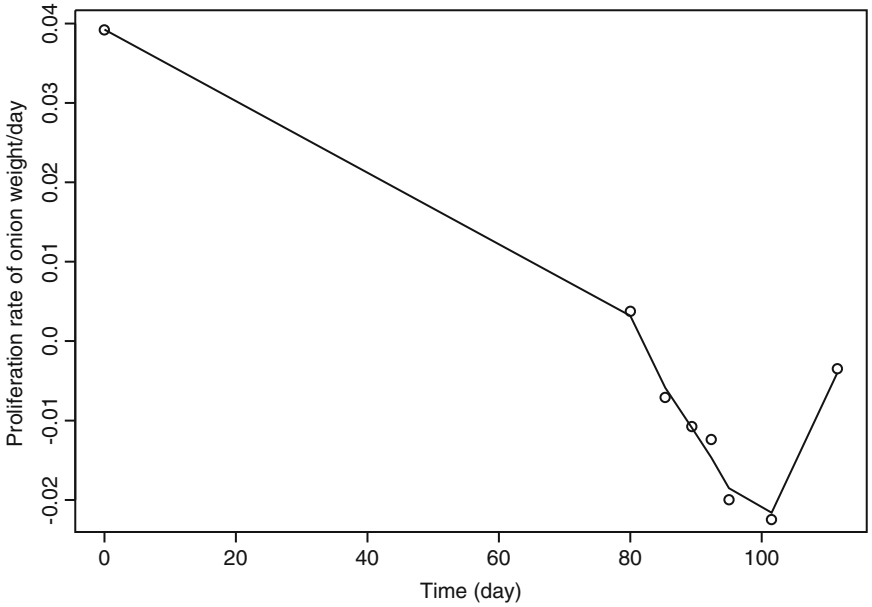


Fig. 5 Proliferation rate of bulb crop yield: trimmed mean, wt. exp(-.01 x); spline

References

Bikelis A (1969) On the estimation of the remainder term in the central limit theorem for samples from finite populations (Russian). Stud Sci Math Hung 4:345–354

Bloznelis M, Götze F (2000) An edgeworth expansion for finite-population U-statistics. Bernoulli 6:729–760

Dasgupta R (1989) Some further results on nonuniform rates of convergence to normality. Sankhyā A 51:144–167

Dasgupta R (1992) Rates of convergence to normality for some variables with entire characteristic function. Sankhyā A 54:198–214

Dasgupta R (1993) Moment bounds for some stochastic processes. Sankhyā A 55:180–152

Dasgupta R (1994) Nonuniform speed of convergence to normality while sampling from finite population. Sankhyā A 38:227–237

Dasgupta R (2006) Nonuniform rates of convergence to normality. Sankhyā 68:620–635

Dasgupta R (2013) Non uniform rates of convergence to normality for two sample U-statistics in non IID case with applications, Chap 4. In: Advances in growth curve models. Topics from the Indian Statistical Institute. Springer Proceedings in Mathematics & Statistics, vol 46. Springer, New York, pp 61–88

Erdös P, Rényi A (1959) On the central limit theorem for samples from a finiite population. Publ Math Inst Hung Acad Sci 4:49–61

Hájek J (1960) Limiting distributions in simple random sampling from a finite population. Publ Math Inst Hung Acad Sci 5:361–374

Höglund T (1978) Sampling from a finite population. A remainder term estimate. Scand J Stat 5:69–71

Hu Z, Robinson J, Wang Q (2007) Cramer type large deviations for samples from a finite population. Ann Stat 35:673–696

Robinson J (1977) Large deviation probabilities for samples from a finite population. Ann Probab 5:913–925

Unbounded Growth Model for Word Frequencies in Political Transition

Ratan Dasgupta

Abstract Frequencies of some words appearing in a vernacular newspaper are studied. Usages of certain politically flavoured words are without any bound during poll time. Cumulative frequencies of some such words appearing in a vernacular daily from West Bengal are modeled by an unbounded growth curve $y(t) = e^{b\exp(ct)}, b > 0, c > 0; t \in (0, \infty)$, resembling the structure of a Gompertz model. The present study is a relook at the same data set considered in Dasgupta (Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York, 2015) in a different approach. Such studies have relevance in prediction of poll results. Estimates of the model parameters are obtained from observed data over the period 2001–2010, covering several elections in India. Unbounded growth models in continuous time and discrete time are discussed in terms of interrelated proliferation rates.

Keywords Gompertz curve • Vernacular daily • Unbounded growth • Proliferation rate

MS subject classification: Primary: 62G05, secondary: 62P25.

1 Introduction

An upper bound in growth models arise mainly due to constraints on available resources, e.g., see Dempster and McLean (1998) on population growth. Growth of tumour spheroids eventually stops in models proposed by Wallace and Xinyue (2013).

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_12

In contrast, time dependent variables may sometimes have sharp growth without an upper bound. As for example, words with political flavour are frequently spoken in election rallies and reported in newspapers, having an impact on public opinion, affecting poll results in a multiparty democracy. Cumulative frequency of these words over time may increase without any bound. Gompertz curve has an upper asymptote, and may not be an appropriate model in such situations.

Gompertz model is generally appropriate for bacterial growth, tumour immune system, etc., where initially there is no competition for resources, followed by decrease in growth rate due to competition for the nutrients as cellular population expands. Tumours are cellular populations growing in a confined space, where availability of nutrients becomes limited after a time lapse from start.

The scenario is different with rampant use of politically flavoured words with no restriction on upper bound during election time. Dasgupta (2015) proposed a Gompertz model over a limited time zone for some words along with inflections appearing in a vernacular newspaper during the period 2001–2010, covering general elections in India. Although the fit is good, the problem of long range forecasting persists in a Gompertz analysis over limited time.

In this paper we address the problem by modifying the parameters in Gompertz model so as to accommodate unbounded growth. Specifically, we fit a Gompertz like growth curve

$$y = y(t) = e^{b\exp(ct)}, b > 0, c > 0; t \in (0, \infty) \tag{1}$$

to the available data. The form retains the structure of Gompertz model although $y(t) \to \infty$, as $t \to \infty$.

In Dasgupta (2015), cumulative relative frequency data on a time span of 10 years for 12 words appearing in a vernacular daily is modeled. Relative cumulative frequencies of each word over time are computed with respect to the total number ($n = 142985088$) of *all* words appearing in that daily newspaper during the years 2001–2010. In the present case we analyse the cumulative frequencies of the words, i.e., a scaled version of earlier data by unbounded growth model (1).

Unbounded models are useful in other studies as well. In a closed quantum system of many interacting particles Bardarson et al. (2012) proposed unbounded growth model in the propagation of entanglement, where entropy develops approximately logarithmically over a diverging time scale.

Proliferation rate $d \log y/dt$ for model (1) is $c \log y,\ c > 0$, this is an increasing function of growth, in contrast to the traditional Gompertz model with decreasing proliferation rate. Fitting the model is done by plotting the growth observations in log log scale over time and ascertaining approximate linearity in plotted data. Estimate of the model parameters is obtained from the slope and intercept of the fitted least squared regression line. Proliferation rates that have relevance in poll prediction are computed from estimated parameters.

In Sect. 2 we analyse the word frequency data considered in Dasgupta (2015) by the unbounded growth model (1). High value of coefficient of determination indicates that the proposed model fits the data well. Sharp growth patterns of

some words are reflected in large values of estimated parameters. In Sect. 3 we discuss continuous and discrete versions of unbounded growth models in terms of proliferation rates.

2 Analysis of Data with Unbounded Growth Model

Data of cumulative frequencies on 12 words including inflections versus time are presented in log log scale in Fig. 1. Approximate linearity with increasing trend is observed for all the words, especially for higher values of time. This indicates that the unbounded growth model (1) may be appropriate in the present case. In Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 we fit least squared regression lines to the data points for each word. Model parameters b and c of (1) are estimated from the exponential of the intercept and slope, respectively, of the fitted line. Accuracy of the model fit is determined by the value of r^2, the coefficient of determination, vide Table 1. The value of r^2 may improve considerably, if some initial data points are purged for some of the words, see Fig. 1.

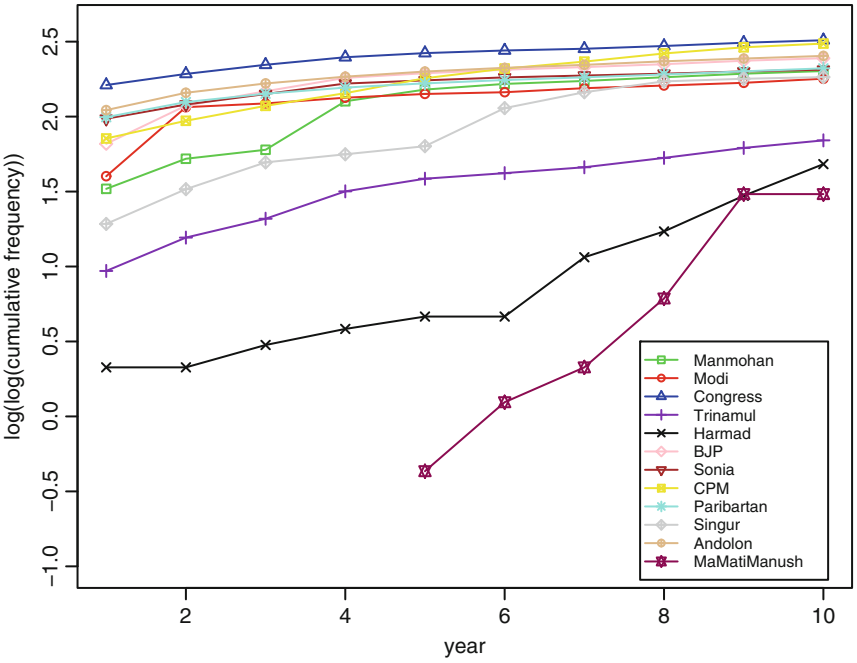


Fig. 1 Unbounded growth model for 12 words appearing in a vernacular daily

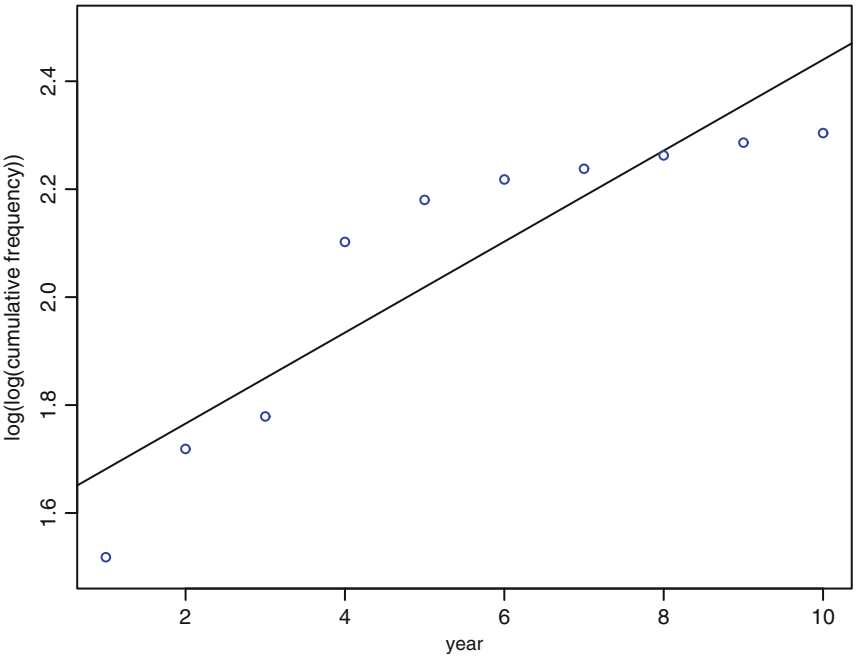


Fig. 2 Unbounded growth fit for cumulative frequency of the word "Manmohan"

In Fig. 2 cumulative frequencies of the word *Manmohan* over 10 years time show increasing trend, fit of the least square regression line would have been better than the present $r^2 = 0.8212$, if three initial points were purged; an increasing trend with remaining data points seems prominent.

Figure 3 plots the data points for the word *Modi* after purging the first data point, the value of r^2 is improved from 0.5844 to 0.9885. The values of the parameter b and c in Table 1 are also changed 6.340578 (7.580256); 0.04722 (.0228655) after deleting the first data point; the least square fit in Fig. 3 is excellent.

Least squared fit shown in Fig. 4 for the word *Congress* is good with a value of $r^2 = 0.9013$. Increasing trend in the data points are prominent.

The same can be said about Fig. 5 for the word *Trinamul.* This is name of a political party like *Congress.* The value of $r^2 = 0.9149$ is slightly higher than that for the latter.

If we ignore the apparent stability of the word *Harmad* during the time period 5–6 years, then an increasing trend in last four data points is prominent for the word as seen in Fig. 6. The value of r^2 is high, $r^2 = 0.9347$.

The coefficient of determination for the word *BJP* in the model fitting is $r^2 = 0.7635$. However, if the initial three data points are purged, the value is much higher $r^2 = 0.993$, see Table 1 and Fig. 7.

Table 1 Estimated parameters of unbounded growth curve $y(t) = e^{bexp(ct)}$, b>0, c>0, t>0; and model accuracy (coefficient of determination) r^2

Parameter	Manmohan	Modi (deleting 1 initial pt.)	Congress	Trinamul	Harmad	BJP (deleting 3 initial pts.)	Sonia	CPM	Paribartan	Singur	Andolon	MaMatiManush
b	4.93878821	6.340578 (7.580256)	9.368622	2.815861	1.001922	4.938166 (8.842974)	7.638109	6.366112	7.6335271	3.654791	8.061503	0.096708885
c	0.08428	0.04722 (.0228655)	0.030052	0.088332	0.15415	0.050772 (.0213180)	0.03242	0.070136	0.03174	0.110099	0.035472	0.39613
r^2	0.8212	0.5844 (.9885)	0.9013	0.9149	0.9347	0.7635 (.993)	0.8406	0.9616	0.9034	0.9406	0.8947	0.9642

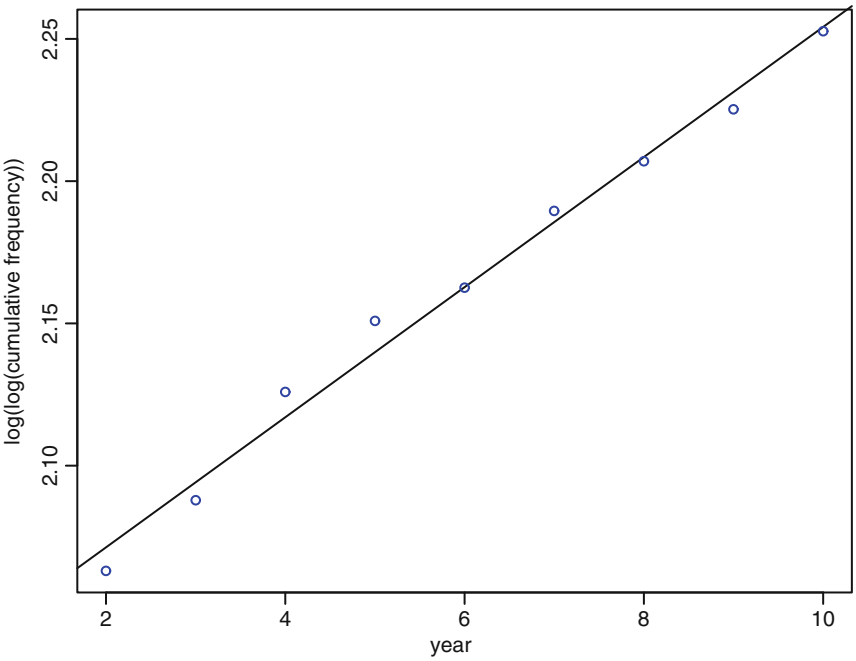


Fig. 3 Modified unbounded growth fit for cumulative frequency of the word "Modi"

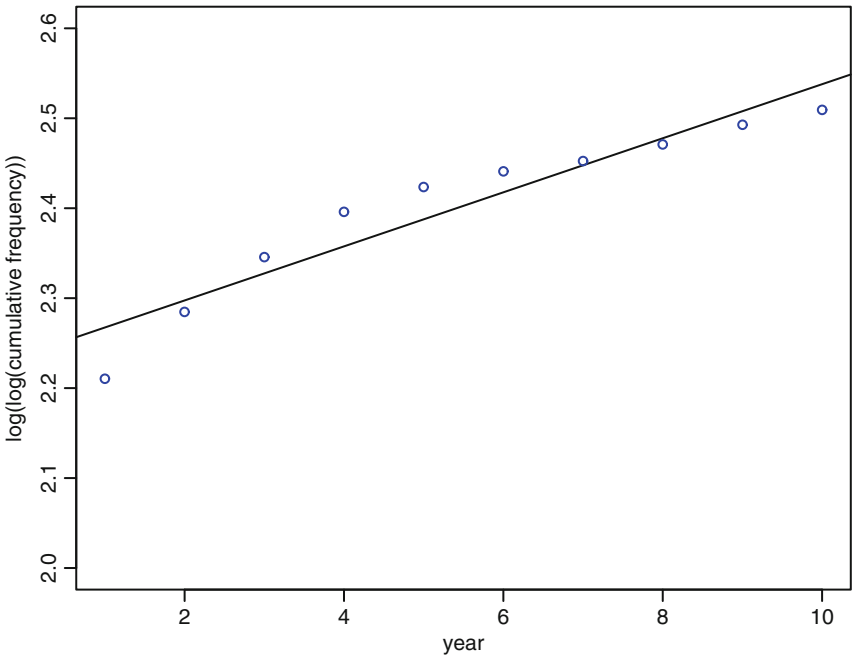


Fig. 4 Unbounded growth fit for cumulative frequency of the word "Congress"

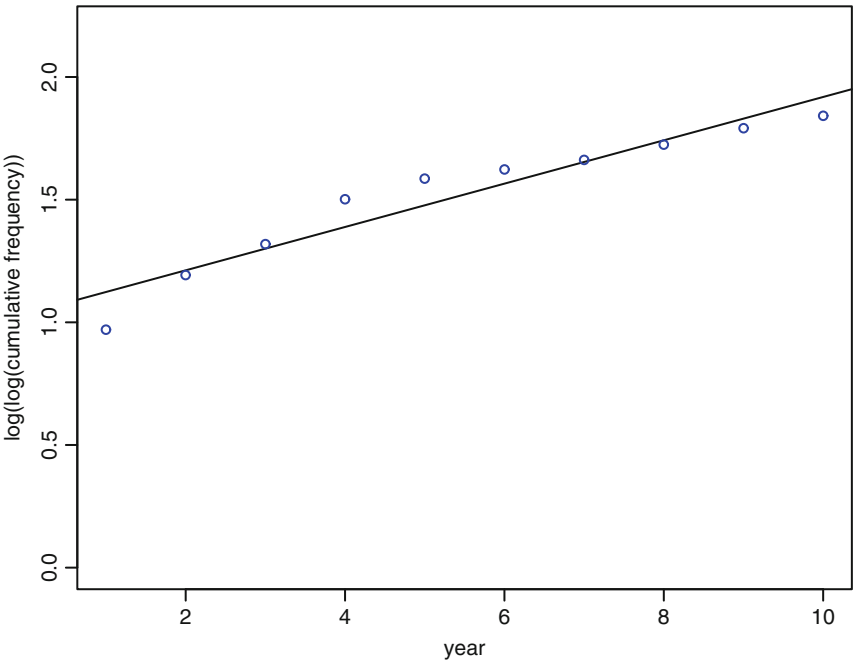


Fig. 5 Unbounded growth fit for cumulative frequency of the word "Trinamul"

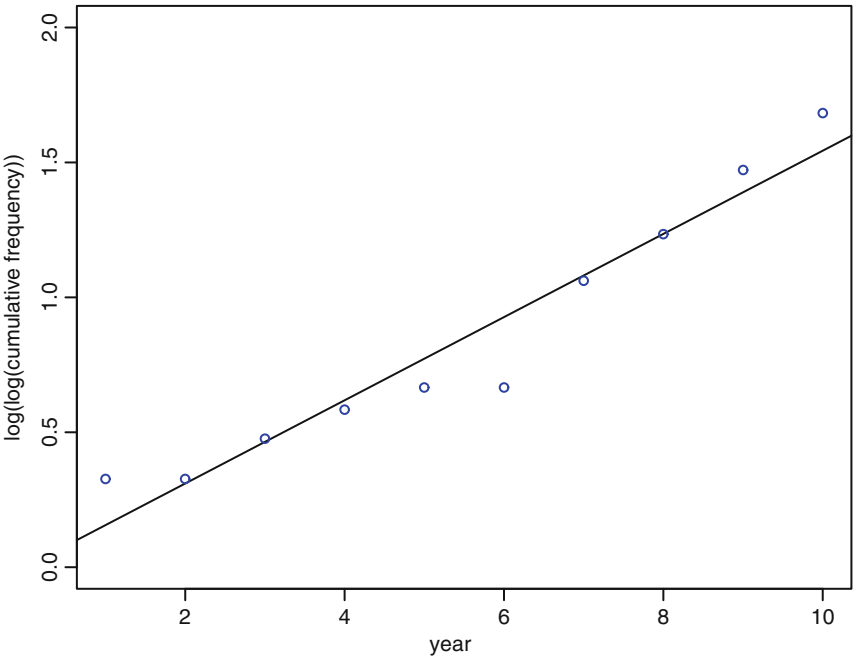


Fig. 6 Unbounded growth fit for cumulative frequency of the word "Harmad"

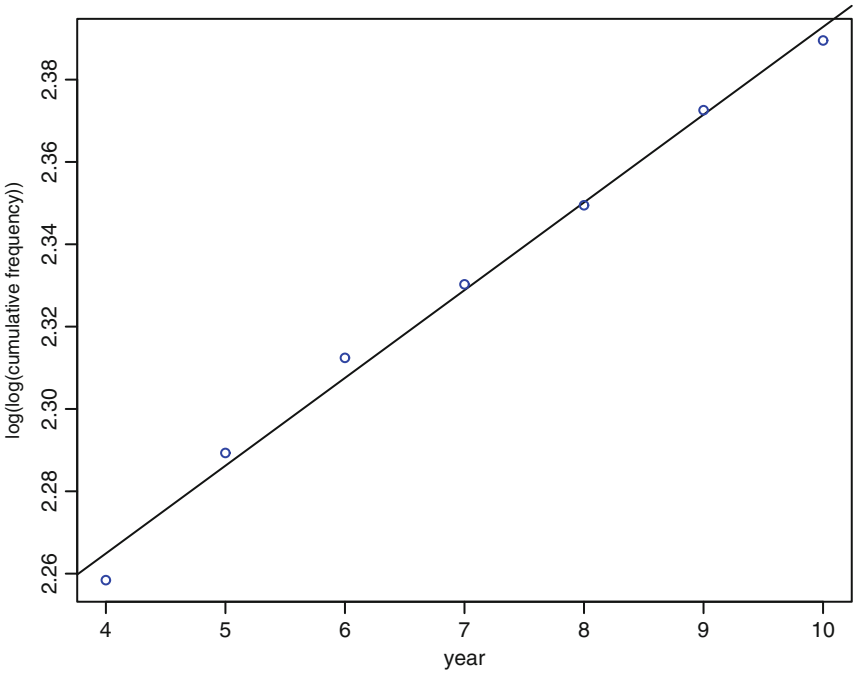


Fig. 7 Modified unbounded growth fit for cumulative frequency of the word "BJP"

Model fit for the word *Sonia* is satisfactory with $r^2 = 0.8406$, see Fig. 8. It appears that a better fit would have been possible, if the first data point was purged in modeling.

Model fit shown in Fig. 9 for the word *CPM* is excellent with a value of $r^2 = 0.9616$, and estimated value of slope $c = 0.070136$ for regression line in log log scale for model (1). Amongst the names of all political parties viz., *Congress, Trinamul, BJP, CPM* considered in this analysis, the word *Trinamul* corresponds to the highest value of the slope of increase with $c = 0.088332$, indicating a rapid growth of the word in the considered time segment.

The word *paribartan* has cumulative frequency that conforms to the model with a high value of $r^2 = 0.9034$. Regression line of Fig. 10 is nearly touching all the data points.

Singur is name of the place from where an agitation started, out of dispute arising from land acquisition for industry. Growth of this word is modeled in Fig. 11. Estimated slope $c = 0.110099$ for *Singur* is high, like that of the word *Harmad* with $c = 0.15415$. The value of r^2 is 0.9406 for the word *Singur*.

Growth of the word *Andolon* is depicted in Fig. 12. The value of r^2 is 0.8947 for the word *Andolan*. The pattern is similar to that of the word *paribartan* as shown in Fig. 10. Values of the estimated slope of these two words are moderate compared to other words.

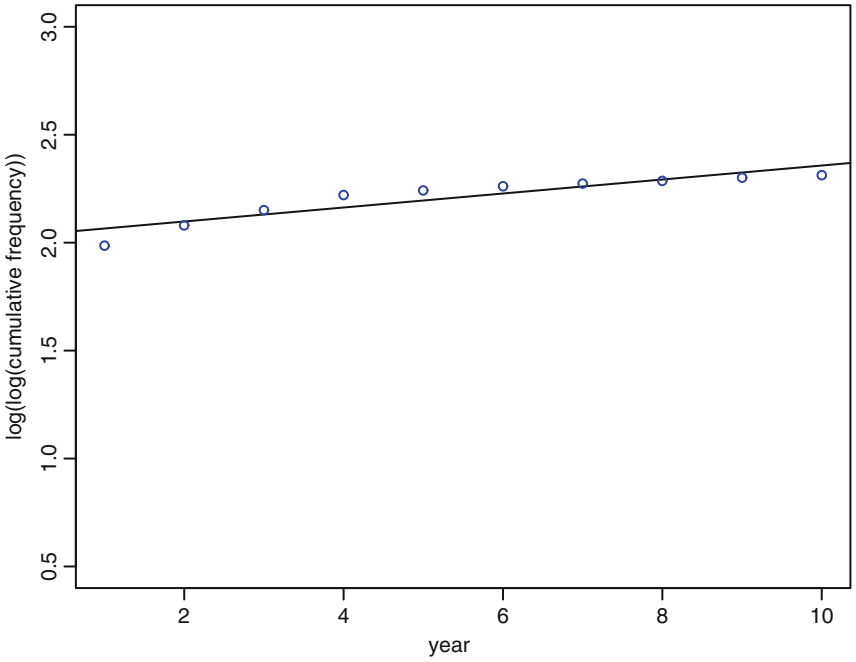


Fig. 8 Unbounded growth fit for cumulative frequency of the word "Sonia"

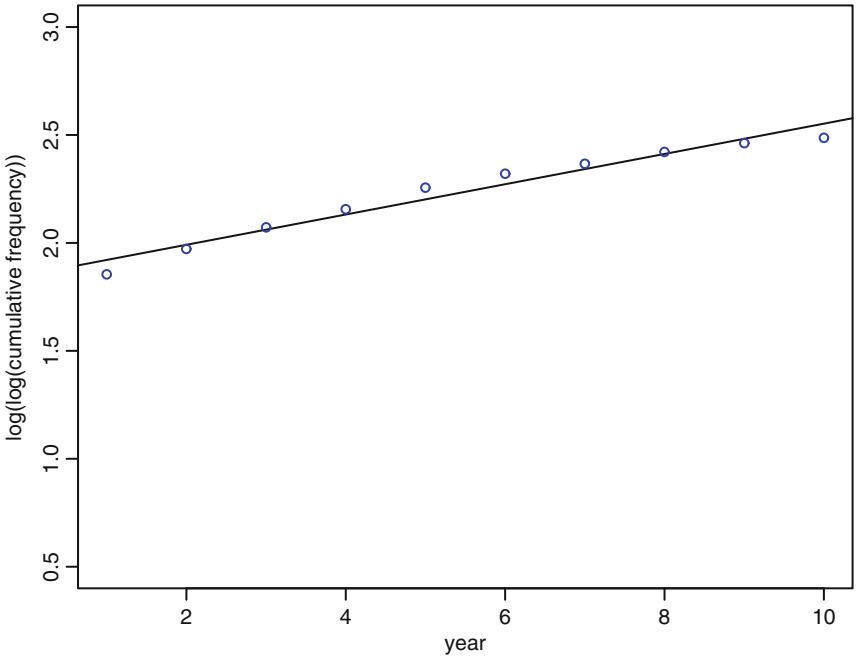


Fig. 9 Unbounded growth fit for cumulative frequency of the word "CPM"

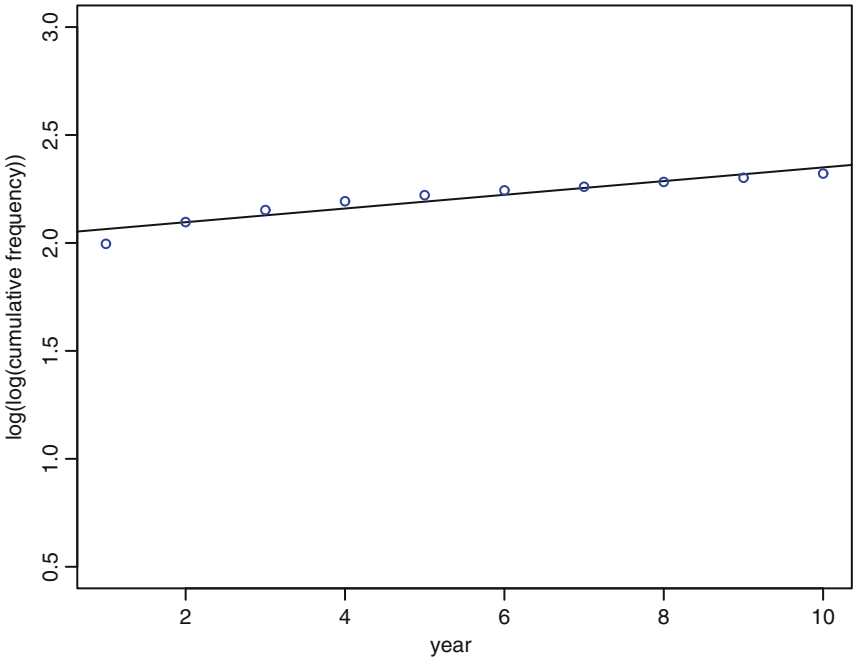


Fig. 10 Unbounded growth fit for cumulative frequency of the word "Paribartan"

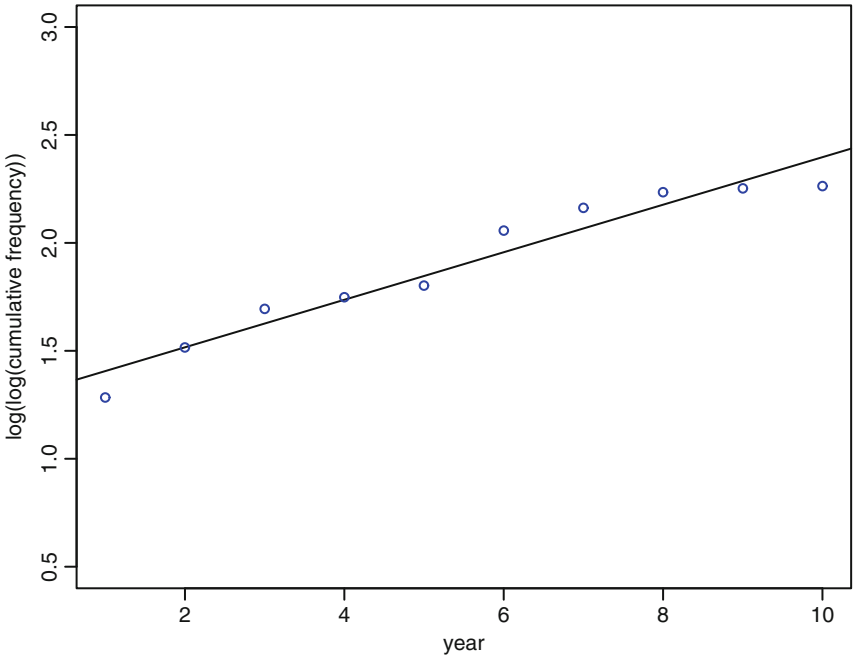


Fig. 11 Unbounded growth fit for cumulative frequency of the word "Singur"

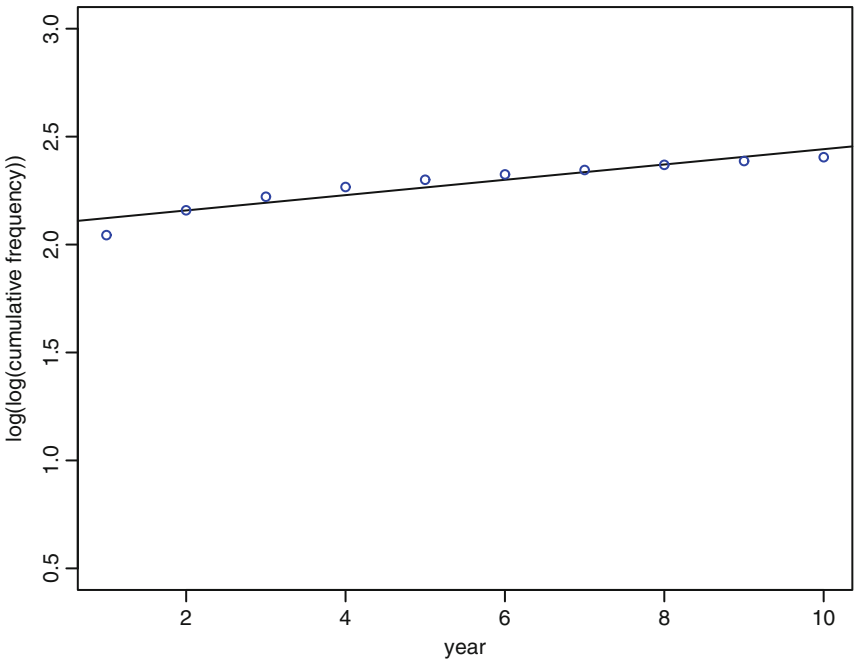


Fig. 12 Unbounded growth fit for cumulative frequency of the word "Andolon"

In Fig. 13 the growth of the word *MaMatiManush* is seen to be steepest among all the 12 words. This is reflected in the highest value of the estimated slope viz., $c = 0.39613$. The value of r^2 is 0.9642 for the word *MaMatiManush*.

The values of r^2 being high, the proliferation rates may be computed from the estimated parameters b and c of the model (1).

3 Continuous and Discrete Versions of Unbounded Model

In Sect. 2 we have seen the adequacy of unbounded growth model to the word frequency data. The proposed model structure resembles Gompertz curve. A number of growth curve families are related to this unbounded growth model in terms of the proliferation rate. Below we discuss properties of the families and discrete time versions of these in terms of proliferation rates in the context of limiting behaviour of parameters involved in the models.

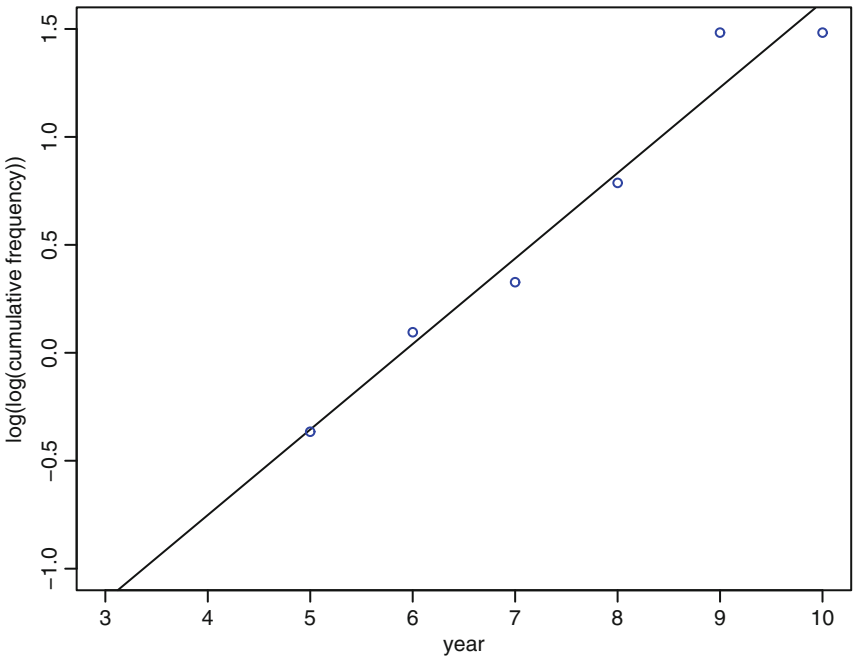


Fig. 13 Unbounded growth fit for cumulative frequency of the word "MaMatiManush"

3.1 Unbounded Gompertz Type Growth Curve & Other Related Growth Models

Proliferation rate $c \log y(t)$ of unbounded curve (1) is relatively slow in growth than that for *unbounded generalised logistic function* having proliferation rate $c\nu[1 - \{\frac{1}{y(t)}\}^{1/\nu}], c > 0, \nu > 0$. The former is logarithmically growing with $y(t)$, whereas the latter is polynomially growing. Here, one uses the fact that

$$\lim_{u\to\infty} u(1 - x^{1/u}) = -\log(x)$$

An exponentially growing proliferation rate may reduce to polynomial growth in limiting form of the model parameters $c > 0, \beta > 0, \nu > 0$, see also Dasgupta (2013) and Dasgupta (2015).

$$c\nu\beta[1 - e^{\{(1/y(t))^{1/\nu}-1\}/\beta}] \to c\nu[1 - \{\frac{1}{y(t)}\}^{1/\nu}],\ \beta \to \infty,$$
$$\to -c\log(1/y(t)),\ \nu \to \infty,\ t > 0 \qquad (2)$$

where $y > 1$. An appropriate unbounded model may be selected by checking steepness of growth from observed data.

3.2 *Discrete Version of Unbounded Growth Models*

In view of the relationship of proliferation rates among growth curve families in continuous time, one may consider discrete version $\Delta y(t)/\{y(t)\Delta t\}$ of the three proliferation rates given in (2) that gives rise to the discrete growth curves for $t \in N_0 = \{0, 1, 2, 3, \ldots\}$ satisfying the following *difference* equations in (3).

$$\frac{y(t+1)}{y(t)} - 1 = c\nu\beta[1 - e^{\{(1/y(t))^{1/\nu}-1\}/\beta}] \to c\nu[1 - \{\frac{1}{y(t)}\}^{1/\nu}],\ \beta \to \infty,$$
$$\to -c\log(1/y(t)),\ \nu \to \infty \quad (3)$$

That is, for $t \in N_0 = \{0, 1, 2, 3, \ldots\}$

$$y(t+1) = [1 + c\nu\beta\{1 - e^{\{(1/y(t))^{1/\nu}-1\}/\beta}\}]y(t) \quad (4)$$
$$y(t+1) = [1 + c\nu\{1 - (y(t))^{-1/\nu}\}]y(t) \quad (5)$$
$$y(t+1) = [1 + c\log y(t)]y(t) \quad (6)$$

For unbounded growth observations recorded at discrete time, the above family of models (4)–(6) with some assigned initial value for $y(0) > 1$ may be appropriate. This family of discrete proliferation rates, like its continuous time counterpart, covers a broad spectrum starting from exponential to logarithmic order of growth in a continuous manner, as the next class in the series (4)–(6) is a limit of the former.

References

Bardarson JH, Pollmann F, Moore JE (2012) Unbounded growth of entanglement in models of many-body localization. Phys Rev Lett 109(1):107–202

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

Dasgupta R (2015) Growth model of some vernacular words during political transition, Chap 10. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Dempster JP, McLean IFG (eds) (1998) Insect populations in theory and in practice. In: Symposium of the Royal Entomological Society of London. Kluwer Academic, Dordrecht

Wallace DI, Xinyue G (2013) Properties of tumor spheroid growth exhibited by simple mathematical models. Front Oncol 3:51

A Statistical Analysis of MicroRNA: Classification, Identification and Conservation Based on Structure and Function

Mohua Chakraborty, Ananya Chatterjee, Krithika S, and Vasulu T.S.

Abstract The microRNAs (miRNAs) are small non-coding RNAs which play an important role in gene regulation and are involved in several biological functions. Studies have shown that there are several hundreds of them across (human) genome. And one miRNA may be involved in several genes and several miRNA may target a gene. In this regard it is interesting to know whether these several known miRNAs show structural and functional similarities. Do they fall into recognisable groups with respect to their structure and function and does the length of miRNA follow evolutionary principles and are highly conserved?. This study with the help of statistical tools explores characterising, identification of (human) miRNA based on their structure and function, network analysis of their relationship and target genes and conservation of their length and sequence structure across species.

Keywords Pre and mature miRNA • Length variation • Clustering • Star graphs • miRNA target • Network analysis • Gene-specific-miRNA • miRNA across species

1 Introduction

In genome biology, there are two major discoveries that have helped us to understand the structure of the 'gene'—the biological unit of heredity that transfers genetic information from parents to offspring among living things. These are: (a) double-helix model proposed for DNA structure (Watson and Crick 1953; Franklin

M. Chakraborty
Assam University, Silchar, India

A. Chatterjee
Tata Consultancy Services, Kolkata, India

Krithika S.
College of London, London

Vasulu T.S. (✉)
Indian Statistical Institute, Kolkata, India
e-mail: vasulu@gmail.com

R. Dasgupta (ed.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_13

and Gosling 1953; Wilkins et al. 1953); and (b) the discovery of non-coding part of the gene the 'introns'—the genes in pieces (Gilbert 1978; Doolittle 1978). The molecular structure of 'gene' is the sequential arrangement of four types of nucleotides of DNA molecules (A, T, G, C) and constitutes 'exons' the coding part which is involved in the transcription and translation responsible for protein formation; and the 'introns' the non-coding part not involved in the synthesis of proteins (Watson 1965; Woese 1967, 2001; Crick 1968, 1988; Jeffreys and Flavel 1977; Gilbert 1978; Doolittle 1978). In human genome, most of the genes constitute 'exons' and 'introns', while a few of the genes have only 'exons'. Overall, genes control the biological fate of living organisms.

How do the genes function? While we are still far from understanding the complex function of genes, at least there are two major discoveries that help us to get an insight into the function of the gene. One is, the so-called, 'central dogma of molecular biology' and 'sequence hypothesis' proposed by Crick which describes the flow of genetic information from DNA to RNA to synthesis of 'proteins' (Crick 1958, 1968, 1970; Fantini 2006; Morange 2006, 2008). Cricks's central dogma of molecular biology describes the genetic mechanism that explains the transcription of information from DNA to messenger RNAs (mRNA) and each mRNA contains program to synthesis or translation of mRNA into a particular protein (Crick 1958, 1970). 'The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid' (Crick 1970). Sequence hypothesis: 'in its simplest form it assumes that the specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and that this sequence is a (simple) code for the amino acid sequence of a particular protein' (Crick 1958, 1970; Fantini 2006; Morange 2006, 2008; Doolittle et al. 2013). The second one is the recent discovery of a variety of non-coding RNAs: microRNA (miRNA), piwiRNAs (piRNA), small interfering RNAs (SiRNA), long non-coding RNAs (lncRNA), Enhancer and Promoter-associated RNAs (PARs) (Lee et al. 1993; Wightman et al. 1993; Lee and Ambros 2001; Lau et al. 2001; Doolittle et al. 2013). While Crick's model is concerned about the function of about 1 % of coded information of the (human) genome, the discovery of non-coding small RNAs explains the role of introns [constitutes a major portion (human) genome] and other transient RNAs in regulating the gene expression.

The small or ncRNAs help us to understand the epigenetic mechanism underlying the RNA splicing and post-transcriptional regulation of gene expression of protein coding genes (Brody and Abelson, 1985). The non-coding RNAs (ncRNAs) can be of two types: infrastructural and regulatory. The infrastructural ncRNAS are more related to housekeeping function whereas regulatory ncRNAS are concerned with epigenetic control of other RNAs.

2 MicroRNA

The miRNA, one of the non-coding RNAs, plays an important role in the regulation of gene expression (Ruvkun 2001). The miRNAs are most extensively characterised in plants and worms (in which they were first recognised). They arise from precursors (about 70–90 nucleotides long) transcribed from non-protein encoding genes. These transcripts contain sequences that form stem loop structures, which are processed by Dicer (or DCLI, for Dicer-like 1, in plants). The miRNAs they produce lead to the destruction (typically the case in plants) or translational repression (in worms) of target mRNAs with homology to the miRNA. The 'mature' miRNA—derived from primary (*pri-miRNA*) and precursor miRNA (*pre-miRNA*)—is short, single-stranded non-coding RNA molecule of length of about 22 base pairs (vary from 17 to 24) found in some viruses, plants, animals and appear to be nonexistent among bacteria.

The miRNAs are transcribed from different genomic locations, for example by RNA polymerase II/III, as long primary transcripts known as (primary) 'pri-miRNA' that form stem-loop structure and range in size from hundreds of nucleotides to tens of kilo bases. This gets cleaved by the complex in nucleus (microprocessor complex, consisting of the RNase III enzyme Drosha, and the double-stranded-RNA-binding protein, Pasha/DGCR8) to single-stranded/double-stranded RNA junction producing ∼70 nucleotide hairpin *precursor* (*pre-miRNA*). The *pre-miRNAs* get cleaved in the cytoplasm to produce *mature miRNA* of length of about 22 nt. The mature miRNA has target sites in hundreds of genes. The human genome harbours a variety of multitude miRNA which plays a crucial role in epigenetic regulation of gene expression of several biological processes. Also, strikingly, 30 % of miRNAs found in worms have close homologous miRNAs in flies and/or mammals. Thus, it seems that miRNAs are an ancient part of programs of gene regulation during development (Watson 1965; Priyapongsa et al. 2007).

2.1 miRNA Function

Since its discovery by Ambros (Lee et al. 1993; Olsen and Ambros 1999; Reinhart et al. 2000; Lee and Ambros 2001), several studies have reported myriad variety of miRNAs and their regulatory role across whole spectrum of biological functions/processes across species including Man (Ambros 2004; Vergoulis et al. 2015). A variety of miRNAs regulate the expression of target genes at post-transcriptional level. They are considered to regulate expression of genes involved in development, cell proliferation, apoptosis, response to stress, etc. (Fig. 1). Recently discovered functions of miRNA include control of cell proliferation, cell death fat metabolism, heart diseases, cancer biology (Brennecke et al. 2003, 2005; Xu et al. 2003; Soifer et al. 2007; van Rooij et al. 2007; Vergoulis et al. 2012; Xie et al. 2013). These (miRNA) form another layer of regulatory circuitry that exists in the cell

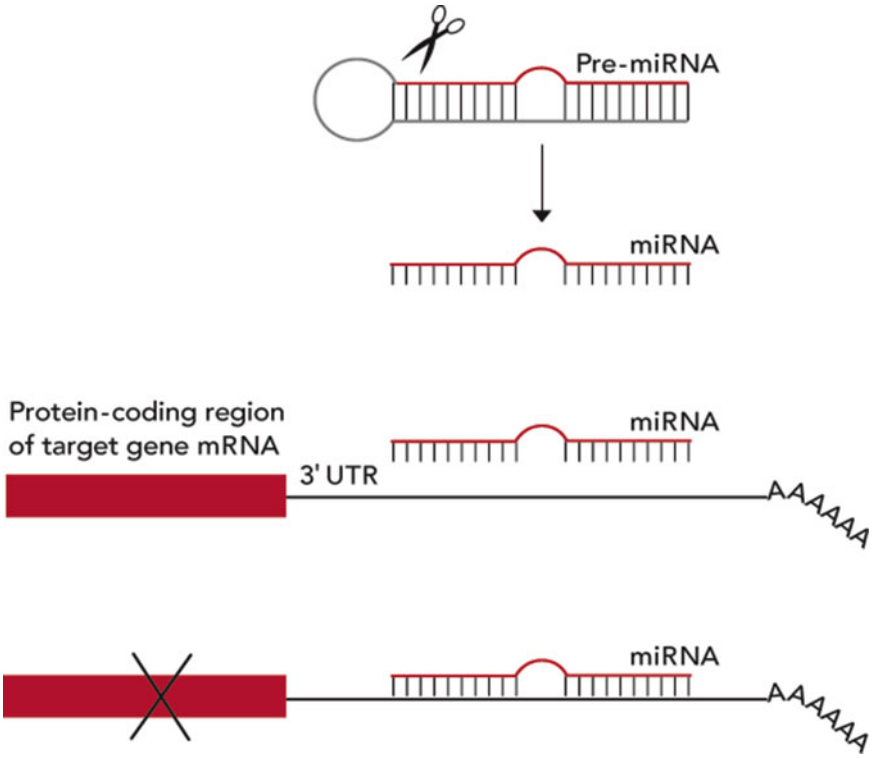


Fig. 1 Illustration showing steps of miRNA binding to the 3′ UTR region of the target mRNA

(Smielewska 2008). Thus any misregulation of miRNAs can lead to great regulatory imbalance in the cell, which in certain cases leads to cancerous phenotypes. In fact it was shown that miRNA profiles are changed in large number of cancer cells and over expression of miRNAs can lead to the development of cancers. The miRNA regulates gene expression either by switching off or by tuning target expression levels. A detailed account on miRNA and its functions can be obtained from a recent publication (Vergoulis et al. 2015).

2.2 miRNA and Target Gene

The miRNAs are produced from either their own genes or from introns. They can be encoded by independent genes, but also be processed (via the enzyme Dicer) from a variety of different RNA species, including introns, 3′ UTRs of mRNAs, long non-coding RNAs, snoRNAs and transposons (Priyapongsa et al. 2007).

The genes encoding miRNAs are much longer than the processed mature miRNA molecule; miRNAs are first transcribed as primary transcripts or pri-miRNA with a

cap and poly-A tail processed to short, 70-nucleotide stem-loop structures known as pre-miRNA in the cell nucleus. This processing is performed in animals by a protein complex known as the Microprocessor Complex, consisting of the nuclease Drosha and the double-stranded RNA-binding protein Pasha (Denli et al. 2004; Gregory et al. 2004; Han et al. 2004). These pre-miRNAs are then processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer, which also initiates the formation of the RNA-induced silencing complex (RISC). This complex is responsible for the gene silencing observed due to miRNA expression and RNA interference. The pathways in plants vary slightly due to their lack of Drosha homologs; instead, Dicer homologs alone effect several processing steps. The pathway is also different for miRNAs derived from intronic stem-loops; these are processed by Dicer but not by Drosha. Either the sense strand or antisense strand of DNA can function as templates to give rise to miRNA.

Efficient processing of pri-miRNA by Drosha requires the presence of extruded single-stranded RNA on both $3'$ and $5'$ ends of hairpin molecule. These ssRNA motifs could be of different composition while their length is of high importance if processing is to take place at all. The Drosha complex cleaves the RNA molecule $\sim$22 nucleotides away from the terminal loop. Most pre-miRNAs do not have a perfect double-stranded RNA (ds RNA) structure topped by a terminal loop. There are a few possible explanations for such selectivity. One could be that dsRNAs longer than 21 base pairs activate interferon response and anti-viral machinery in the cell. A miRNA may target more than one gene, often in several sites, and that one gene may be targeted by many miRNAs acting cooperatively. Individual miRNAs and their targets can share common regulators, also miRNAs and transcription factors (TFs) co-regulate their target genes. MiRNA in this motif stabilises the feedback loop to resist environmental perturbation (Sarazin and Voinnet 2014). Two classes of miRNAs exist with distinct preference for network subgraph: the first class is regulated by a large number of transcription factors while the second class of miRNAs regulates TFs. These two classes have different biological roles (Lee et al. 2004; Yu et al. 2008).

The epigenetic role of miRNAs in several biological functions at the cellular level implies that they have evolved along with other genes and organisms from lower organisms to *Homo sapiens sapiens*. Studies indicate that several miRNAs related to cell regulation appears to be common to animals and therefore must have co-evolved across the species.

2.3 miRNA Database

Since the discovery of miRNA and its role in the regulation of gene expression, in recent years, studies have come out with hundreds and hundreds of a variety of miRNA and their active role in gene expression related to different biological processes. Soon there has been catalogue of miRNA and the list has identified more than one thousands miRNAs and their role and target gene (Griffiths-Jones

et al. 2005; Soifer et al. 2007; Jiang et al. 2009; Vergoulis et al. 2012; Xie et al. 2013). This has led to development of databases like miRBase, miRecord, mirPub, mirCancer, etc., http://www.microrna.gr/mirpub/, Griffiths-Jones et al. 2003, 2005). This has opened up further issues of its identification, bioinformatics methods for motifs related to network pathways, regulatory loops linking diseases (Zamore 2002; Zeng et al. 2002; Zhang et al. 2015) and studies on population genetics issues concerning the evolutionary significance, selection and the conservation of miRNA across species.

Myriad variety of miRNAs has raised some of the problems of prediction, identification of miRNA especially due to factors like similar structure and some are even 'pseudo miRNA' which mimic but are not related to gene target (Mathews et al. 1999). Various methods of classification and categorization of miRNAs into families have been employed based on various parameters such as sequence similarity, homology, seed sequence similarity, etc., (e.g., Sinha et al. 2009). Some of the widely popular classification methods include the naming convention used in miRBase and division into conserved, non-conserved family used in TargetScan and the Phylogeny-Bootstrap-Cluster (PBC) pipeline.

The miRBase classified miRNAs with similar mature forms together and assigned them the same id numbers. In recent releases of miRBase, a 'miRNA family (miFam)' feature was present, which clustered similar miRNA precursors together based on computational analysis and manual inspection. In miRBase miRNAs are named using the 'mir' prefix and a unique identifying number (e.g., miR-1, miR-2, ..., miR-89, etc.). The identifying numbers are assigned sequentially, with identical miRNAs having the same number, regardless of organism. Nearly identical orthologs are also given the same number. Identical or very similar miRNA sequences within a species are given the same number, with their genes distinguished by letter and/or numeral suffixes, according to the convention of the organism. A uniform system for miRNA annotation can be found from Ambros et al. (2003).

In this regard, a few bioinformatics softwares have been developed to recognise and identify the miRNA types. For example, the 'TargetScan' software for Human miRNAs is categorised into conserved and non-conserved families. The categorization is in the following manner: broadly conserved: conserved across most vertebrates, usually to zebra fish and conserved: conserved across most mammals, but usually not beyond placental mammals and third category is poorly conserved: it includes all others. Whereas the 'Phylogeny-Bootstrap-Cluster (PBC) pipeline' identifies miRNA families based on branch stability in the bootstrap trees derived from overlapping genome-wide miRNA sequence sets. A 'Vote' algorithm was designed to automate the process of identifying and evaluating potential families (Huang and Gu 2007). The above computational methods are limited to identification and classifying miRNA and do not consider categorization of miRNA for investigating the patterns or identification based on structure or target or other criterion.

There are some underlying problems with the current methods. First, the mature miRNAs were often used as the classification criteria in the methods. This can

decrease the sensitivity of finding paralogous miRNAs, as the mature part of a duplicated copy of a miRNA is not necessarily under strong selective pressure. Secondly, due to the short length of the mature forms, false classification caused by convergent evolution is very likely to happen. A classification method based on a wide set of parameters rather than one single property will provide a better classification scheme for miRNAs. Since mature miRNAs are derived from precursor miRNA, variation in properties of precursor miRNA should also be taken into account while defining a classification scheme. In the present study we investigate the classification and characterization of miRNA (microRNA database) using statistical methods so as to get patterns, and examine the network analysis of miRNA function. The study compares a couple of miRNAs sequences across species to investigate the conservative nature or evolutionarily stable structure.

2.4 Objectives of the Study

The study investigates three aspects of miRNA structure and function: (1) Classification and characterization, (2) Interrelationship between miRNA and target genes and (3) Evolutionary conservation of miRNA across species. This has been attempted by using various statistical techniques used to observe and understand the pattern of distribution of these parameters in the entire human miRNA dataset derived from miRBase and the biological significance of the variation of these parameters within different human miRNAs.

Here we have attempted to identify, classify and characterise human miRNAs into certain groups based on various parameters *viz.*, length of the precursor miRNA, chromosomal distribution of miRNA, length of the mature miRNA, target genes and function of target genes. Further, miRNA and target interaction is studied by building networks that represent association of miRNAs and their target genes. Each of the parameters shows a particular distribution pattern within the human miRNA dataset, based on which we tried to cluster the miRNAs into groups. The study investigates the conserved nature of miRNA across species by comparing a set of same miRNAs related to a set of corresponding common genes across selected species.

3 Materials and Methods

3.1 Data Source

MiRNA sequences of human (both pre-miRNA and mature miRNA) were downloaded from miRBase of 2010.The validated targets of respective miRNA were retrieved from miRECORDS.

3.1.1 miRBASE

The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.

3.1.2 miRECORD

The 'miRecord' is resource for animal miRNA–target interactions. miRecords consists of two components. The Validated Targets component is a large, high-quality database of experimentally validated miRNA targets resulting from meticulous literature curation. The Predicted Targets component of miRecords is an integration of predicted miRNA targets produced by 11 established miRNA target prediction programs (Xiao et al. 2009). Identical or very similar miRNA sequences within a species, which are either present in different genomic loci or have distinct precursor sequences, have the same number with genes distinguished by letter or numeral suffixes.

3.2 Methods

For examining the variation among human miRNA datasets, we first consider length of the precursor miRNA as a parameter. For the present study, we have considered two separate datasets: the first set includes all such miRNA subtypes and named as the 'Redundant group' (R type) while the second set has been cleared off the redundancy caused by inclusion of these subtypes and only one member from each subtype has been included, this group is named as the Non-redundant group (NR type).

As mentioned above the entire human miRNA dataset retrieved from miRBASE is divided into (1) Redundant (R type) and (2) Non-redundant groups (NR type). Members of these groups are further classified separately based on length variation into four different clusters of 96, 109, 81 bp length, respectively, and the remaining pre-miRNAs are put into the fourth group. We then analyse these groups for the presence of patterns. We have used several softwares for statistical and bioinformatics analyses: Mega 4.1, for nucleotide substitution, composition, distance matrix and clustering analysis; SPSS, for principal component analysis and other statistical estimates (Kumar 2007); OligoSCAN, for estimating the GC content; Matlab, for counting the length of sequences and PAJEK (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) for obtaining miRNA network and target gene analysis.

3.2.1 Analysis of miRNA Subgroups Classified According to Variation in Length

The members of each group were subject to different analyses. All the miRNAs of each (length wise clustered) group were analysed for nucleotide composition, nucleotide substitution, conserved motif and GC content. Mutational variation between members of each length wise clusters was analysed by calculating the pairwise distance between these miRNAs through MEGA 4.1 and by generating the distance matrix, which is further used to construct tree by neighbour joining method. This helps us to identify and cluster together closely related miRNAs. The distance matrix is also used for Principal Component Analysis through SPSS .The miRNAs having high loadings under the same principal component were further clustered into smaller groups. These clusters were viewed with the help of scatter plot diagram (built by SPSS) and tree diagram (constructed by neighbour joining method through MEGA).

3.2.2 Analysis of miRNA Subgroups of Similar Length Classified According to Variation in Chromosomal Location

All the miRNAs of each (length wise clustered) group were further clustered according to their chromosomal location and similar set of analyses were carried out to collect information regarding nucleotide composition, nucleotide substitution, conserved motif and GC content. This was done to check if any pattern or similarity exists between miRNAs of same length located on same chromosome. In this regard Redundant miRNA dataset is considered because similar miRNAs located in different chromosome are given the same ID followed by numeral suffixes.

3.2.3 Analysis of miRNA Based on Chromosomal Distribution

Chromosome wise distribution of the miRNA in human is studied by counting the total number of miRNA located in each chromosome .The frequency distribution graph of miRNA in each chromosome summarises the scenario of distribution of miRNA into different chromosomes in human genome.

3.2.4 Analysis of Mature miRNA

Length variation among all human mature miRNA is studied by analysing the frequency distribution graph where total number of miRNA having a given length is plotted for each length. The average nucleotide composition of mature miRNA is calculated from the sum of each nucleotide through the entire dataset of mature

human miRNA. The variation of length of mature miRNA derived from precursor miRNA of equal length is studied by analysing the length variation within mature miRNA derived from precursor miRNAs of length 96 and 109 bp.

3.2.5 Analysis of miRNA and Targets

Validated targets of all human miRNAs were downloaded from miRECORDS. There are 774 targets for a total of 127 miRNAs. A network showing the association of these miRNAs with their respective target genes is then constructed through PAJEK. Various clusters of miRNA and their associated target genes were identified and the related function of each of the target gene is retrieved from 'GeneCards' along with its associated GO-Identity. Based on the type of function and cellular location of the site of action, the functions are further grouped into general classes, so as to provide a way to cluster the targets of each miRNA into further groups which describe the functions of the group members.

3.2.6 Comparison of miRNA Across Species

1 Collection of DNA nucleotide sequence: For the construction of phylogenetic tree, five species, namely *Homo sapiens sapiens* (human), *Gallus gallus* (chicken), *Canis familiaris* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat) had been selected.

   Then five genes had been selected which are common to all the five species; *viz*., carbonic anhydrase II (ca II), phosphofructokinase (muscle) (pfkm), cleavage and polyadenylation specific factor 2 (cpsf2), zinc finger CCCH type containing 15 (zc3h), Coagulation Factor II (Thrombin) receptor like II (cf2rl2). The above information was obtained from the NCBI website (www.ncbi.nlm.gov).
2 Collection of miRNA sequence of respective genes: The number of miRNAs have been obtained from www.miRDB.org. The miRNA sequences have been obtained from http://microsanger.ac.uk/, which target all the genes which are obtained above.
3 The given table (Table 1) indicates the particulars of five selected genes with their Gene IDs, among five species and their number of miRNAs targeting each gene.
4 Softwares used for the analysis of data obtained from databases for the above five genes from five species:- For across species comparison we have used SWORDS. SWORDS is a statistical software designed to handle large genome sizes of individual species data and does statistical analysis of DNA-WORD frequency distribution, clustering of species and plots based on 'nj-clustering' and 'star graphs' based on 'DNA-word frequencies' (Chaudhuri and Das 2001, 2002; Basu et al. 2003). In this study we have used SWORDS to compare genomes of five species and mature miRNA sequences of a few specific genes across five species.

Table 1 List of five genes across five species with details of gene identity and number of miRNA targeting the genes (obtained from ncbi)

Sl. no.	Species name	Gene ID	No. of MiRNAs targeting the gene
Gene name: (A) Carbonic anhydrase II			
1	*Homo sapiens*	GI:760	21
2	*Gallus gallus*	GI:396257	4
3	*Canis familiaris*	GI:477684	8
4	*Mus musculus*	GI:12349	11
5	*Rattus norvegicus*	GI:54231	10
Gene name: (B) Cleavage and polyadenylation specific factor 2			
1	*Homo sapiens*	GI:53981	26
2	*Gallus gallus*	GI:423416	10
3	*Canis familiaris*	GI:480230	4
4	*Mus musculus*	GI:51786	19
5	*Rattus norvegicus*	GI:299256	10
Gene name: (C) Phosphofructokinase (muscle)			
1	*Homo sapiens*	GI:5213	7
2	*Gallus gallus*	GI:374064	1
3	*Canis familiaris*	GI:403849	2
4	*Mus musculus*	GI:18642	3
5	*Rattus norvegicus*	GI:65152	2
Gene name: (D) Coagulation factor II (thrombin) receptor—like II			
1	*Homo sapiens*	GI:2151	23
2	*Gallus gallus*	GI:768449	1
3	*Canis familiaris*	GI:607963	10
4	*Mus musculus*	GI:14064	15
5	*Rattus norvegicus*	GI:29636	2
Gene name: (E) Zinc finger CCCH type containing 15			
1	*Homo sapiens*	GI:55854	7
2	*Gallus gallus*	GI:423992	12
3	*Canis familiaris*	GI:478831	6
4	*Mus musculus*	GI:69082	14
5	*Rattus norvegicus*	GI:362154	7

4 Results

4.1 Length Variation of Pre-miRNA

The length variation of precursor miRNAs is analysed in two ways. In the first method we took into account all the miRNA subtypes, while in the second one we took only one miRNA from each subtype. The redundant dataset consists of all similar miRNA types, some of them have been identified as subtypes (e.g., has-mir-584a-1, has-mir-584a-2, has-mir550-1, has-mir-550-2, etc.). This way we got two datasets, the former one designated as 'redundant miRNA dataset (R-dataset)' while the latter as 'Non-redundant dataset' (NR-dataset).

Length of human precursor miRNAs is found to vary within the range of 43–148 bp for Redundant dataset while for Non-redundant dataset it varies between 47 and 140 bp. The length variation in both Redundant and Non-Redundant dataset is seen to vary in a near about uniform manner, with maximum number of miRNAs having length between 80 and 109 and very few miRNAs lying in the extreme edges.

In redundant dataset length of 109 bp has the highest count of miRNAs, ranking up to a number of 41 followed by length 96 bp with a count of 37 miRNAs, while in case of Non-Redundant dataset 96 bp length holds the highest count of 42 miRNAs followed by 109 bp length with a total of 40 miRNAs (Fig. 2a, b).

The obtained database of about 800 human pre-miRNA shows variation in length from a range of 43 bp length to about 150 bp with variable copy numbers across the human chromosomes. Several studies have shown that a number of miRNA and its target are conserved across species (Piriyapongsa et al. 2007; Kamanu et al. 2013; Warnefors et al. 2014). This implies that the length of miRNA is also under natural selection and influenced by other evolutionary forces. If so, the distribution pattern of the length variation in R and NR types of pre-miRNA should reflect the evolutionary significance. Figure 2a, b, shows a similar distribution pattern for both R and NR types, in case of R type pre-miRNA shows higher range of length variation than NR types. Overall, the distribution of length variation of pre-miRNA shows similar pattern in R and NR types. This possibly suggests that the additional duplicate or similar copies of pre-miRNA types that co-occur in the human genome do not appear to change the overall distribution significantly. The distribution of pre-miRNA length shows a trend of multiple modes with high frequency in case of 85 bp length, 96 and 109 bp being the highest of the three pre-miRNA types. The more frequent occurrence of 109 bp length pre-miRNA suggests its relative importance in several biological processes in Man. To further investigate the pattern in each class of miRNA types in human genome, we have selected two most frequently occurring pre-miRNA types' viz., (1) pre-miRNA lengths of 96 bp [with a frequency of 52 redundant (R) and 35 non-redundant (NR) varieties] and (2) 109 bp (with a frequency of about 40 bp length).

4.1.1 Length 96 bp Pre-miRNA

The miRNAs having 96 bp length, one of the frequently occurring pre-miRNA is considered for analysis. From the database of miRNA having 96 bp length we have obtained a total of 52 varieties (has-miRNA) of redundant type (R type) and 35 varieties of non-redundant types (NR) which are associated with different target genes. The nucleotide composition, nucleotide substitution, chromosome wise distribution of the miRNAs in each cluster varies. Four miRNA are located at Chromosome 7 and 8, whereas three miRNAs occur at chromosomes; 5, 9, 12, 14, 15 and X, respectively. The GC content varies from 33 to 60 % with an average of 44 % for both redundant (R) and non-redundant (NR) 96 bp length miRNA (Fig. 3). The frequency (%) distribution of GC content in both the datasets shows a trend of bimodal with two equal peaks at 35 and 50 % in R and NR types with a cut off

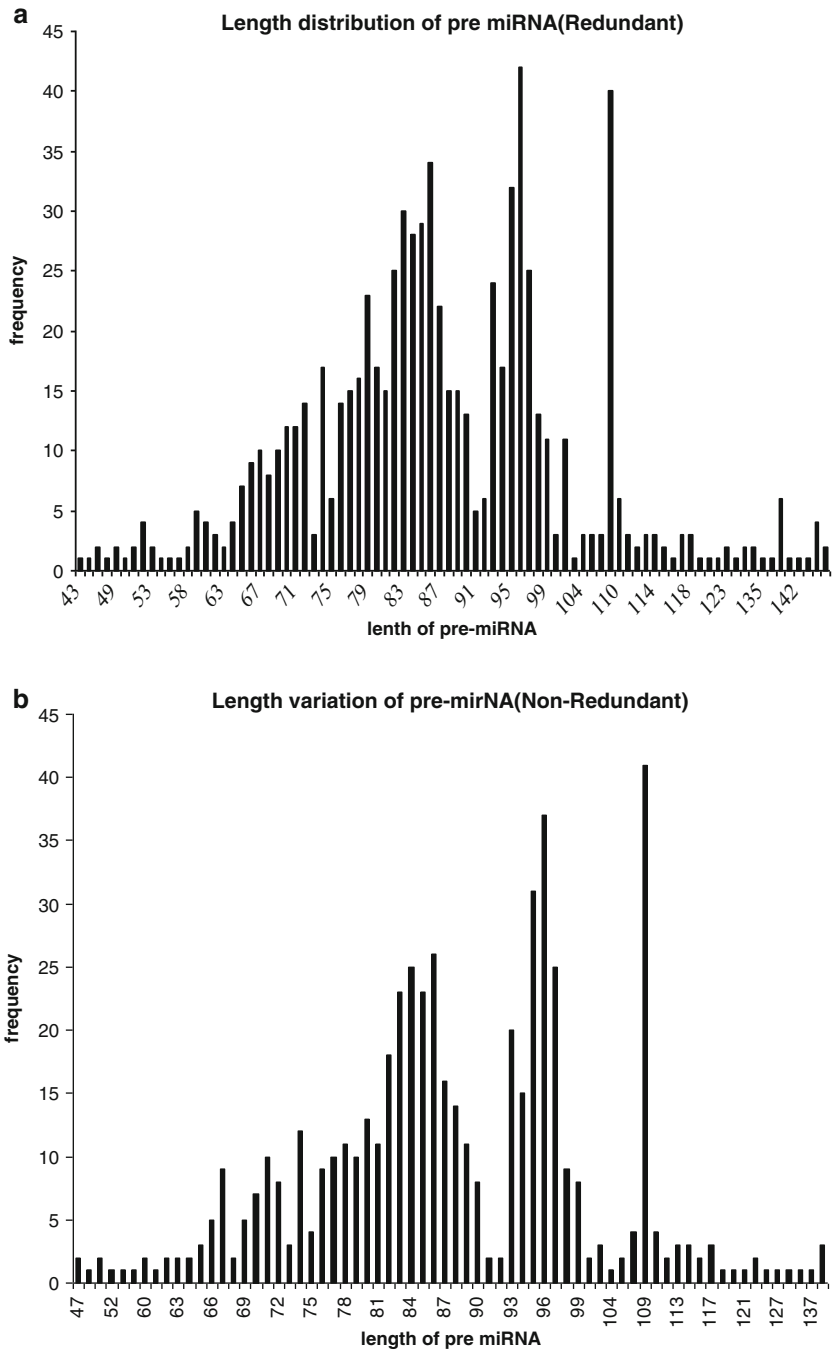


Fig. 2 (**a**) Length variation of pre-miRNA (R dataset including all the subtypes). (**b**) Length variation of pre-miRNA (NR dataset including one from each subtype)

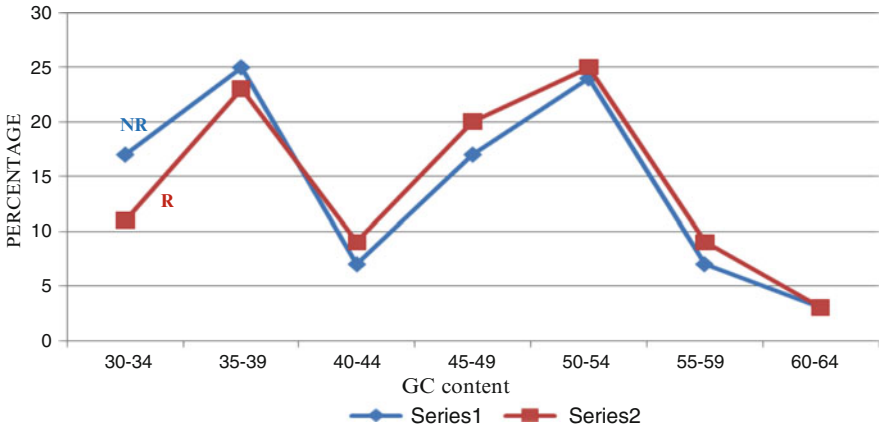


Fig. 3 Percentage distribution of GC content in case of 96 bp length human pre-miRNA (R) redundant (*red colour*—series 2) and non-redundant (NR) (*blue colour*—series 1) of 52R and 35NR varieties, respectively

two clusters at 40 bp length which shows least GC (6 %) content. It is interesting to observe two clusters with different percentage of GC content: 30–40 % and 40–60 % (consisting of about 20 types of miRNA in the second and about 15 in the first clusters).

Clustering of Pre-miRNA of Length 96 bp

To investigate the further classification and groups among the 35 pre-miRNA 96 bp length type, we have considered clustering based on distance analysis and by principal component analysis. The neighbour joining tree clustering obtained by MEGA 4.1 shows two major clusters (Fig. 4). The first major cluster has 12 members and second with 23. The second major cluster has two subclusters with further two subclusters with 17 and 6 members. The principal component analysis displays about ten components showing ~82 % of variance. Further based on the first two components we have obtained a scatter plot (Fig. 5). Based on the high loadings of different miRNAs of 96 bp length under each principal component, all the miRNAs are grouped into five separate clusters. Each of the five clusters was further analysed for investigating further subgroupings. A PC plot based on principal component analysis showed clusters, miRNA types identified by their numbers 6, 8, 10, 11 and 21 forms a cluster. In another cluster, miRNA members with numbers 25, 27 and 32 form a group.

By comparing the results obtained by scatter plot and tree diagram we observe considerable similarity in the distribution pattern of the miRNAs of each cluster in both the figures. Based on the structural similarity several of the 35 different varieties of pre-miRNA of 96 bp length do show clustering. Studies indicate that

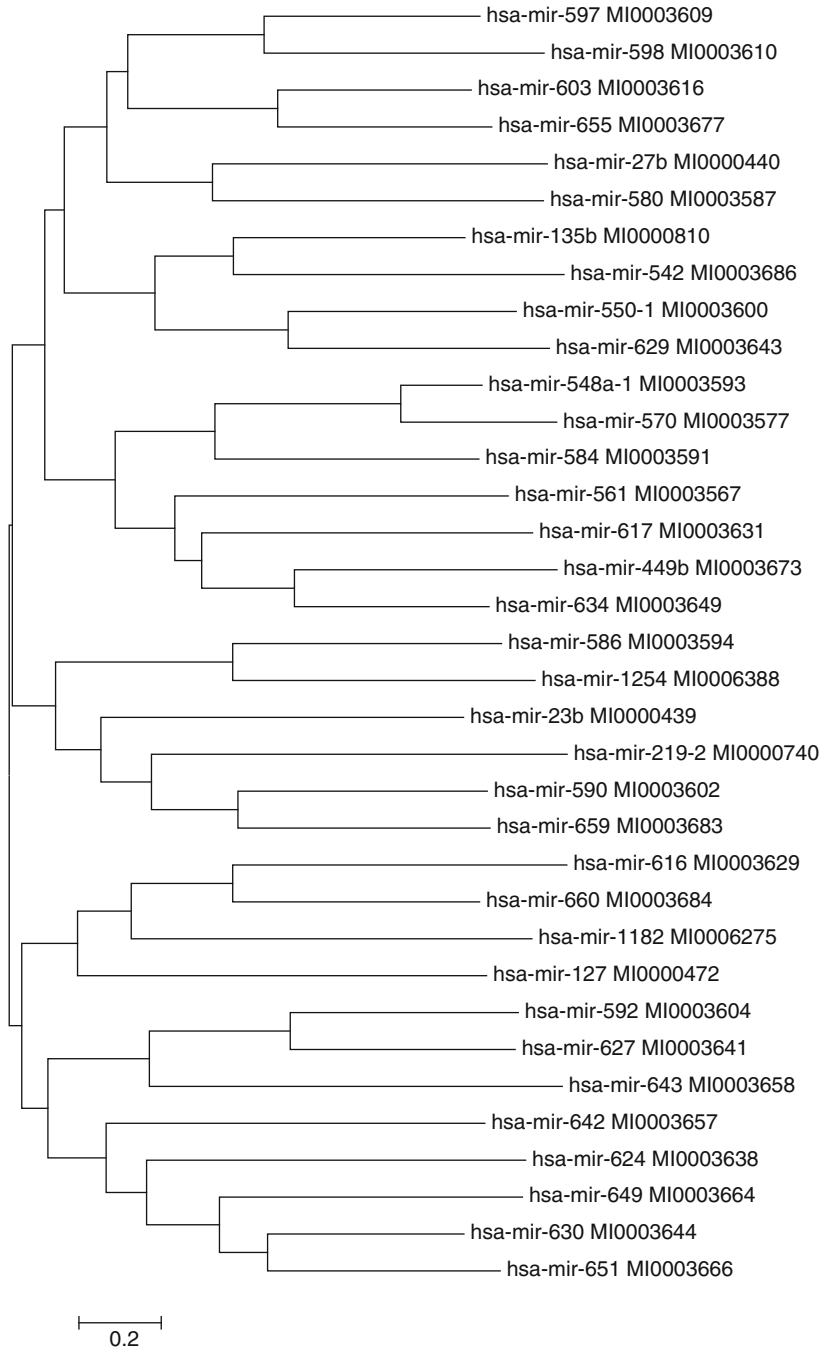


Fig. 4 Clustering of 96 bp length pre-miRNA by rooted nj-tree clustering

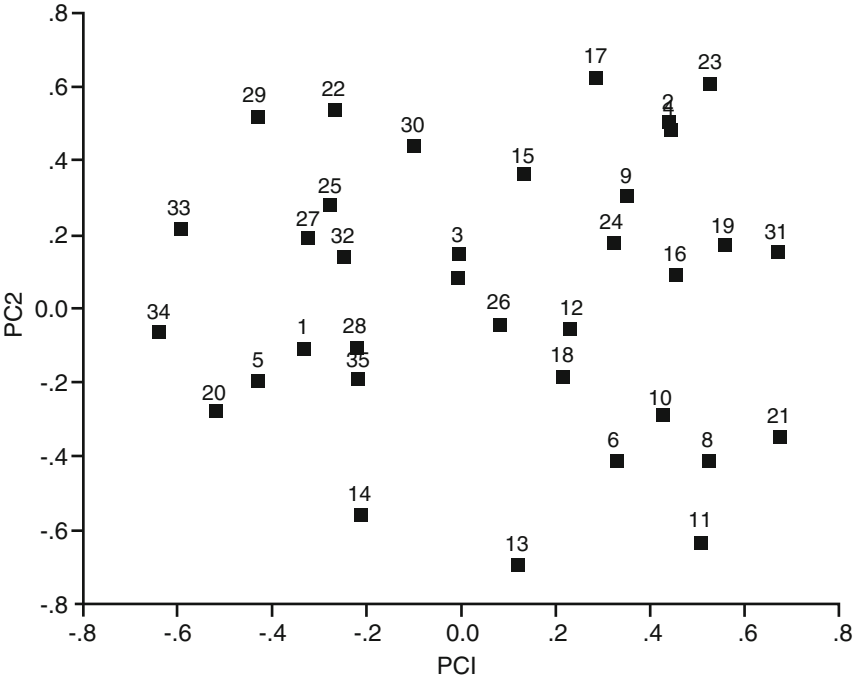
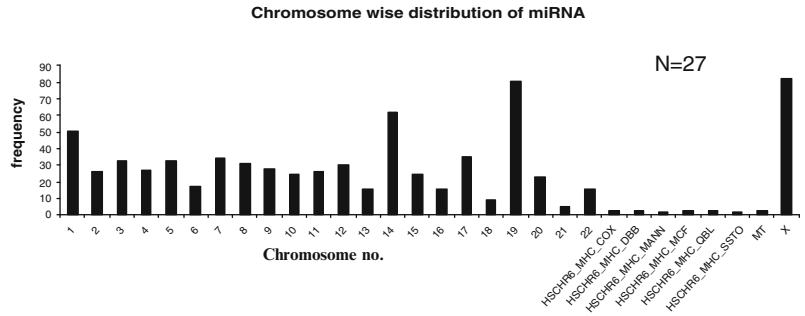


Fig. 5 Scatter plot (PC Plot) showing distribution of miRNA types of fixed 96 bp based on PC1 and PC2

the miRNA are in general conserved through evolutionary scale. As the results shown below gives credence to the possibility that similar miRNA types that target genes associated with similar functions across the species must have co-evolved, and were selectively neutral thus conserving the sequence structure of miRNA types associated with the biological functions.

4.1.2 Length 109 bp Pre-miRNA

The pre-miRNA with 109 bp length has 35 varieties of redundant type and 26 non-redundant types distributed across several chromosomes, the first chromosome has seven members and chromosomes 9 and x has four members (Fig. 6). In both the groups the GC content varies from 35 to 59 % with an average of 51 % (Fig. 7). The percentage distribution of GC content among the class of 109 bp pre-miRNA type shows a trend of single distribution, despite a dent at 50 % of GC content, whereas 96 bp length shows more than 70 % GC content. We want to investigate whether the miRNA located in the same chromosome does show structural similarity. For example, seven members of has-mir-181a-1 type located in chromosome 1 show similar structure and form a cluster! The clustering



Ch 1 – has-mir-34a,181a1,b1,199a2,205,214, 215
Ch 2 – has-mir-10b,216a,217
Ch 3 – has-mir-720
Ch 4 – has-mir-218-1,
Ch 5 – has-mir-218-2,
Ch 6 – has-mir-219,
Ch 7 – has-mir-182, 183,
Ch 9 – has-mir-7-1, 181a2,199b, 204,
Ch 10 – has-mir-202,
Ch 11 – has-mir-192, 210
Ch 12 – has-mir-196a2
Ch 14 – has-mir-203
Ch 15 – has-mir-7-2, 211
Ch 17 – has-mir-10a, 212
Ch 19 – has-mir-7-3, 181c
Ch X – has-mir-220a, 221, 222, 223

Fig. 6 Chromosome wise location of pre-miRNA in man

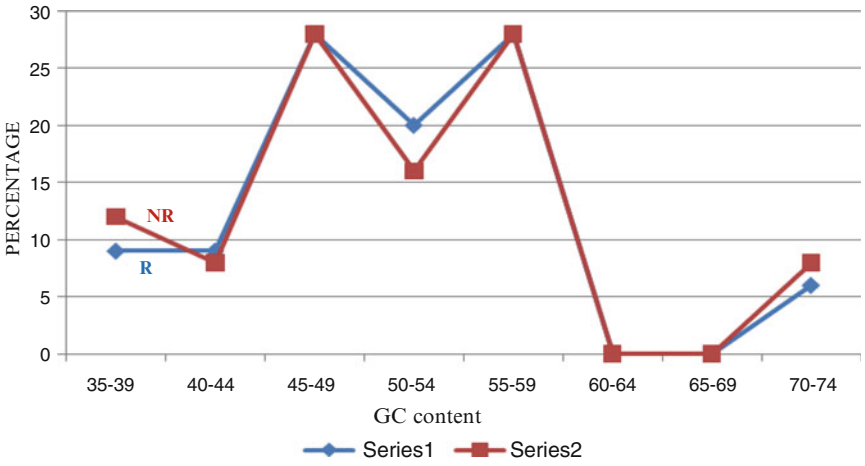


Fig. 7 Percentage of GC content in case of miRNA of length 109 bp—redundant (R) (*blue*—series 1) and non-redundant (NR) (*red*—series 2)

tree obtained from the distance matrix shows (Fig. 8) two major clusters, the second major cluster with two subclusters. The clustering pattern shows miRNAs located in different chromosomes do cluster together. The first cluster has members from chromosome 19, 6, 10, 1, 2, 5 and 11. Similar pattern emerges in case of other clusters. Interestingly it does show some outliers as well. The miRNA in chromosome 12 (has-mir-196a), 9 (has-mir-204) and 3 (has-mir-720) tend to form

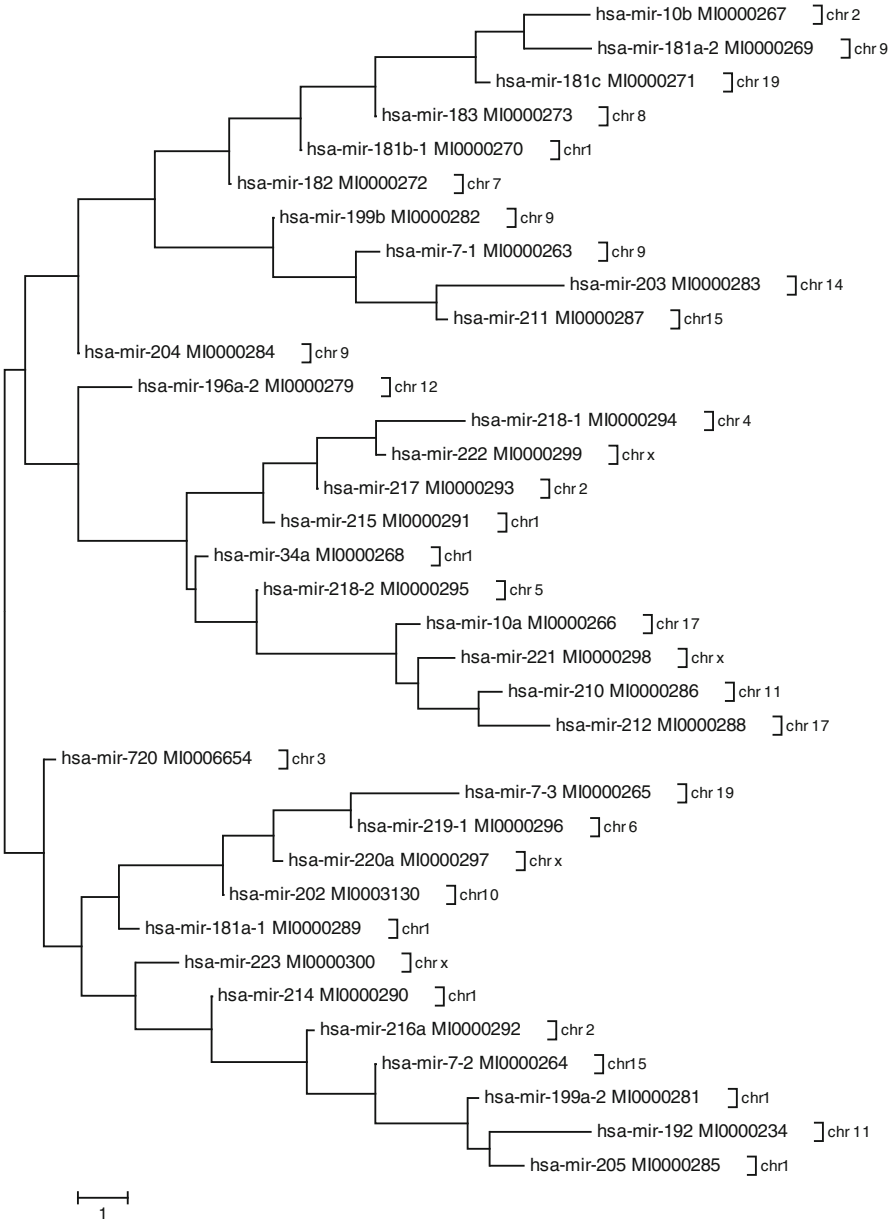


Fig. 8 Clustering of 109 bp length pre-miRNA represented through rooted tree (n–j method)

outliers in respective clusters (Fig. 8). Similar clustering pattern is obtained in case of redundant dataset, though it does not show outliers like the one that has been observed in non-redundant dataset. Chromosome wise location of pre-miRNA shows unequal distribution: X-chromosomes and ch-19 show the highest density miRNA located, whereas a few miRNA types are located in chromosomes ch-18, ch-23, 24.

The principal component analysis of 27 types of pre-miRNA which has a fixed length 109 bp shows about ten components (covering of about 82 % of variance). Further based on first two components we have obtained a scatter plot (Fig. 9). The figure shows clusters involving two to seven members spread throughout the plot. For example, 14, 16, 25, 8, 24, 21 form a cluster. Similarly 26, 11, 27, 23, 2, 4 form another cluster. Both the tree- clustering and pca-plot overall show a trend of similar clustering of miRNA types, though several of them form different clusters, for example, pre-miRNA 210 and 212, 196-a-2 and pre-miRNA 223, 720 are in the same cluster in both the clustering patterns.

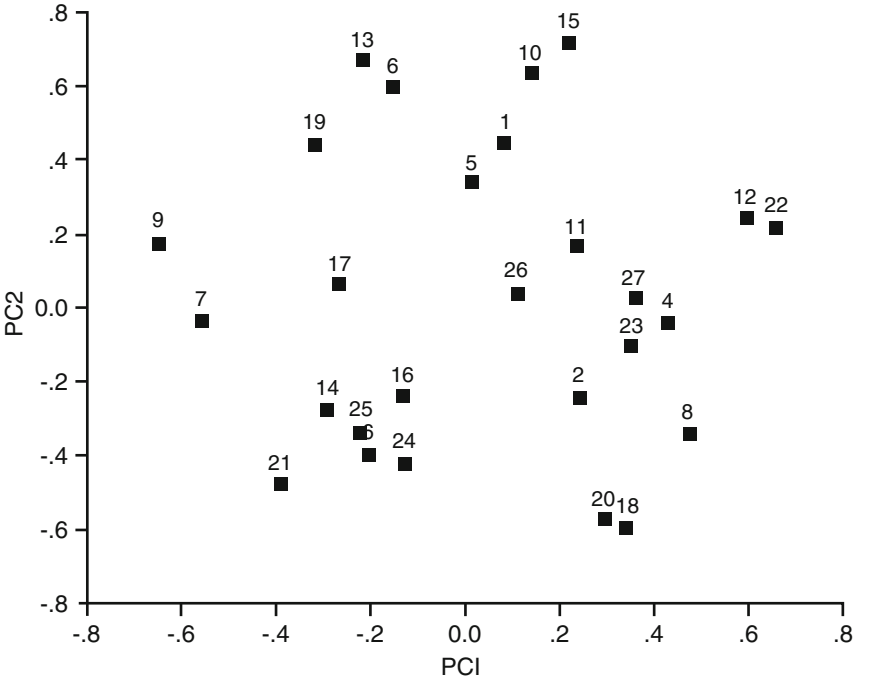
4.2 Mature miRNA

4.2.1 Length Variation of Mature-miRNA

The length of mature miRNA is seen to vary between 17 and 27 bp. The percent frequency distribution graph of length variation of mature miRNA (derived from non-redundant pre-miRNA) shows normal distribution pattern with the peak at 22 bp (45 %) and decreasing uniformly on both the sides of the peak. Overall, the variation in length of different types of mature miRNA in Man shows normal distribution. This is an indication of its evolutionarily conserved nature. The percentage composition of each of the four nucleotides in mature miRNA also shows an overall uniform distribution (A: 22.3 %, C: 23.0 %, G: 26.2 and T: 28.5).

Mature miRNA of Length 96 bp and 109 bp

Figure 10 indicates that in general, the mature miRNA (irrespective of variation in length of its corresponding pre-miRNA 96 or 109 bp) shows normal distribution. It will be interesting to investigate whether the subtypes of mature miRNA (derived from pre-miRNA with length 96 and 109 bp) also show normal distribution! Figure 11 shows the distribution of length (# bp) variation of mature miRNAs derived from pre-miRNA 96 and 109 types, respectively. Both differ in length: mature miRNA of pre-miRNA 96 bp vary from a range of 21 to 24 bp, whereas the mature miRNA of pre-miRNA 109 bp shows wide variation from 17 to 24 bp. In both the cases the mature miRNA 22 bp length shows a maximum frequency (45 and 35) of occurrence. Both show a single mode with normal distribution, the former shows a trend of positively skewed and the latter negatively skewed. This



1-hsa-mir-7-1				22-hsa-mir-219-1
2-hsa-mir-10a	7-hsa-mir-192	12-hsa-mir-204	17-hsa-mir-214	23-hsa-mir-220[a]
3-hsa-mir-34a	8-hsa-mir-196-a-2	13-hsa-mir-205	18-hsa-mir-215	24-hsa-mir-221
4-hsa-mir-181a	9-hsa-mir-199a-2	14-hsa-mir-210	19-hsa-mir-216	25-hsa-mir-222
5-hsa-mir-182	10-hsa-mir-202	15-hsa-mir-211	20 -hsa-mir-217	26-hsa-mir-223
6-hsa-mir-183	11-hsa-mir-203	16-hsa-mir-212	21-hsa-mir-218-1	27-hsa-mir-720

Fig. 9 Distribution of 27 miRNAs of fixed length (109 bp) by PC1 and PC2 plot

suggests that the length variation (number of bp) of mature miRNA is independent of its corresponding pre-miRNA lengths. However the results suggest that 22 bp length of mature miRNA is evolutionary stable and is under selection.

4.3 Mature miRNA and Function

The studies indicate a very complex nature of variety of functions for each mature miRNA type. Each of the mature miRNA type is involved in different biological process (one miRNA—many functions) and each biological function is governed by numerable number of miRNAs (one function-many miRNAs). Here for the study we have considered two miRNA-target clusters: hsa-miR-1 and hsa-miR-124. Both show very complex interactions of involving several functions and target genes.

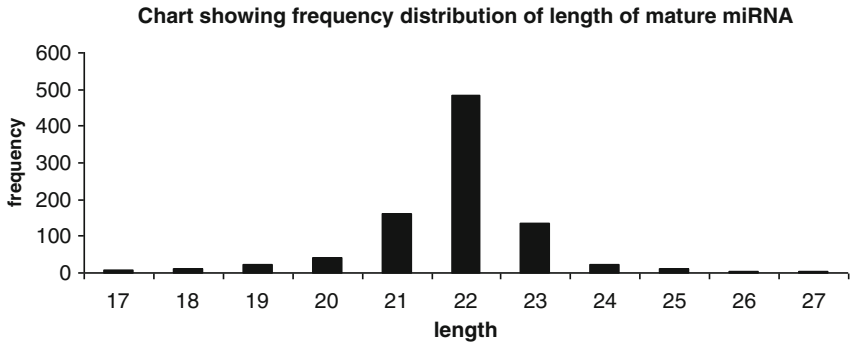


Fig. 10 Length (number of bp) variation of mature miRNA

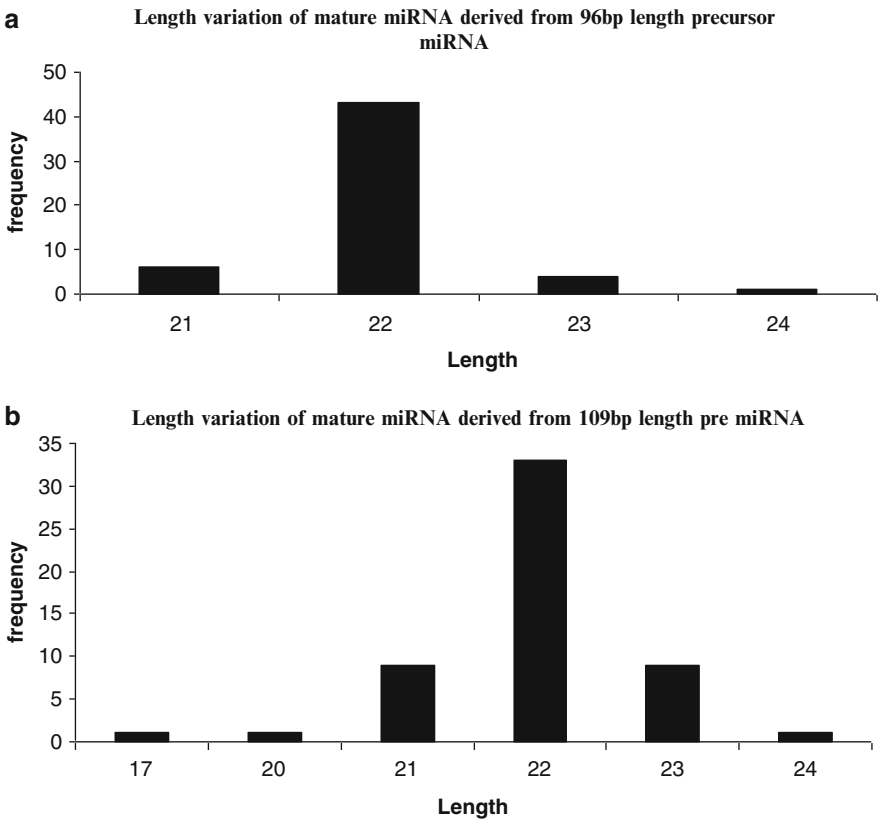


Fig. 11 Length (number of bp) variation of mature miRNA 96 bp and 109 bp

The function of the targets of hsa-miR-124 and hsa-miR-1 basically belongs to the general category that has been analysed from the information obtained from online sources (as given in Table 2). While grouping the functions into classes, related functions have been grouped into certain general classes like cell transformation, cell differentiation have been grouped into a general class of Cell development; cell membrane filaments, nuclear membrane filaments, membrane filaments related enzymes are grouped within Membrane filaments class; all immune cells related proteins and enzyme that is related to either T cell, B cell or complement system are grouped within Immune cells related protein class. Table 2 shows has-mir-1 is involved in at least 28 major functions while has-mir-124 is involved in at least in 36 functions and both play a crucial role in a few common functions (e.g., membrane filament and cell development). In some biological functions they were reported to be associated more frequently than others. For example, hsa-mir-1 is more active in case of cell development (10 times) and signal transduction (9 times) while has-mir-124 is more active in case of TF (17 times), membrane filament protein (16 times) and cell development (13 times) (Table 2 and Fig. 12a, b).

4.4 miRNA and Targets

From the information available (Table 2) about the function and frequency of occurrence of the has-miRNAs we have built a network of miRNAs interrelationship between variety of functions and its corresponding target genes. Figure 13a, b shows complex interaction of miRNA and its target genes. A large number of clusters are seen where one single miRNA regulates more than one target, hsa-miR-124 controls about 201 targets and hsa-miR-1 influences 101 target genes followed by hsa-miR-373 which has control over 64 target genes. Each miRNA is seen to control target genes having different function. An analysis of two miRNA and its associated gene cluster shows how a single miRNA can influence gene related to a wide different variety of function.

4.4.1 A miRNAs Targeting ‘VEGF’

We take an example of network relationship between single genes which is targeted by several miRNA. For example, VGEF gene is targeted by a large number of miRNAs. The phylogenetic clustering of the variety of miRNAs that target VGEF it is given below (Fig. 14). *Vascular endothelial growth factor* (*VEGF*) is a chemical signal produced by cells that stimulates the growth of new blood vessels. It is part of the system that restores the oxygen supply to tissues when blood circulation is inadequate. VEGF’s normal function is to create new blood vessels during embryonic development, new blood vessels after injury, muscle following exercise, and new vessels (collateral circulation) to bypass blocked vessels.

Table 2 Different type of functions of hsa-mir-1 and hsa-miR-124 and related frequency of occurrence of each type

Serial number	Functions of hsa-mir-1	Frequency of occurrence	Functions of hsa-mir-124	Frequency of occurrence
1	Cell development	1	Biodegradative pathway	1
2	Protein trafficking	1	Biomineral formation	1
3	(Metalloproteinases)	1	Biosynthetic and degradative pathway	1
4	RAS GTPase superfamily	1	Biosynthetic pathway	14
5	Adenosine deaminase	1	Cell cycle	9
6	Biosynthetic/degradative pathway	1	Cell cycle/Protein transport	1
7	Biosynthetic pathway	7	Cell development	13
8	Cell cycle	1	DNA binding protein	1
9	Cell development	10	Energy metabolism	3
10	DNA repair enzyme	1	Enzyme	1
11	DNA-mediated transposons	1	G proteins	2
12	Enzyme	1	G-protein dependent	2
13	GTP binding	2	Guanine dependent enzyme	1
14	Histone protein	2	Immune cells related protein	6
15	HSP	2	Inhibitory protein	2
16	Immune cells dependent protein	1	Membrane filament protein	16
17	Inhibitory protein	1	Membrane protein	6
18	Membrane filament protein	9	Membrane trafficking	1
19	Membrane protein	5	Metabolic pathway	1
20	Phospho-transferase	1	Mitochondrial matrix enzyme	1
21	Post TF	3	Mitochondrial oxidase	3
22	Protein trafficking	1	Post replication enzyme	1
23	Protein-DNA and protein–protein interactions	1	Post TF	3
24	RAS GTPase superfamily	1	Protein trafficking	2
25	Signal transduction	9	Regulatory protein	2
26	TF	6	Replication protein.	2

(continued)

Table 2 (Continued)

Serial number	Functions of hsa-mir-1	Frequency of occurrence	Functions of hsa-mir-124	Frequency of occurrence
27	TR	2	Repressor in Signal transduction	1
28	Transport protein	5	Signal transduction	10
	Total	78	TF	17
	Function not defined	31	TF(RNA pol)	1
	Grand total	109	TF/TR	1
			TR	4
			Transferase activity	1
			Transport protein	9
			Tumour suppressor	1
			Total	141
			No information available	55
			Total including blanks	196

When VEGF is over-expressed, it can contribute to disease. Solid cancers cannot grow beyond a limited size without an adequate blood supply; cancers that can express VEGF are able to grow and metastasize. Over expression of VEGF can cause vascular disease in the retina of the eye and other parts of the body. VEGF expression is regulated by multiple miRNAs. Some miRNAs share a common binding site, whereas other miRNAs have their own binding sites in the 3′-UTR of VEGF. Different miRNA combination patterns can produce coordinate action, which increases the repressive effect of miRNAs on VEGF translation, or competitive action, which failed to generate further repression of VEGF translation. Competitive action among miRNAs can weaken the total repressive power of a miRNA group. 12miRNA types related to the function form a single cluster showing structural similarity, while another 25 miRNA types basically cluster into three clusters, each with subclusters.

4.5 miRNA: Across Species

4.5.1 Comparison of Genome Sequence Across Species

Evolution is primarily changed in the genetic structure across species. However, the rate of evolutionary change varies widely, it varies between species and between regional populations within species. Evolutionary rate also varies with respect to some genetic traits or genes within and between species—some are slow, fast and some are stable and conserved. Some of the biological functions which are common across the species similar across the species and the genes that are involved

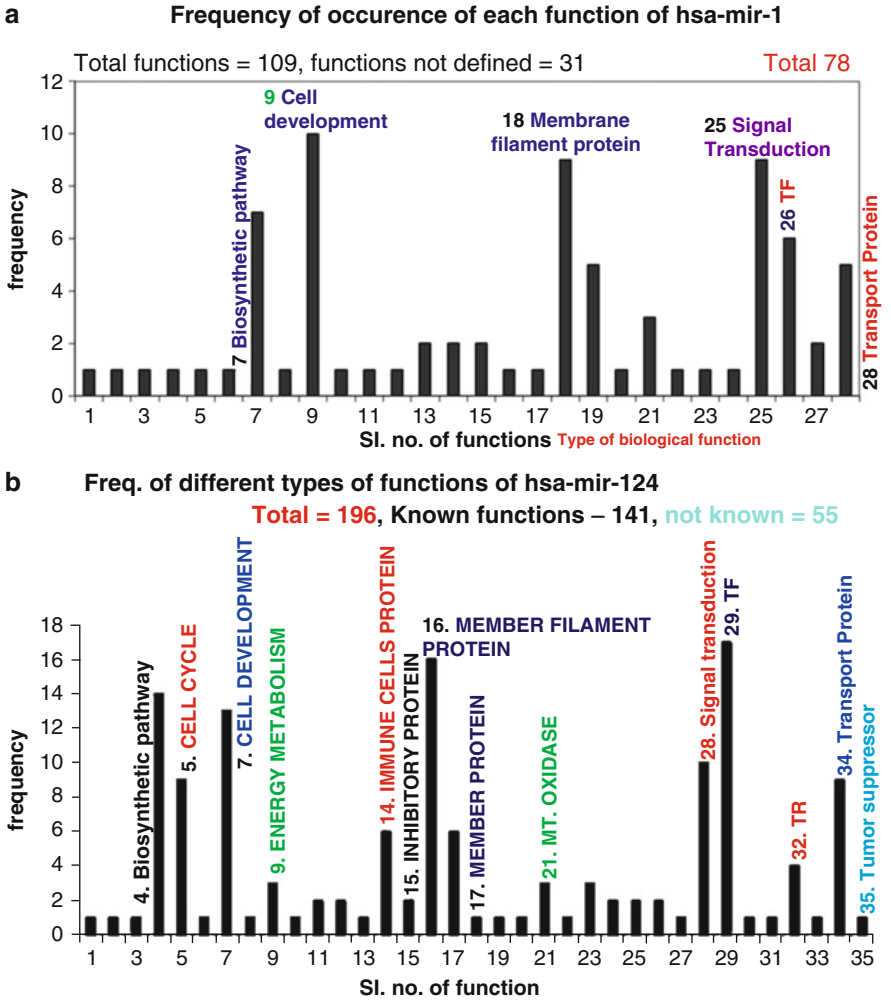


Fig. 12 (**a**) Frequency variation of mature miRNA—hsa-mir-1 and its multiple functions. (**b**) Frequency variation of mature miRNA—hsa-mir-124 and its multiple functions

in these common biological functions therefore are stable across species and are evolutionarily conservative. And some of the miRNAs involved in these genes and their functions are also stable and found to be conserved across species. The normal distribution observed in case of mature miRNA is an indication that the length 22 bp for mature miRNA is stable and is selectively neutral across species. To further illustrate the stability of miRNA across species we have investigated four miRNA that are common among four/five species. Figure 15a shows the comparison of five species with respect to four genes (Carbonic anhydrase II—(Ca II), Cleavage polyadenyhlation specific factor 2 (cpsf2), Phosphofructokinase (muscle)—(Pfkm)

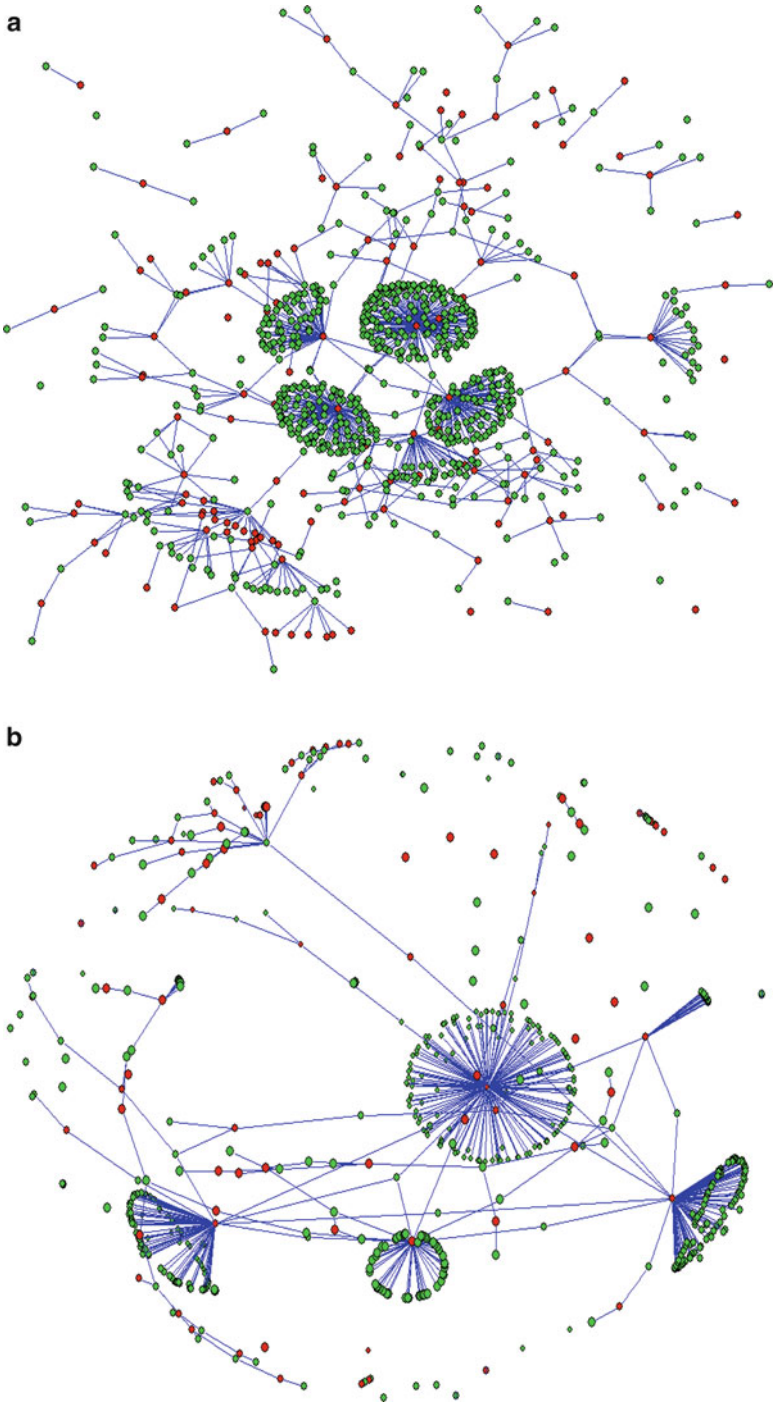


Fig. 13 (**a**) A 2D network representing interaction between miRNA (*red dots*) and their target genes (*green dots*). (**b**) A 2D network representing interaction between miRNA (*red dots*) and their target genes (*green dots*)

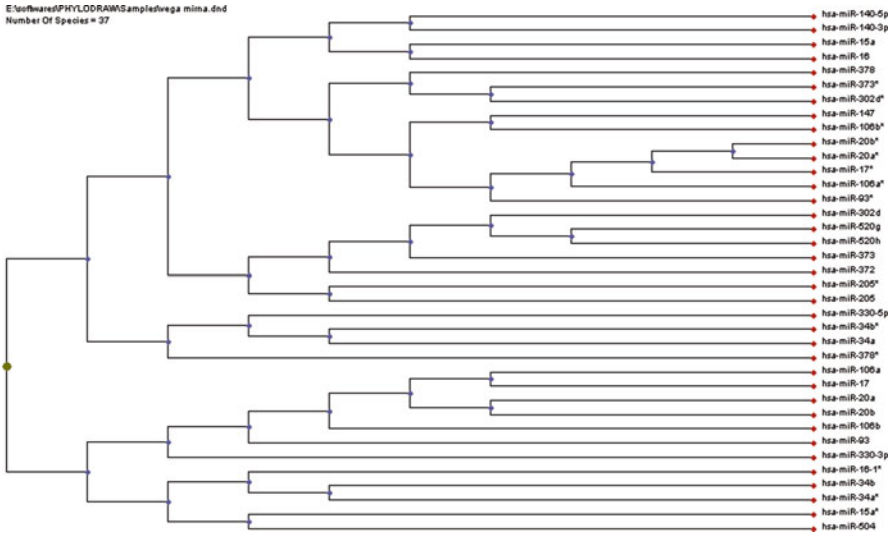


Fig. 14 Tree showing distribution of miRNA targeting VEGA based on pair wise distances

and Zinc Finger CCCH type—(zfc3h)) which are common to the species. The star graphs shown are based on trinucleotide DNA word frequency and nj-tree shows the clustering of five species based on genomes of the five species (*Homo sapiens* (Hs), *Canis familiaris* (cf), *Mus musculus* (Mm), *Rattus norvegicus* (Rt) and *Gallus gallus* (Ga). The nj-tree and star graphs based on genome sequence comparison of the five species for the gene CA II show two clusters (nj-tree) where *Homo sapiens* clusters with Canis familiaris, Ms and Rt form another cluster, whereas Ga departs as an outlier of the group. Same clustering pattern is observed in case of rest three genes. Similar pattern is also seen in case of star graphs. The results suggest that the evolutionary relationship between the five species remains the same with respect to each of the four selected genes.

4.5.2 Comparison of Gene Specific miRNA Across Species

Figure 15b shows the clustering pattern of four/five species based on gene specific miRNA for each of the four genes which are common to four/five species compared. In general the clustering pattern of the four/five species differs with respect to the miRNA and the clustering pattern based on genome sequences. In case of Zfc3h gene Mm, Rn and Cf form a very close cluster, except in Man with little differences in their miRNA for the gene. The mature miRNA shows evolutionarily stable sequence in case of three species and it differs in case of Man. The star graphs show similar pattern. In case of CaII the clustering pattern shows a single cluster with increasing differences from Mm, Rt, Cf, Ga to Hs. A comparison of the star graphs (drawn from trinucleotide miRNA sequences) for the gene Zfc3h across the

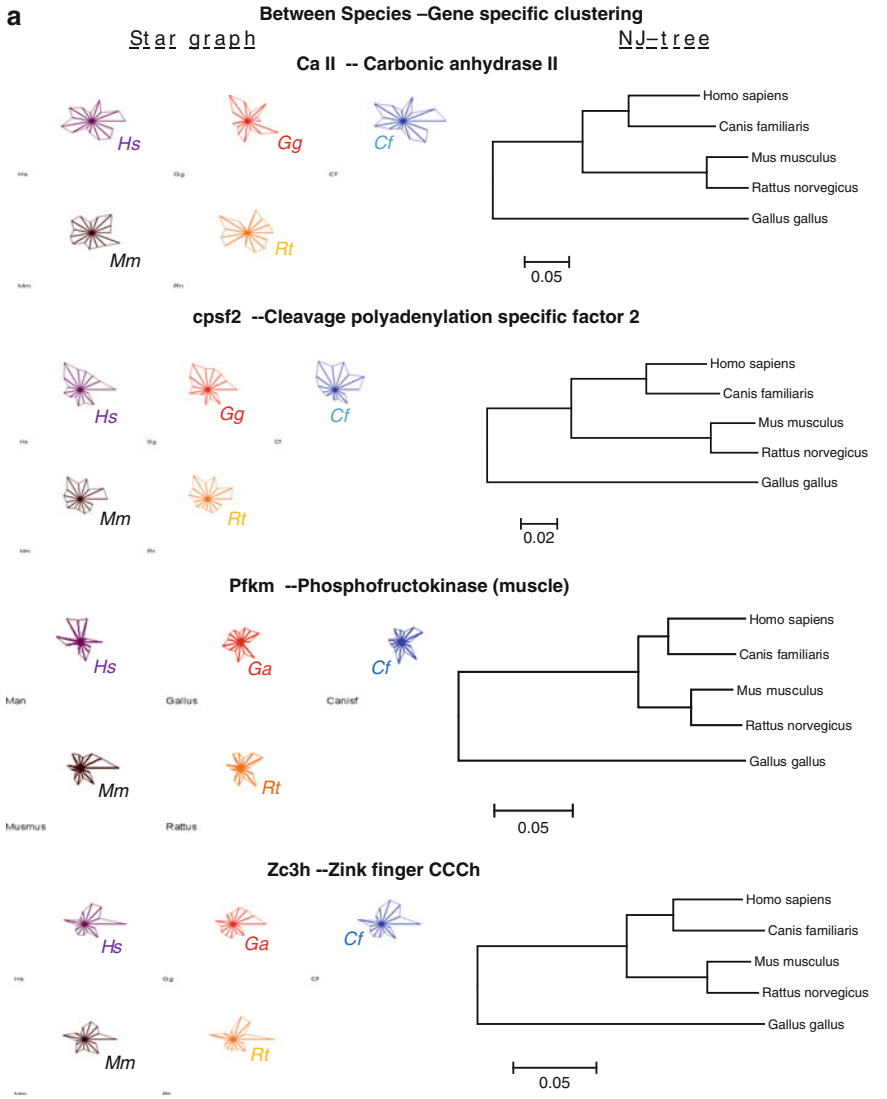


Fig. 15 (**a**) Gene (sequence) specific NJ-tree Clustering and star graphs of (based on trinucleotide word frequency distribution) species—*Homo sapiens* (*Hs*), *Canis familiaris* (*Cf*), *Mum musculus* (*Mm*), *Ratus norvegicus* (*Rn*) and *Gallus gallus* (*Gg*) *for four genes* (*Call*, *cpsf2*, *pfkm and z3ch*). (**b**) Star graphs (based on trinucleotide word frequency distribution) and NJ-tree draw based on (mature) four miRNA sequences of three target genes (z3ch, call and cpsf2) among 4/5 species: *Homo sapiens* (*Hs*), *Canis familiaris* (*Cf*), *Mum musculus* (*Mm*). *Ratus norvegicus* (*Rn*) and *Gallus gallus* (*Ga*)

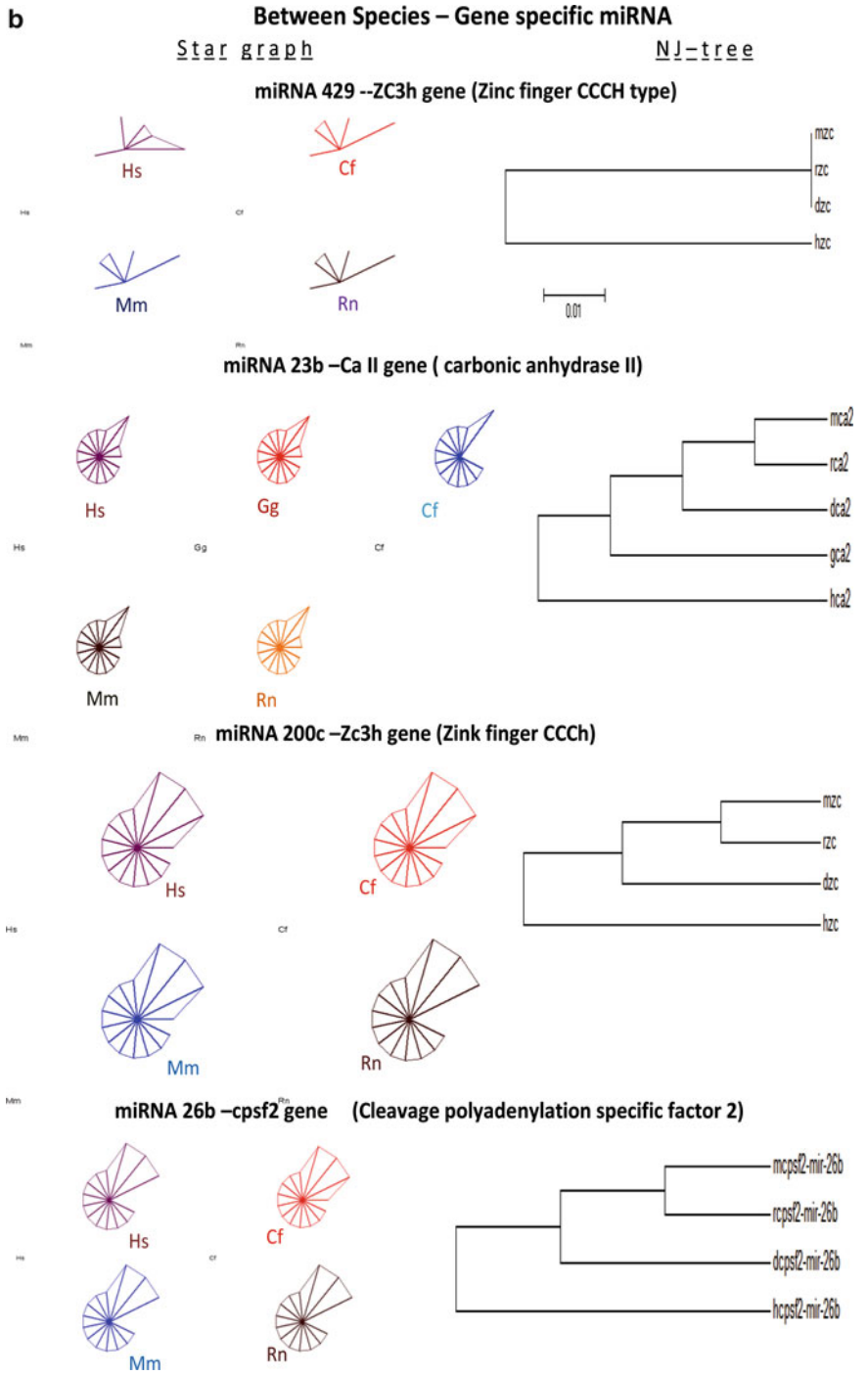


Fig. 15 (Continued)

five species show almost identical pattern indicating the conserved nature of the miRNA across the species. Both Zfc3h and Cpsf2 genes show a single tree with increasing differences across the species in case of nj-clustering. However the star graphs show remarkable shape suggesting stability of miRNA associated with the two genes across the species. Overall, the nj-tree based on gene specific miRNA shows that the phylogeny tree is different from the miRNA tree in case of specific genes.

4.5.3 Comparison of mature microRNA across species

A comparison of sequence lengths of three selected mature miRNAs viz., miR-26b-5p, miRNA-429 and miR-200c-3p that are common to 6 species (obtained from online sources) have been shown in Figs. 16a, 16b, 16c. In case of miR-26b-5p the length of mature miRNA varies between 21 – 22bp: 21 bp length among the three species: *Homo sapiens, Mus musculus and Rattus norvegicus and Macaca mulatta* and is 22bp in case the rest four species (*petromyzon marinus, Cricetulus grseus, Salmo salar and Tupaia chinensis*). In all the eight species compared for the miR-26b-5p the entire length is identical and the extra bp (U), an addition as the 22nd bp in the latter four species. Therefore the eight species shows structural identity from 1-21bp length and are highly conserved, and when compared to man, the four species show one difference – extra (insertion) bp in four species. A comparison of the sequence lengths of mature miR-429 across the six species shows conserved nature. It is identical and a consistent length of 22bp except *Xenopus*, where it differ by one nucleotide – deletion of U at 22nd position. The figure 16c shows comparison of structure of miR-200c-3p across six species. The length varies from 21bp (in Rattus norvegicus), 22bp in two species (Macaca mulatta and Mondephis domestica) to 23bp in three species (Homo sapiens, mus musculus, Tupaia chinensis). The sequence is identical in all the 23bp length in three species (Homo sapiens, mus musculus, Tupaia chinensis) and in all the 21bp and 22bp lengths in two species (Rattus norvegicus and Mondephis domestica) with deletion of 2bp (end bps GA) and 1bp (end bp A) respectively. The species Macaca mulatta shows least identical of the six species compared and differs at 12 positions.

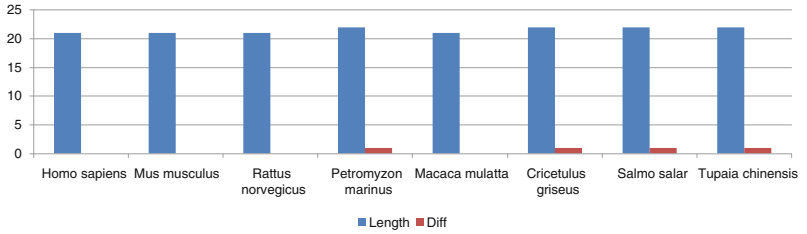
5 Conclusions

A statistical analysis of classification, characterization and conservative nature of miRNA based on the structure and function reveals: PRE-miRNA:1. Length of human precursor miRNAs (pre-miRNA) are found to vary within the range of 43bp-148bp for Redundant data set (R) while for Non-redundant dataset (NR) it varies between 47bp to 140bp, with a maximum number of miRNAs vary a length between 80 to 109bp. The length variation of both the types appears to follow binormal distribution with an antimode between 90bp to 93bp length. 2. The distribution of

a

miR-26b-5p: Across species

	miR-26b-5p	Length	Diff
UUCAAGUAAUUCAGGAUAGGU	Homo sapiens miR-26b	21	
UUCAAGUAAUUCAGGAUAGGU	Mus musculus miR-26b	21	0
UUCAAGUAAUUCAGGAUAGGU	Rattus norvegicus miR-26b	21	0
UUCAAGUAAUCCAGGAUAGGUU	Petromyzon marinus miR-26b	22	1
UUCAAGUAAUUCAGGAUAGGU	Macaca mulatta miR-26b	21	0
UUCAAGUAAUUCAGGAUAGGUU	Cricetulus griseus miR-26b	22	1hamster
UUCAAGUAAUCCAGGAUAGGUU	Salmo salar miR-26b	22	1
UUCAAGUAAUUCAGGAUAGGUU	Tupaia chinensis miR-26b	22	1treeshrew



miR-26b-5p : across species

b

miR-429 Across species

	miR-429	Length	Diff
UAAUACUGUCUGGUAAUGCCG	Xenopus	21	1
UAAUACUGUCUGGUAAAACCGU	human	22	0
UAAUACUGUCUGGUAAUGCCGU	mouse	22	0
UAAUACUGUCUGGUAAUGCCGU	Canis familiaris miR-429	22	0
UAAUACUGUCUGGUAAUGCCGU	Fugu rubripes miR-429	22	0
UAAUACUGUCUGGUAAUGCCGU	Tetraodon nigroviridis miR-429	22	0

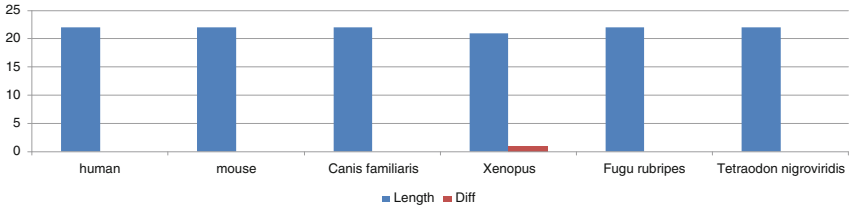


Fig. 16 (**a**, **b**) Comparison of length variation and number of differences (with respect to Homosapiens sapiens) across 8/6 species in case of miR-26-b, miR-429. (**c**) Comparison of length variation and number of differences (with respect to Homosapiens sapiens) across six species in case of miR-420

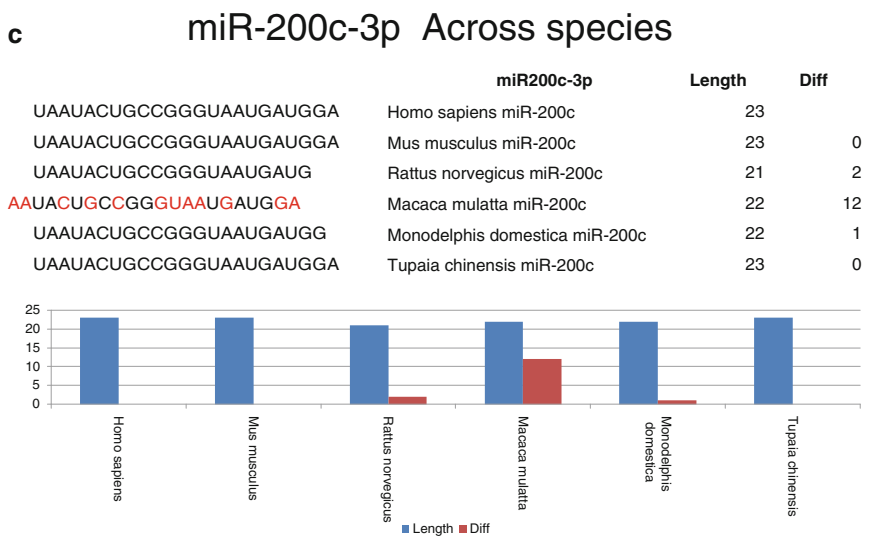


Fig. 16 (Continued)

GC content of the pre-miRNA 96bp length shows binormal distribution with modes at 35-39 and at 50-54, whereas pre-miRNA 109bp length shows a trend of normal distribution with a tendency of antimode at 50-54bp. Both 96bp length and 109 bp length pre-miRNAs types can be further classified into several clusters based on sequence similarity. 3. Classification based on chromosomal location suggests about a maximum of them are located at 19th and X-chromosomes. Mature miRNA: 1. The mature miRNA in Man shows a length variation between 17bp to 27bp with a peak at 22bp and follows a normal distribution. 2. In case of the mature miRNA derived from 96bp pre-miRNA and 109bp pre-miRNAs however show positively and negatively skewed distribution s respectively. The miRNA Function: 1. The has-mir-1 and has-mir-124 are known to be associated with 31 and 141 different functions with varying frequency respectively. 2. The net work relationship between miRNA and its functions reveal that has-mir-124 controls about 201 targets and has-mir-1 influences 101 targets. 3. Whereas VEGF gene (involved in the growth of blood vessels) is targeted by about 36 miRNAs. The miRNA across species: Comparison of the mature miRNA sequences across species and with respect specific genes reveals that the length of mature miRNA and the sequence structure is evolutionarily conserved across species.

References

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW (2005) Sequence variants in SLITRK1 are associated with Tourett's syndrome. Science 310(5746):317–320

Agostini F, Dapas B, Farra R, Grassi M, Racchi G, Klingel K, Kandolf R, Heidenreich O, Mercatanti A, Rainaldi G et al (2006) Potential applications of small interfering RNAs in the cardiovascular field. Drugs Future 31(6):513–525

Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355

Ambros V, Bartel B, Bartel DP, Burge CB et al (2003) A uniform system for microRNA annotation. RNA 9:277–279

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism and function. Cell 116(2): 281–297

Basu S, Burma DP, Chaudhuri P (2003) Words in DNA sequences: some case studies based on their frequency statistics. J Math Biol 46:479–503

Bouamar H, Jiang D, Wang L, Lin AP, Ortega M, Aguiar RC (2015) MicroRNA 155 control of p53 activity is context dependent and mediated by Aicda and Socs1. Mol Cell Biol 35(8): 1329–1340

Brennecke J, Hipfner DR, Stark A, Russel RB, Cohen SM (2003) *Bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell 113:25–36

Brennecke J, Stark A, Russel RB, Cohen SM (2005) Principles of micro-RNA-target recognition. PLoS Biol 3:e85

Brody E, Abelson J (1985) The "spliceosome" yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. Science 228(4702):963–967

Chaudhuri P, Das S (2001) Statistical analysis of large DNA sequences using distribution of DNA words. Curr Sci 80:1161–1166

Chaudhuri P, Das S (2002) SWORDS: a statistical tool for analyzing large DNA sequences. J Biosci 27:1–6

Crick FHC (1958) On protein synthesis. Symp Soc Exp Biol 12:138–163

Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379

Crick FHC (1970) Central dogma of molecular biology. Nature 227:561–563

Crick FHC (1988) What made pursuit. Basic Books, New York

Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the microprocessor complex. Nature 432(7014):231–240

Doolittle WF (1978) Genes in pieces: were they ever together? Nature 272:581–582

Doolittle WF, Fraser P, Gerstein MB, Graveley BR (2013) Sixty years of genome biology. Genome Biol 14(4):113–119

Dusl M, Senderek J, Müller JS, Vogel JG, Pertl A, Stucka R, Lochmüller H, David R, Abicht A (2015) A 3′-UTR mutation creates a microRNA target site in the GFPT1 gene of patients with congenital myasthenic syndrome. Hum Mol Genet 24(8). doi:10.1093/hmg/ddv090

Fantini B (2006) History of central dogma of molecular biology and its epistemological status today, Geneva, February 22–23, 2007. Hist Philos Life Sci 28:487–609

Lian F, Cui Y, Zhou C, Gao K, Wu L (2015) Identification of a plasma four-microRNA panel as potential noninvasive biomarker for osteosarcoma. PLoS One 10(3):e0121499. doi:10.1371/journal.pone.0121499

Franklin RE, Gosling RG (1953) Molecular configuration in sodium thymonucleate. Nature 171(4356):740–741

Gilbert W (1978) Why genes in pieces? Nature 271(5645):501

Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, Shiekhattar R (2004) The microprocessor complex mediates the genesis of microRNAs. Nature 432(7014):235–240

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31:439–441

Griffiths-Jones S et al (2005) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34(Suppl 1):D140–D144

Gu W, Wang X, Zhail C, Xie X, Zhou T (2012) Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol 29:3037–3044

Han J, Han J, Lee Y, Yeom K, Nam J, Heo I, Rhee J, Sohn SY, Cho Y, Zhang BT, Kim VN (2004) The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 18(24):3016–3027

Huang Y, Gu X (2007) A bootstrap based analysis pipeline for efficient classification of phylogenetically related animal miRNAs. BMC Genomics 8:66. doi:10.1186/1471-2164-8-66

Jeffreys AJ, Flavel RA (1977) The rabbit beta-globin gene contains a large large insert in the coding sequence. Cell 12(4):1097–1108

Jiang Q et al (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 37(Database issue):D98–D104

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. PLoS Biol 2:e363. doi:10.1371/journal.pbio.0020363, pmid:15502875

Kamanu TKK, Radovanovic A, Archer JAC, Bajic VB (2013) Exploration of miRNA families for hypotheses generation. Nat Sci Rep 3:2940. doi:10.1038/srep02940

Kefas B, Comeau L, Floyd DH, Seleverstov O, Godlewski J, Schmittgen T et al (2009) The neuronal microRNA miR-326 acts in a feedback loop with notch and has therapeutic potential against brain tumors. J Neurosci 29:15161–15168. doi:10.1523/JNEUROSCI.4966-09.2009, pmid:19955368

Kefas B, Comeau L, Erdle N, Montgomery E, Amos S, Purow B (2010) Pyruvate kinase M2 is a target of the tumor-suppressive microRNA-326 and regulates the survival of glioma cells. Neuro Oncol 12:1102–1112. doi:10.1093/neuonc/noq080, pmid:20667897

Kruger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 34:W451–W454. doi:10.1093/nar/gkl243, pmid:16845047

Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294(5543):858–862

Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. Science 294:862–864

Lee RC, Feinbaum RL, Ambros V (1993) The C elegans heterochronic gene lin-4 encodes small RNAs with artisense complementarity to lin-14. Cell 75:843–854

Lee Y et al (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23: 4051–4060

Lorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Ménard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res 65(16):7065–7070

Ma L, Teruya-Feldstein J, Weinberg RA (2007) Tumour invasion and metastasis initiated bu microRNA-10b in breast cancer. Nature 449(7163):682–688

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940

miRBase (2010) Sanger miRBase sequence database. http://microrna.sanger.ac.uk/sequences/

Mendell JT (2008) myRiad roles for the miR-17-92 cluster in development and disease. Cell 133(2):217–222

miRBase (2010) Sanger miRBase sequence database. microrna.sanger.ac.uk/sequences/

Morange M (2006) The protein side of the central dogma:permanence and change. Hist Philos Life Sci 28:513–524

Morange M (2008) What history tells us XIII. Fifty years of central dogma. J Biosci 33(2):171–175

Mulder C, Arrand JR, Delius H, Keller W, Pettersson U, Roberts RJ, Sharp PA (1975) Cleavage maps of DNA from adenovirus types 2 and 5 by restriction endonucleases EcoRI and HpaI. Cold Spring Harb Symp Quant Biol 39(Pt 1):397–400

Olsen PH, Ambros V (1999) The *lin-4* regulatory RNA controls developmental timing in *C. elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol 2:671–680

Piriyapongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337

Prabhakar S, Noonan JP, Pääbo S, Rubin EM (2007) Accelerated evolution of conserved noncoding sequences in humans. Science 314(5800):786

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in Caenorhabditis elegans. Nature 403:901–906

Rhee J-K, Shin S-Y, Zhang B-T (2013) Construction of microRNA functional families by a mixture model of position weight matrices. Peer J 1:e199. doi:10.7717/peerj.199

Ruvkun G (2001) Molecular biology: glimpses of a tiny RNA world. Science 294:797–799

Sarazin A, Voinnet O (2014) Exploring new models of easiRNA biogenesis. Nat Genet 46(6):530. doi:10.1038/ng.2993

Sharp PA, Sugen B, Sambrook J (1973) Detection of two restriction endonuclease activities

Sinha S, Vasulu TS, Rajat KD (2009) Performance and evaluation of microRNA gene identification tools. J Proteomics Bioinform 2(8):336–343

Smielewska MM (2008) The role of miRNAs and PiRNAs in planarian regeneration. UMI, Ann Arbor

Soifer H et al (2007) MicroRNAs in disease and potential therapeutic applications. Mol Ther 15:2070–2079

Chung SH, Gillies M, Sugiyama Y, Zhu L, Lee S-R, Shen W (2015) Profiling of microRNAs involved in retinal degeneration caused by selective Müller cell ablation. PLoS One 10(3):e0118949

van Rooij E et al (2007) MicroRNAs: powerful new regulators of heart disease and provocative therapeutic targets. J Clin Invest 117:2369–2376

Vergoulis T et al (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Res 40:D222–D229

Vergoulis T, Kanellos I, Kostoulas N, Georgakilas G, Sellis T, Hatzigeorgiou A, Dalamagas T (2015) mirPub: a databse for searching microRNA publications. Bioinformatics 31(2):1–3

Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Lorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM (2006) A microRNA expression signature of human solid tumor cancer gene targets. Proc Natl Acad Sci U S A 103(7):2257–2261

Watson JD (1965) Molecular biology of the gene. W A Benjamin, New York

Watson JD, Crick FH (1953) Molecular structure of nucleic acids: a structure for deoxy rebose nucleic acid. Nature 153:737–738

Wightman B, Ha I, Ruvkun G (1993) Posttransacriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 75(5):855–862

Wilkins MHF, Stokes AR, Wilson HR (1953) Molecular structure of deoxypentose nucleic acids. Nature 171(4356):738–740

Woese CR (1967) The genetic code: the molecular basis for genetic expression. Harper and Row, New York

Woese CR (2001) Translation: in retrospect and prospect. RNA 7:1055–1067

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37(Database issue):D105–D110. doi:10.1093/nar/gkn851, Epub 2008 Nov 7

Xie B et al (2013) MirCancer: a microRNA-cancer association database constructed by text mining literature. Bioinformatics 29:638–644

Xu P, Vernooy SY, Guo M, Hay BA (2003) The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. Curr Biol 13(9):790–795

Yu X, Lin J, Zack DJ et al (2008) Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. Nucleic Acids Res 36(20):6494–6503

Yu X, Lin J, Zack DJ et al (2008) Analysis of regulatory network topology reveals functionally distinct classes of microRNAs. Nucleic Acids Res 36(20):6494–503

Zamore PD (2002) Ancient pathways programmed by small RNAs. Science 296:1265–1269

Zeng Y, Wagner EJ, Cullen BR (2002) Both natural and designed microRNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell 9:1327–1333

Zhang H-M, Kuang S, Xiong X, Gao T, Liu C, Guo A-Y (2015) Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. Brief Bioinform 16:45–58

Zhang Y, Li M, Wang H, Fisher WE, Lin PH, Yao Q et al (2009) Profiling of 95 microRNAs in pancreatic cancer cell lines and surgical specimens by real-time PCR analysis. World J Surg 33:698–709. doi:10.1007/s00268-008-9833-0, pmid:19030927

Zhang Z-L, Bai Z-H, Wang X-B, Bai L, Miao F, Pei H-H (2015) miR-186 and 326 predict the prognosis of pancreatic ductal adenocarcinoma and affect the proliferation and migration of cancer cells. PLoS One 10(3). doi:10.1371/journal.pone.0118814

Longitudinal Growth of Elephant Foot Yam and Some Characterisation Theorems

Ratan Dasgupta

Abstract We estimate growth curve of yam in a longitudinal study with 60 plants comprising of three groups, each group having twenty plants corresponding to 500, 650 and 800 g of seed weight. The study is relevant for appropriate choice of initial weight for plantation and harvest time of yam. Longitudinal growths of Elephant-foot-yam are studied by taking yam from the ground with care in the middle of a production season, then measure underground growth by Archimedean principle and replant the structure for further growth in remaining lifetime. Yam growth curve has a spike and takes a sharp upturn towards the end. Harvest of the crop at mature stage of plant lifetime increases yam yield substantially, as seen in sharp increase of growth curve towards end. Growth slopes before and after intervention while taking interim observation and difference of these two slopes for each plant are considered. These yam characteristics are linear combinations of yield observations, and are seen to follow normal distribution in quantile–quantile plot; the finding has implications in testing of hypothesis concerning slopes. Experimental results obtained in the present study from 60 plants supplement and reconfirm earlier findings of Dasgupta (Growth curve and structural equation modeling, Chap. 1, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York, 2015) on yam growth based on six plants, two plants per above-mentioned seed weights. Growth curve corresponding to seed weight 650 g indicates superior yield when crops are harvested at the end of season. Growth curve variation in each group of seed weight in parametric and nonparametric set-up is also studied. Seed weight 650 g corresponds to growth curve with minimal variation. Above ground biomass of yam, in the middle of crop season, is higher for seed weight 800 g than that for 500 g. Prediction problem of underground yam weight on the basis of observable above ground biomass and plant lifetime is discussed. Accuracy of prediction is high in logarithmic scale, suggesting that the relationship is nonlinear. Growth of plant lifetime that has applications in crop harvest rate and forecasting market supply of yam is approximated by a Weibull model. For some specific choice of parameters, hazard rate of Weibull distribution is shown to be the limiting form

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_14

of that for folded normal variable, characterising the limiting distribution. An error bound of hazard rate approximation is obtained. Discrete versions of the folded normal distribution and Weibull distribution are also discussed.

Keywords Amorphophallus paeoniifolius • Elephant foot yam • Longitudinal study • Cross sectional study • Archimedean principle • Quantile • Wilcoxon 2-sample U statistics • Ornstein–Uhlenbeck process • Burr • Hazard rate • Weibull distribution • Folded normal distribution

MSC classification 2010: Primary 62P10; Secondary 62J02

1 Introduction

We study longitudinal growth curve of Elephant foot yam (*Amorphophallus paeoniifolius*) over plant lifetime for different seed weights, with applications to selection of initial seed weight and harvest time. Growth curve with less variation with respect to the choice of seed weight is also of interest. This helps the experimenter to select appropriate seed weight with ascertained yield during the growth period.

This is a confirmatory follow-up study with a relatively large number of plants compared to an experiment with six plants considered in Dasgupta (2015), where seed weight 650 g turned out to be the recommended weight for cut yam corm. Plant sensitivity to inadvertently induced hurt is also examined therein.

Sixty plants consisting of three groups, each group having twenty plants corresponding to 500, 650 and 800 g of seed weights are considered in the experiment. Cut yam of above-mentioned weights as seed corm were planted on 2 April 2014 at the Indian Statistical Institute, Giridih Farm, Jharkhand.

In the middle of production season plants were uprooted with care, volume of underground yam attached to plant is measured by Archimedean principle and an indirect estimate of underground yam weight is obtained by multiplying volume with yam density, following a procedure adopted in Dasgupta (2015). We also record the total weight of uprooted plant. Relationship of above ground biomass with underground yam is also studied.

We observe presence of a spike towards end of yam growth curve; see also Dasgupta (2013a), Dasgupta (2015).

Yam plant lifetime may extend approximately to 7 months. Harvesting the mature crop provides approximately five times the initial weight in Giridih farm. *The yam yield is doubled in about 75 days when harvested early by farmers for monetary reasons.* Tender yam shoots and stems from early harvest also have a market value. *Above ground biomass of yam is significantly higher for seed weight 800 g than that for 500 g, in the middle of the crop season, when compared by Wilcoxon 2-sample U statistics.*

Growth slopes before and after intervention for taking interim readings of yam, and difference of growth slopes may be expressed as linear combinations of observations. Distributions of these characteristics are of interest in testing of hypothesis on growth pattern. *Analysed data indicates that the growth slopes and differences of slopes are normally distributed.*

Variation of the growth curves in each group of seed weight is also of interest, less variation in curve makes the corresponding seed weight preferable. *Seed weight 650 g has less residual variation in yam growth.*

The problem of predicting underground yam weight based on above ground biomass and plant lifetime is also studied. *Accuracy of yam yield prediction over time is higher in log scale compared to usual scale, suggesting that the relationship is nonlinear.*

Yam plant lifetime may be well approximated by Weibull distribution, see e.g. Dasgupta (2014). This is of interest in studying crop maturity, and subsequent market supply prediction. We prove a characterisation result of Weibull distribution in terms of limiting hazard rate of folded normal distribution. Discrete version of these distributions are of similar properties.

In the next section we describe the experimental layout, longitudinal data on yam, analysis of data and the results. In Sect. 3 we show that Weibull model with some specific choice of parameters is the only distribution that has the limiting hazard rate of a folded normal variable $Y = |X|$, where $X \sim N(0, \sigma^2)$. The distribution has applications in contexts where signs of characteristics are ignored. Error bound in hazard rate approximation is obtained. Discrete versions of the distributions are considered. Proximity of these distributions with Weibull model is of interest in survival analysis and plant lifetime modeling. Discussion of the results and conclusions are stated in Sect. 4.

2 Experimental Layout, Analysis of Yam Data and the Results

The experimental layout consists of six columns, in each columns there are ten equidistant pits at a distance of 1 m. First two columns are for seed weight 500 g, next two are for seed weight 650 g, and the last two columns are for plants with seed weight 800 g. Column to column distance is also 1 m. Little bit of organic manure like cow dung was given in the pits while planting the fungicide treated cut yam seed corms of specified weights at the start of the experiment on 2 April 2014 in Giridih farm of the Indian Statistical Institute. Little amount of vermicompost was further added at the time of replanting after taking interim observations on yam.

Table 1 provides 60 plant characteristics recorded in the growth experiments viz., initial yam weight, weight during intermediate lifetime, above ground biomass, final weight, etc.

Table 1 Growth data on yam

Plant no.	Seed wt (kg)	Sprouting date	Interim obs. date	Interim yam wt (kg)	Above ground biomass (kg)	Date till plant alive	Final wt (kg)	Plant life (day)
1	0.50	27-05-14	04-09-14	0.445068	1.205932	02-11-14	0.812	155
2	0.50	16-05-14	04-09-14	1.48356	3.84444	06-12-14	2.736	200
3	0.50	19-05-14	04-09-14	0.935784	2.711216	06-12-14	1.761	197
4	0.50	20-05-14	20-09-14	2.270988	3.481012	21-11-14	2.926	181
5	0.50	19-05-14	20-09-14	2.202516	4.056484	28-11-14	3.045	189
6	0.50	19-05-14	20-09-14	1.791684	3.087316	28-11-14	2.593	189
7	0.50	15-05-14	20-09-14	2.16828	3.70972	28-11-14	2.976	193
8	0.50	29-05-14	20-09-14	0.91296	2.68304	28-11-14	1.994	179
9	0.50	26-05-14	20-09-14	1.392264	2.267736	21-11-14	2.237	175
10	0.50	27-05-14	20-09-14	0.524952	1.760048	21-11-14	1.294	174
11	0.50	02-05-14	04-09-14	2.693232	3.253768	18-09-14	2.964	136
12	0.50	07-05-14	04-09-14	0.79884	2.76516	12-11-14	1.595	185
13	0.50	12-05-14	04-09-14	2.51064	4.18836	28-11-14	3.402	196
14	0.50	25-05-14	04-09-14	1.266732	2.582268	06-12-14	2.429	191
15	0.50	16-05-14	20-09-14	2.008512	2.681488	05-11-14	2.993	169
16	0.50	28-05-14	20-09-14	1.586268	3.480732	28-11-14	2.913	180
17	0.50	15-05-14	20-09-14	1.962864	3.767136	28-11-14	3.086	193
18	0.50	19-05-14	20-09-14	1.837332	3.787668	12-11-14	2.577	173
19	0.50	31-05-14	20-09-14	0.239652	1.150348	28-11-14	0.742	178
20	0.50	28-05-14	20-09-14	1.19826	2.44874	05-11-14	1.061	157
21	0.65	19-05-14	04-09-14	1.380852	3.537148	14-10-14	1.874	145
22	0.65	23-05-14	04-09-14	1.59768	3.31532	14-10-14	2.623	141
23	0.65	19-05-14	04-09-14	1.380852	1.342148	28-11-14	2.354	189
24	0.65	29-05-14	20-09-14	0.79884	1.73316	06-12-14	1.578	187
25	0.65	28-05-14	20-09-14	2.373696	4.485304	28-11-14	3.37	180
26	0.65	17-06-14	20-09-14	0.821664	2.193336	06-12-14	1.495	169
27	0.65	26-04-14	20-09-14	2.51064	5.15036	28-11-14	4.158	212
28	0.65	26-05-14	20-09-14	0.924372	3.246628	21-11-14	2.196	175
29	0.65	15-05-14	20-09-14	2.05416	3.85484	12-11-14	3.185	177
30	0.65	26-05-14	20-09-14	1.48356	3.20544	28-11-14	2.408	182
31	0.65	05-05-14	04-09-14	2.179692	1.820308	18-09-14	2.399	133
32	0.65	15-05-14	04-09-14	2.807352	5.049648	28-11-14	3.896	193
33	0.65	06-05-14	04-09-14	2.79594	4.47506	25-10-14	3.569	169
34	0.65	02-05-14	04-09-14	2.213928	3.505072	21-11-14	3.301	199
35	0.65	25-04-14	21-09-14	2.45358	1.78642	14-10-14	2.932	169
36	0.65	28-05-14	21-09-14	2.122632	5.383368	21-11-14	3.698	173
37	0.65	27-05-14	21-09-14	1.563444	4.624556	28-11-14	2.796	181
38	0.65	27-04-14	21-09-14	3.446424	4.966576	21-11-14	4.105	204
39	0.65	27-05-14	21-09-14	1.48356	3.49744	21-11-14	2.235	174

(continued)

Table 1 (continued)

Plant no.	Seed wt (kg)	Sprouting date	Interim obs. date	Interim yam wt (kg)	Above ground biomass (kg)	Date till plant alive	Final wt (kg)	Plant life (day)
40	0.65	16-05-14	21-09-14	1.038492	1.994508	05-11-14	1.611	169
41	0.80	11-06-14	04-09-14	1.928628	5.126372	06-12-14	2.818	175
42	0.80	19-05-14	04-09-14	2.544876	5.706124	25-10-14	3.381	156
43	0.80	16-05-14	04-09-14	2.407932	3.933068	14-10-14	3.368	148
44	0.80	19-05-14	21-09-14	1.688976	2.614024	14-10-14	2.059	145
45	0.80	07-05-14	21-09-14	3.1383	4.4257	14-10-14	4.116	157
46	0.80	24-04-14	21-09-14	1.643328	3.118672	14-10-14	2.383	170
47	0.80	29-05-14	21-09-14	2.704644	5.968356	28-11-14	3.355	179
48	0.80	25-05-14	21-09-14	1.723212	2.627788	14-10-14	2.327	139
49	0.80	12-05-14	21-09-14	2.487816	4.794184	28-11-14	3.508	196
50	0.80	24-04-14	21-09-14	3.663252	3.463748	14-10-14	4.431	170
51	0.80	27-04-14	04-09-14	3.925728	4.634272	14-10-14	4.603	167
52	0.80	15-05-14	04-09-14	1.186848	1.825152	25-10-14	1.546	160
53	0.80	16-05-14	04-09-14	1.015668	3.632332	21-11-14	2.012	185
54	0.80	01-06-14	04-09-14	1.65474	4.25526	28-11-14	2.716	177
55	0.80	27-05-14	21-09-14	2.213928	3.634072	21-11-14	3.048	174
56	0.80	30-05-14	21-09-14	2.39652	4.01748	14-10-14	3.003	134
57	0.80	27-04-14	21-09-14	2.5677	4.0933	28-11-14	2.987	211
58	0.80	27-05-14	21-09-14	2.05416	4.44484	25-10-14	2.807	148
59	0.80	28-04-14	21-09-14	1.974276	3.139724	25-10-14	2.84	177
60	0.80	19-05-14	21-09-14	2.62476	3.88824	12-11-14	3.483	173

In Fig. 1, individual growth curves joining the yam weights by straight lines for each of the 60 yam plants are shown. Upward movement in growth is generally seen in individual curves. Circular points in red colour are median of y values over individual growth curves for grid spacing of 1 day for time on x axis. These provide a robust estimate of response curve. *Indication of steep growth towards end is seen.* The picture is messy with curves from many plants. Next, partition of the curves in groups is made.

Consider the plants with seed weight of 500 g each; Fig. 2 shows the individual growth curves of twenty such plants. The response curve is shown in red colour, the median of y values on grid spacing are joined by straight lines. Plant number 11 and 13 have same trajectory up to an extent as seen in the uppermost curve.

In a similar manner we obtain Fig. 3 that represents plants with seed weight 650 g. The response curve shown here in red is superior than the corresponding curve for seed weight 500 g; elevation in the response curve is prominent in Fig. 3 for 650 g compared to that in Fig. 2 for 500 g.

Figure 4 corresponds to the growth trajectories of 20 yams having seed weight 800 g, the picture shows an upper trend of growth. However, the response curve

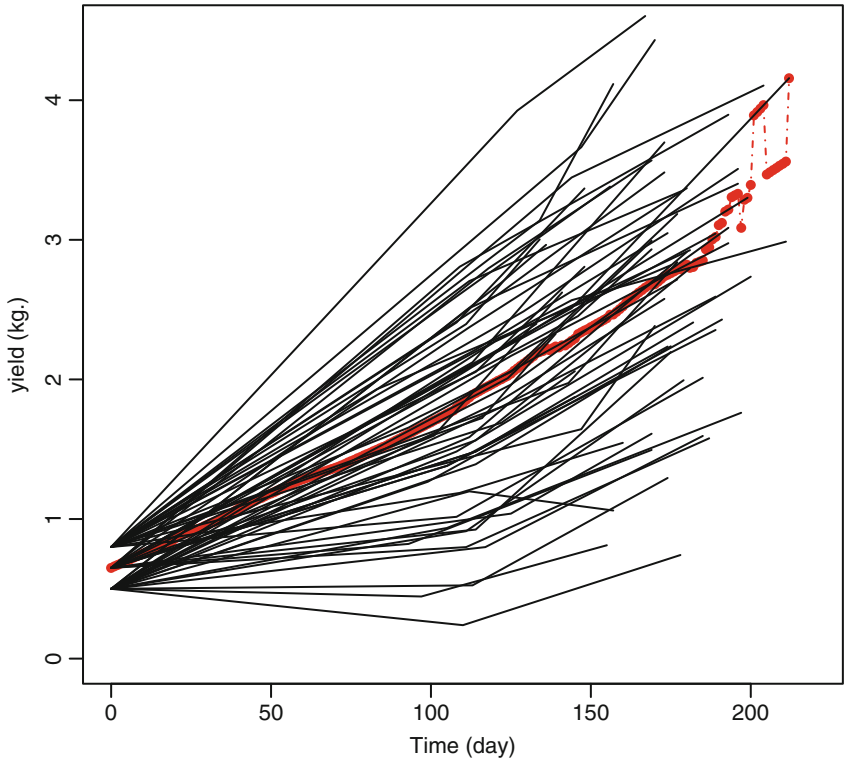


Fig. 1 Individual growth of yam and median response curve

shown in red colour in this case does not seem to be superior than that for 650 g of seed weight.

By nonparametric lowess regression, see, e.g., Cleveland (1981), on median values plotted in red colour in Figs. 2, 3, and 4, we estimate the three growth curves in Fig. 5 corresponding to three different seed weights. We also compute the overall response curve from the median values denoted by red points in Fig. 1, representing the 60 plants combining all seed weights. The lowermost curve in Fig. 5 represents the response for 500 g of seed weight by lowess technique with $f = 0.18$. Index f represents the fraction of observations used by lowess regression at a particular point. The next prominent curve in blue is response curve for seed weight 650 g, obtained by lowess technique with $f = 0.18$. The dot-dash curve merged with it in the beginning, and separated towards end represents the overall response curve, obtained by lowess with $f = 0.18$. The uppermost curve from start in black colour in Fig. 5 corresponds to the seed weight 800 g. This is obtained by lowess regression with $f = 0.2$. *The black curve in Fig. 5 corresponding to seed weight 800 g, although starts from a higher value, has comparatively lower rate of growth and crosses the two curves below it slightly after 150 days of plant lifetime.*

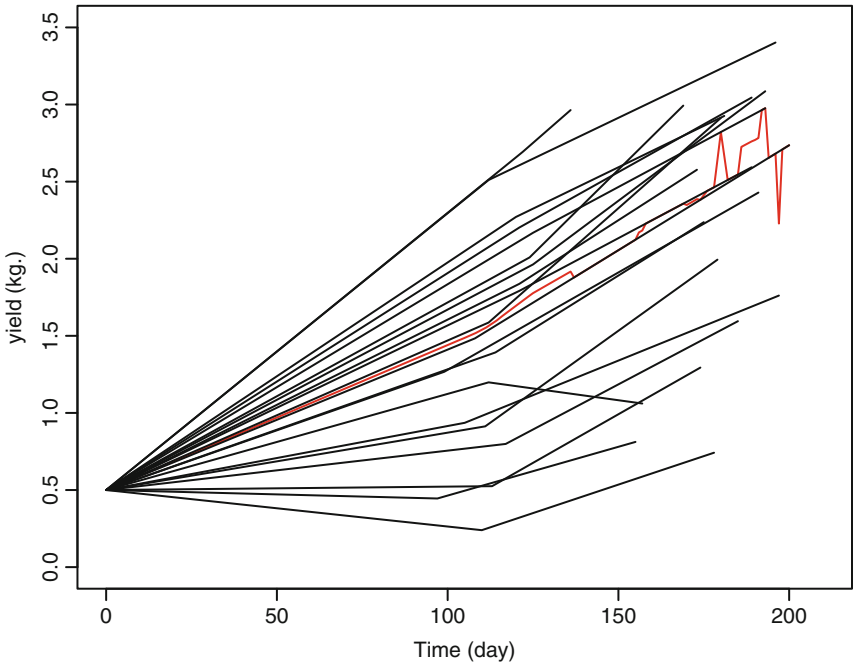


Fig. 2 Individual growth of yam and median response for seed wt. 500 g

Farmers sometime opt for early harvest due to monetary reasons. *Figure 5 indicates that if the yam harvest time is less than 5 months, then a higher seed weight 800 g is preferred compared to 650 or 500 g. However, this may not be the case if a farmer prefers to wait till the crop matures. From Fig. 5, it appears that the growth curve corresponding to seed weight 650 g is superior.* This confirms earlier findings stated in Dasgupta (2015), to recommend 650 g as seed weight of yam to farmers of Giridih, Jharkhand.

Growth slopes before and after the intervention, when yams are taken out of the ground, are estimated for each plant, see also Dasgupta (2015). The slopes and the difference of second slope from first slope for each plant are shown in Table 2.

The normal quantile plots for the first slopes, categorised by date of intervention, and next combined for all dates are shown in Fig. 6. The coefficient of determination is exceptionally high in each of the cases, indicating a strong possibility of normal distribution for these characteristics in hypothesis testing problems.

Figure 7 describes the second slope in normal quantile plot classified according to the date of intervention, and then combined for all dates. Like the previous figure, the coefficients of determinations r^2 are high in each of the cases, indicating a possibility of normal distribution. Presence of a few outliers is also observed. The values of r^2 are slightly lower than previous figure.

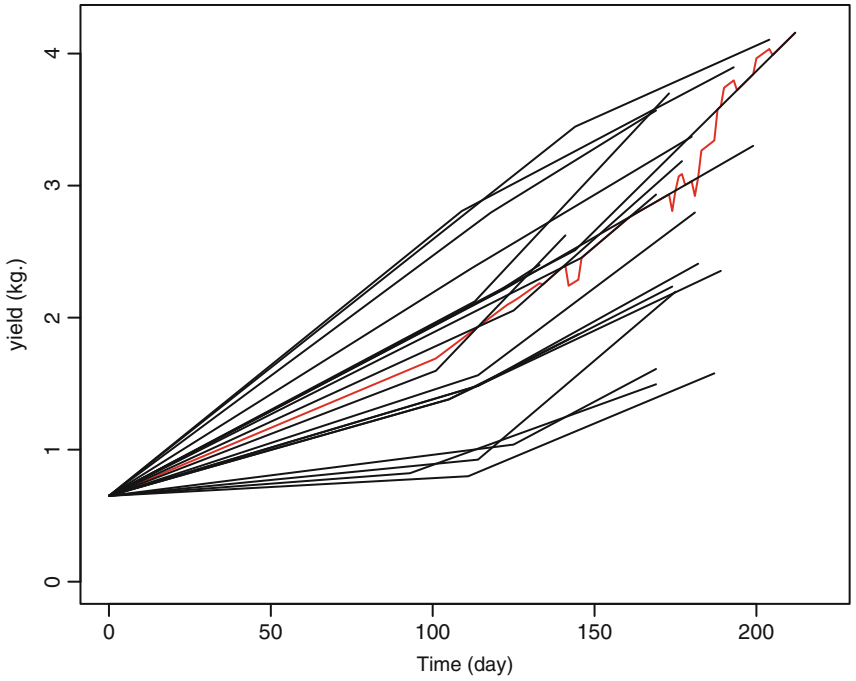


Fig. 3 Individual growth of yam and median curve for seed wt. 650 g

Difference of slopes i.e., first slope minus second slope are shown in Fig. 8 with normal quantile plot, categorised with respect to date of intervention. Figure 8a–c, categorised with respect to date of intervention, and Figure 8d (for combined dates, with $r^2 = 0.9903433$) indicate normal distribution for slope differences as well.

In Dasgupta (2015) fluctuations in growth curves are modeled by Ornstein–Uhlenbeck (O–U) process $V(s)$. An *almost sure* type estimate $\tilde{\sigma}_v = \max_{0 \le s \le t} \mid \hat{V}(s) \mid / \sqrt{2 \log t}$ of the asymptotic standard deviation of the process is available from the maximum fluctuation of observed process $\hat{V}(s)$ on time segment $[0, t]$, around mean response curve. Figure 9 shows the maximum fluctuation (in absolute value) of the curves in Figs. 1, 2, 3, and 4 over grid spacing of 1 day on X axis for time i.e., for each day Fig. 9 plots the maximum variation in Y axis among the growth curves shown in Figs. 1, 2, 3, and 4 for various seed weights. The time range is almost same for plants corresponding to different seed weights viz., $t = 201, 213, 212$ days for seed weights 500, 650, 800 g, respectively; In Fig. 9 the peak of the lower three curves viz., 2.532274, 2.340403, 2.938477 is conservative estimate of $\sqrt{2 \log t}$ times the corresponding asymptotic standard deviation of the underlying O–U process, $t \approx 200$. None of the three curves with different seed weights uniformly dominates the other for all values of time. *In Fig. 9 the curve corresponding to seed weight 650 g lies below the curves of other seed weights for a larger time region and have lower magnitude of peak, it is apparent that the*

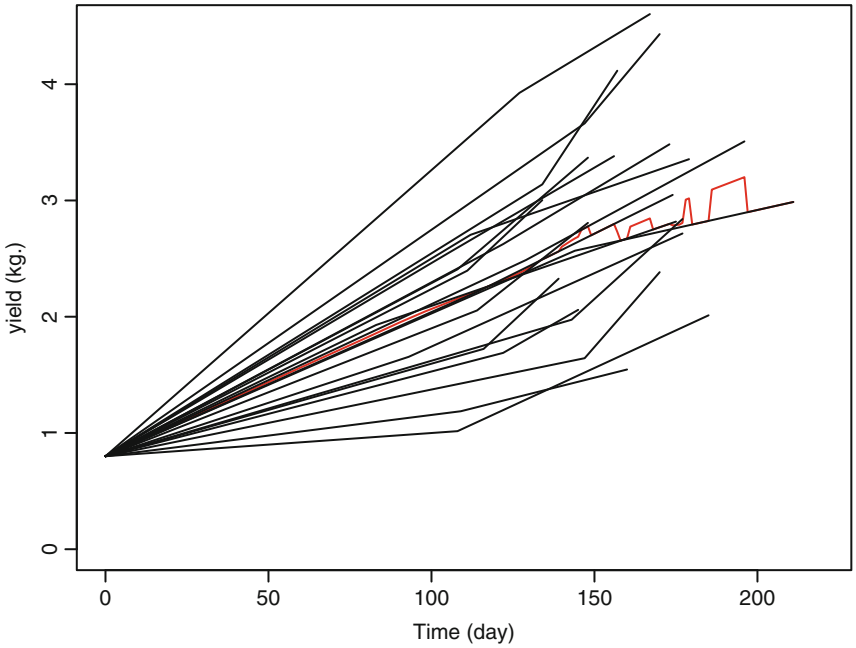


Fig. 4 Individual growth of yam and median response curve for seed wt. 800 g

seed weight 650 g is superior in terms of lowering variation in growth curve as well. These spread index of process fluctuations are nonparametric in nature without model assumptions, as these are based on maximum fluctuation of curves.

The topmost curve in Fig. 9 corresponds to fluctuation of the curves shown in Fig. 1 i.e., for all seed weights combined. As such, this curve has the highest fluctuation of magnitude 3.942262, the topmost curve starts from $(0.8 - 0.5) = 0.3$ kg for y value, the maximum difference between seed weights.

Yet another nonparametric method for comparison of fluctuations around the response curve is available by comparing the magnitudes of fluctuations on n grid points of time, $n_1 = 201, n_2 = 213, n_3 = 212$ for seed weight 500 g, 650 g, and 800 g, respectively. For curves corresponding to seed weights 500 and 650 g, probabilistic ordering of absolute values of residual variables u, v can be quantified in terms of Wilcoxon two sample U statistic of the form $U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(|u_i| > |v_j|)$. While comparing growth curve fluctuations, or absolute deviations $d = d_{500}, d_{650}$ for seed weight 500 g with that for 650 g over time grids, a point estimate of $P(d_{500} > d_{650})$ is $\hat{P}(d_{500} > d_{650}) = U/(n_1 n_2) = 25344/42813 = 0.5919697$. The value is significantly different from 0.5 as the standardised value of U is $U^* = (U - \frac{n_1 n_2}{2})/\{\frac{n_1 n_2 (n_1+n_2+1)}{12}\}^{1/2} = 3937.5/1216.8058 = 3.236$. When compared with a normal deviate, one sided p-value of significance is 0.0006.

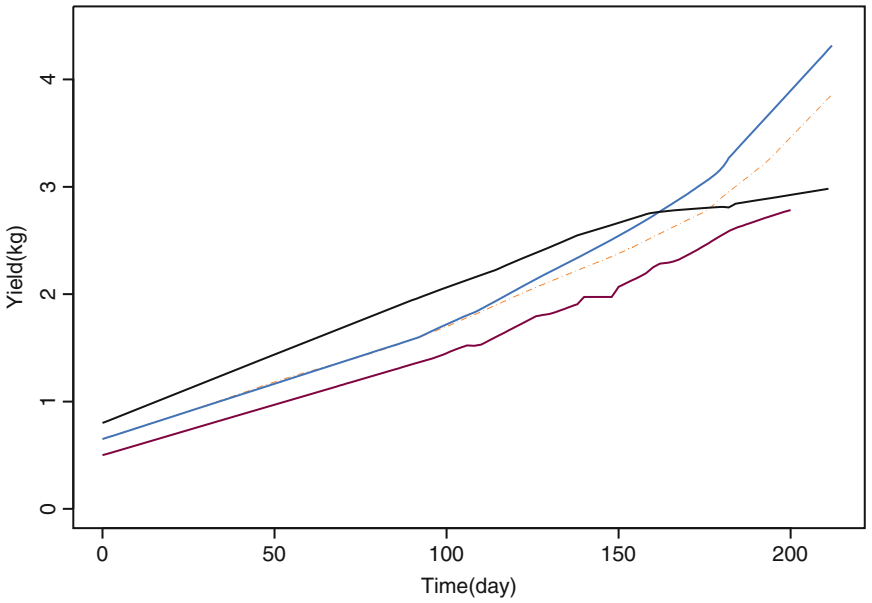


Fig. 5 Growth curve of yam. By nonparametric lowess regression on median values plotted in red color in Figures 2–4, we estimate the three growth curves in Figure 5 corresponding to three different seed weights. The lowermost curve in Figure 5 represents the response for 500 g of seed weight by lowess technique with $f = 0.18$. The next prominent curve in blue is response curve for seed weight 650 g, obtained by lowess technique with $f = 0.18$. We compute the overall response curve from the median values denoted by red points in Figure 1, representing the 60 plants combining all seed weights. The dot-dash curve merged with the blue curve in the beginning, and then gradually separated towards the end represents the overall response curve, obtained by lowess with $f = 0.18$. The uppermost curve from start in black color in Figure 5, corresponds to the seed weight 800 g. This is obtained by lowess regression with $f = 0.2$. *The black curve in Figure 5 corresponding to seed weight 800 g, although starts from a higher value, has comparatively lower rate of growth and crosses the two curves below it slightly after 150 days of plant lifetime. Figure 5 indicates that if the yam harvest time is less than 5 months, then a higher seed weight 800 g is preferred compared to 650 g or 500 g. However, this may not be the case if a farmer prefers to wait till the crop matures. From Figure 5, it appears that the growth curve corresponding to seed weight 650 g is superior*. This confirms earlier findings stated in Dasgupta (2015), to recommend 650 g as seed weight of yam in Giridih, Jharkhand

Similarly, to compare d_{800} with d_{650} we compute $\hat{P}(d_{800} > d_{650}) = U/(n_3 n_2) = 25290/45156 = 0.5600585$. The standardised value statistic in this case is $U^* = 2712/1266.1114 = 2.142$. The value is significant, one sided p-value is 0.0161.

Next, to compare d_{800} with d_{500} we compute $\hat{P}(d_{800} > d_{500}) = U/(n_3 n_1) = 21719/42612 = 0.5096921$. The standardised value of U in this case is $U^* = 413/1211.0173 = 0.341$. The value is insignificant when compared to normal deviate.

Thus fluctuation of growth curve corresponding to seed weight 650 g is significantly smaller compared to those for other two seed weights considered.

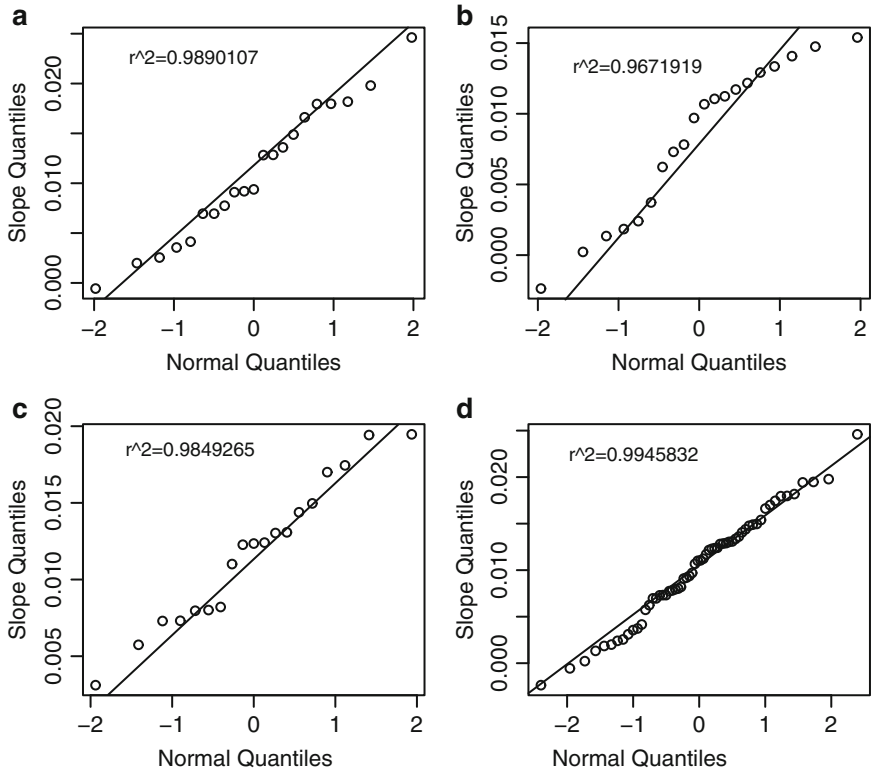


Fig. 6 Normal Quantile plot. (**a**) First slope (4/9/14); (**b**) first slope (20/9/14); (**c**) first slope (21/9/14); (**d**) first slope (combined)

Next consider the problem of predicting underground yam weight y on the basis of above ground biomass x. Figure 10 plots linear regression of these two variables in usual scale. Coefficient of determination $r^2 = 0.6816$ is high for seed weight 500 g. The regression line being $y = -0.4385 + 0.6620x$.

For seed weight 650 g and 800 g, the value of r^2 is 0.3594 and 0.2153, respectively.

For all seed weight combined $r^2 = 0.445$ and the estimated regression line is $y = 0.30879 + 0.45646x$.

Accuracy of regression is improved in logarithmic scale. Figure 11 plots linear regression of two variables x and y in log scale. Coefficient of determination $r^2 = 0.8073$ is higher for seed weight 500 g. The regression line being $\log y = -1.3388 + 1.5602 \log x$.

For seed weight 650 g and 800 g, the value of r^2 in log scale is 0.301 and 0.3026, respectively.

For all seed weight combined $r^2 = 0.5294$ in log scale and the estimated regression line is $\log y = -0.6430 + 0.9933 \log x$.

Table 2 Growth slopes and slope difference for 60 yams

Plant no.	First slope	Second slope	Slope difference
1	−0.0005663093	0.0063264140	−0.0068927233
2	0.0091070370	0.0136134780	−0.0045064410
3	0.0041503238	0.0089697390	−0.0048194152
4	0.0147582333	0.0107379020	0.0040203313
5	0.0140703802	0.0123894710	0.0016809092
6	0.0106750744	0.0117840590	−0.0011089846
7	0.0133462400	0.0118782350	0.0014680050
8	0.0037203604	0.0158976470	−0.0121772866
9	0.0078268772	0.0138481310	−0.0060212538
10	0.0002208142	0.0126073440	−0.0123865298
11	0.0179773115	0.0193405710	−0.0013632595
12	0.0025541880	0.0117082350	−0.0091540470
13	0.0179521429	0.0106114290	0.0073407139
14	0.0077447677	0.0126333480	−0.0048885803
15	0.0121654194	0.0218775110	−0.0097120916
16	0.0096988214	0.0195107650	−0.0098119436
17	0.0117029120	0.0165167060	−0.0048137940
18	0.0110523306	0.0142243850	−0.0031720544
19	−0.0023668000	0.0073874710	−0.0097542710
20	0.0062344643	−0.0030502220	0.0092846863
21	0.0069604952	0.0123287000	−0.0053682048
22	0.0093829703	0.0256330000	−0.0162500297
23	0.0069604952	0.0115850950	−0.0046245998
24	0.0013409009	0.0102521050	−0.0089112041
25	0.0153901429	0.0146515290	0.0007386139
26	0.0018458495	0.0088596840	−0.0070138345
27	0.0129211111	0.0242258820	−0.0113047709
28	0.0024067719	0.0208463610	−0.0184395891
29	0.0112332800	0.0217469230	−0.0105136430
30	0.0073119298	0.0135947060	−0.0062827762
31	0.0128545546	0.0156648570	−0.0028103024
32	0.0197922202	0.0129600950	0.0068321252
33	0.0181859322	0.0151580390	0.0030278932
34	0.0128190820	0.0141178180	−0.0012987360
35	0.0123532877	0.0208008700	−0.0084475823
36	0.0130321416	0.0262561330	−0.0132239914
37	0.0080126667	0.0183963580	−0.0103836913
38	0.0194196111	0.0109762670	0.0084433441
39	0.0073119298	0.0125240000	−0.0052120702
40	0.0031079360	0.0130115450	−0.0099036090
41	0.0135979277	0.0096670870	0.0039308407
42	0.0166178667	0.0163945880	0.0002232787

(continued)

Table 2 (continued)

Plant no.	First slope	Second slope	Slope difference
43	0.0148882593	0.0240017000	−0.0091134407
44	0.0072866885	0.0160880000	−0.0088013115
45	0.0174500000	0.0425086960	−0.0250586960
46	0.0057369252	0.0321596520	−0.0264227268
47	0.0170057500	0.0097068060	0.0072989440
48	0.0079587241	0.0262516520	−0.0182929279
49	0.0130838450	0.0152266270	−0.0021427820
50	0.0194779048	0.0333803480	−0.0139024432
51	0.0246120315	0.0169318000	0.0076802315
52	0.0035490642	0.0070421960	−0.0034931318
53	0.0019969259	0.0129393770	−0.0109424511
54	0.0091907527	0.0126340480	−0.0034432953
55	0.0124028772	0.0139012000	−0.0014983228
56	0.0143830631	0.0263686960	−0.0119856329
57	0.0122756944	0.0062582090	0.0060174854
58	0.0110014035	0.0221423530	−0.0111409495
59	0.0082117203	0.0254624710	−0.0172507507
60	0.0149570492	0.0168282350	−0.0018711858

The absolute residual versus the predictor in usual scale and log scale are shown in Fig. 12 and Fig. 13, respectively. These do not show any specific pattern. For seed weight 500 g, four predicted values of yam weight were negative; as such these are not seen in Fig. 13a.

Similar plots of absolute residual versus yam weight, the dependent variable in usual scale and log scale are shown in Fig. 14 and Fig. 15, respectively. These figures do not show any specific pattern.

Next we consider multiple regression of interim yam weight based on above ground biomass and interim plant lifetime (t), in days. Multiple regression squared $R^2 = 0.7439$ is high for seed weight 500 g. The regression line being $y = -3.01163 + 0.53668x + 0.02571t$.

For seed weight 650 g and 800 g, the value of R^2 is 0.6003 and 0.5105, respectively.

For all seed weight combined, $R^2 = 0.6425$ and the estimated regression line is $y = -2.570078 + 0.425820x + 0.025493t$.

Accuracy of regression is improved in logarithmic scale. Multiple regression squared $R^2 = 0.8204$ is high for seed weight 500 g. The regression line being $\log y = -6.8715 + 1.4311 \log x + 1.1957 \log t$.

For seed weight 650 g and 800 g, the value of R^2 in log scale is 0.5359 and 0.5573, respectively.

For all seed weight combined, $R^2 = 0.6496$ in log scale and the estimated regression line is $\log y = -8.0125 + 0.9284 \log x + 1.5655 \log t$.

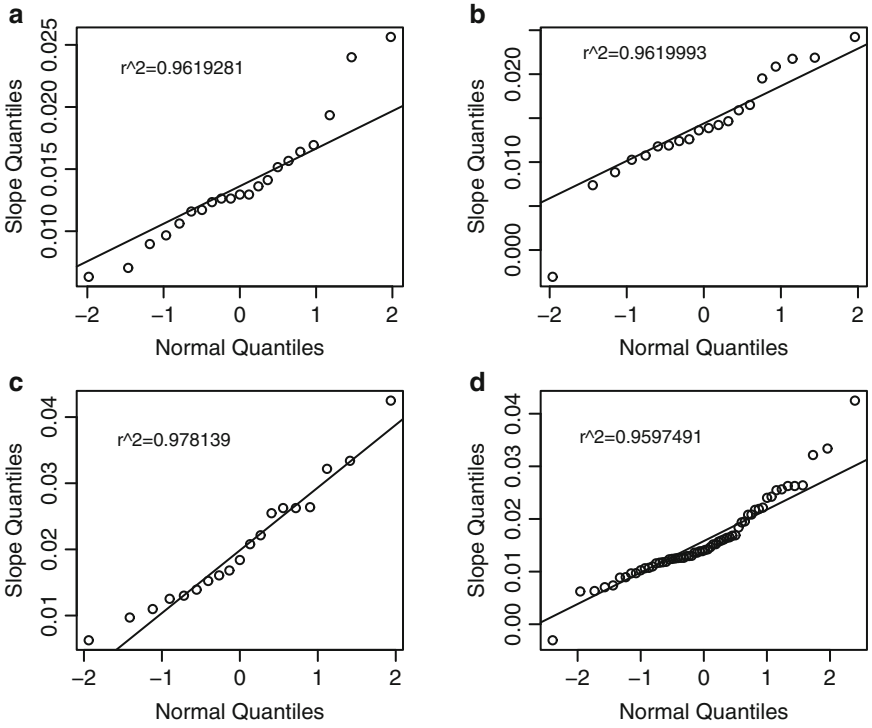


Fig. 7 Normal Quantile plot. (**a**) Second slope (4/9/14); (**b**) second slope (20/9/14); (**c**) second slope (21/9/14); (**d**) second slope (combined)

The absolute residual versus the interim yam weight in usual scale and log scale are shown in Fig. 16 and Fig. 17 respectively. Once again, these do not show any specific pattern.

Farmers have an idea about final yield based on the above ground biomass at the middle of season. Consider multiple regression on final yield based on above ground biomass and plant lifetime till the end. For seed weight 500 g, R^2 of multiple regression is 0.774. The regression line being $y = 0.383917 + 0.841309x - 0.003094t$. Biomass is highly significant with $p = 1.76 \times 10^{-06}$.

For seed weight 650 g and 800 g, the value of R^2 is 0.6635 and 0.3052, respectively. Biomass is significant in both the cases.

For all seed weight combined $R^2 = 0.5858$ and the estimated regression line is $y = 0.465162 + 0.570168x + 0.001582t$. Biomass is highly significant with $p = 4.94 \times 10^{-12}$.

Accuracy of multiple regression is improved in logarithmic scale. For seed weight 500 g, R^2 of multiple regression in logarithmic scale is 0.8121. The regression line being $y = -0.07992 + 1.14006 \log x - 0.06510 \log t$. Biomass is highly significant with $p = 3.68 \times 10^{-07}$. For seed weight 650 g and 800 g, the value of R^2 is 0.5357 and 0.4697, respectively, in logarithmic scale. Biomass is significant in both the cases.

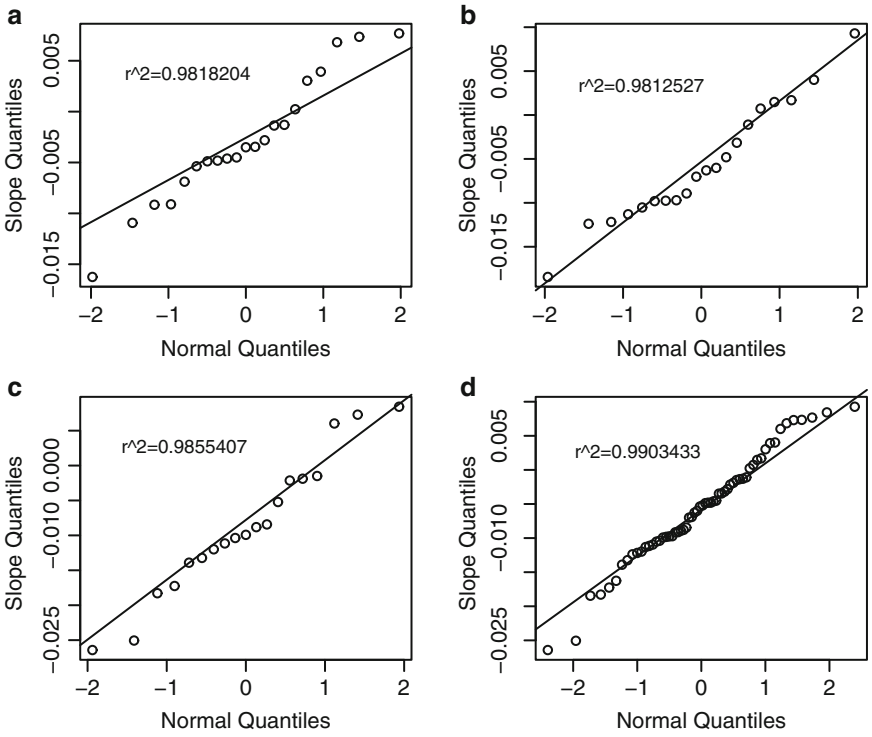


Fig. 8 Normal Quantile plot. (**a**) Slope difference (4/9/14); (**b**) slope difference (20/9/14); (**c**) slope difference (21/9/14); (**d**) slope difference (combined)

For all seed weight combined $R^2 = 0.623$ and the estimated regression line is $y = -0.37435 + 0.78323 \log x + 0.07546 \log t$. Biomass is highly significant with $p = 2.5 \times 10^{-13}$. Absolute residual plots (not shown in figures) do not exhibit any pattern in this case.

Above ground biomass is seen to be significant in all cases as a predictor of final yield of the crop.

Since the growth curve corresponding to seed weight 650 g is superior, it is of interest to examine the proliferation rate of this. In Fig. 18 we plot the proliferation curve obtained by a technique described in Dasgupta (2013b). To start with, median of the (normalised) exponentially weighted individual 213 slope estimates obtained from the lowess estimate of growth (the blue curve in Fig. 5) are considered for a particular time point. These median values of slope estimates for 213 time points are smoothed by SPlus smooth.spline with spar=0.2 to obtain Fig. 18.

Proliferation rate of yam growth shown in Fig. 18 for seed weight 650 g decreases in the beginning and gradually takes an upturn slightly before 100 days, and then increases sharply towards the end indicating the final days in yam lifetime are important in yam growth.

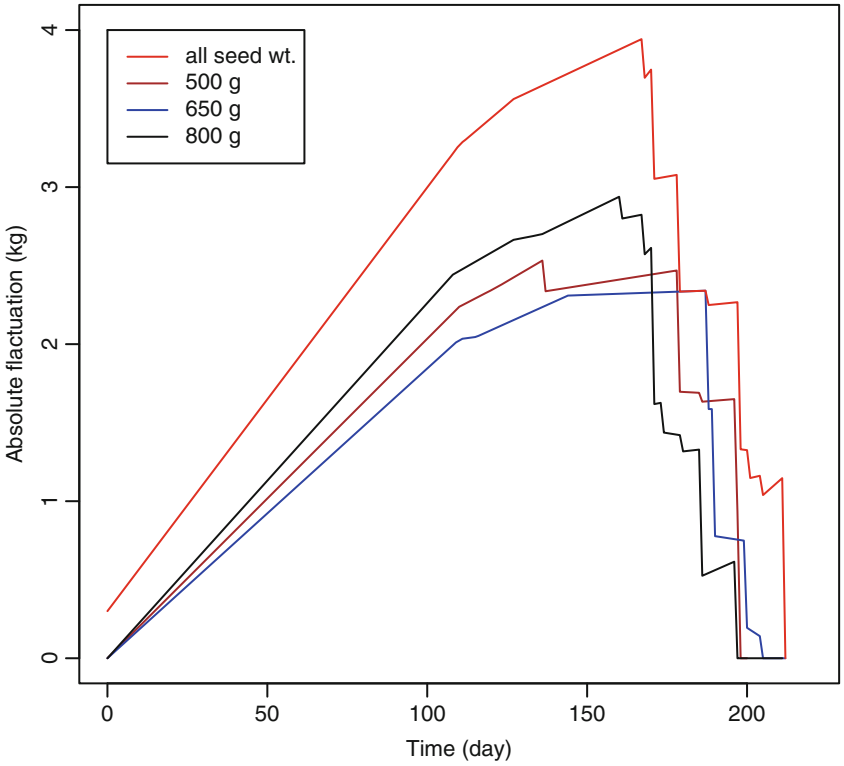


Fig. 9 Maximum fluctuation of growth curves over time

One may compare above ground biomass over different seed weights. Tender yam stems have a market value. To test whether biomass b_w is higher for higher seed weight w, $w = 500, 650, 800$ g, consider first the Wicoxon 2 sample U statistic with kernel $I(b_{650} > b_{500})$, based on column 6 of Table 1, computed on the basis of 20 plants in each group of seed weight. The standardised value of $U = U_{650,500}$ for comparing seed weight 650 g with 500 g is $U^* = (U - 200)/\{\frac{400(20+20+1)}{12}\}^{1/2} = (250 - 200)/36.97 = 1.35$, and $\hat{P}(b_{650} > b_{500}) = U/400 = 250/400 = 0.625$. When compared with a normal deviate, one sided p-value of significance for U^* is 0.0885.

Similarly, $\hat{P}(b_{800} > b_{650}) = U/400 = 242/400 = 0.605$. The standardised value of U is $U^* = (U - 200)/\{\frac{400(20+20+1)}{12}\}^{1/2} = (242 - 200)/36.97 = 1.14$, one sided p-value of significance is $p = 0.127$.

Finally, to compare biomass b_{800} with b_{500}, one has $\hat{P}(b_{800} > b_{500}) = 310/400 = 0.775$, with $U^* = (310 - 200)/36.97 = 2.97$; one sided p-value of significance for U^* is 0.0015.

The above ground yam biomass for seed weight 800 g is significantly higher than that for 500 g.

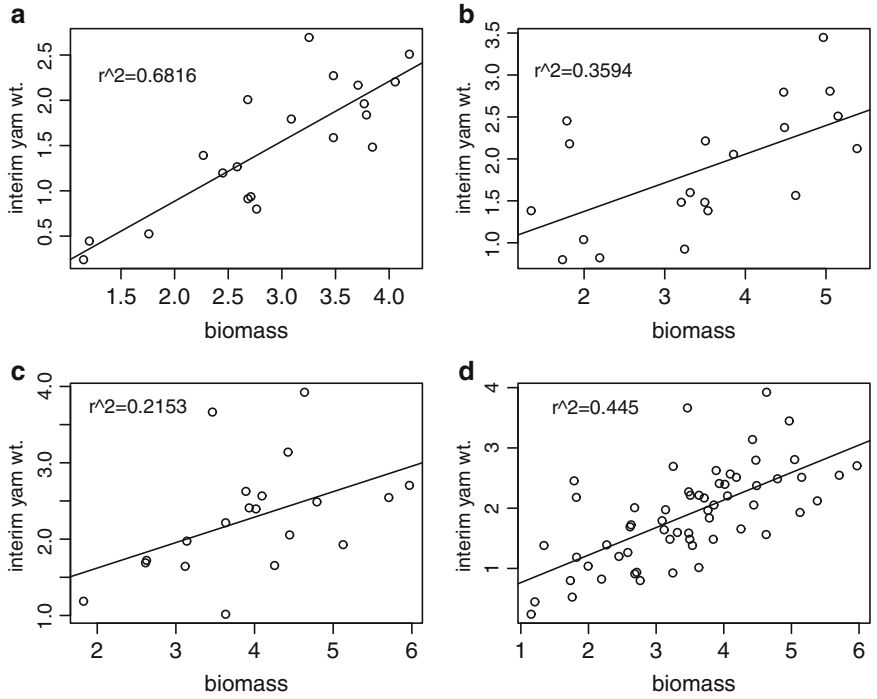


Fig. 10 Regression of interim yam wt. on above ground biomass. (**a**) Seed wt. 500 g; (b) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

3 Plant Lifetime, Folded Normal Distribution and Weibull Model

(i) Continuous variable: Growth curve of yam plant lifetime may be modeled by a Weibull distribution e.g., see Dasgupta (2014).

Hazard rate of plant lifetime distribution is associated with crop harvest rate and market supply. Figure 19 shows 60 yam plant lifetimes in Weibull probability plot. Data points adhere to 95 % confidence band, justifying Weibull model for plant lifetime. Growth curve of yam plant lifetime is shown in Fig. 20. The curve shows a sharp rise towards end.

Weibull distribution has wide applications in industrial context and is a candidate model for burr, see Dasgupta (2011). Effect of crossing a (large) threshold on the quantiles of Weibull distribution is considered in Dasgupta (2013c). Such studies are of interest for analyzing excessively large yam plant lifetime, following a Weibull model.

Consider the hazard rate of distribution G for folded normal variable $Y = |X|$, where $X \sim N(0, \sigma^2)$. The distribution of Y is positively skew. Folded normal distributions also have applications in modeling industrial job characteristics when

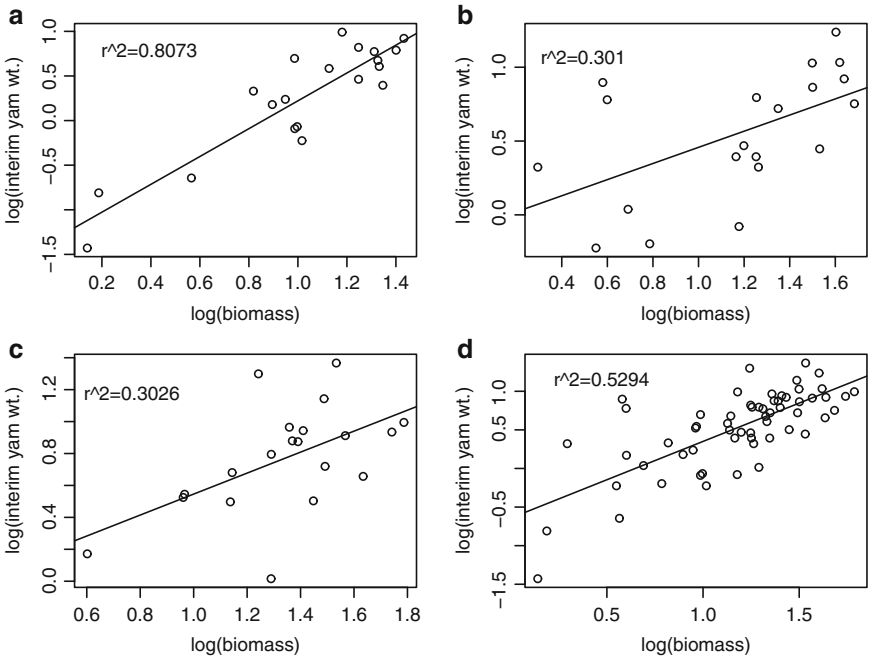


Fig. 11 Regression of interim yam wt. on above ground biomass in log scale. (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) 800 g; (**d**) all seed wt.

signs of the variables are ignored; see, e.g., Dasgupta (2005) and Leone et al. (1961). From normal tail probability expansion up to third order for large values of t, one may write for hazard rate $h(t)$ of an *approximate* folded normal variable $Y_n \sim Y(1 + o_p(1))$ with density $2\phi(t/\sigma)(1 + o(1)), t \geq 0$ as

$$h(t) = \frac{\phi(t/\sigma)}{\Phi(-t/\sigma)}(1 + o(1)) \approx \frac{\phi(t/\sigma)}{(\frac{1}{t/\sigma} - \frac{1}{(t/\sigma)^3} + \frac{3}{(t/\sigma)^5})\phi(t/\sigma)}(1 + o(1))$$
$$= \frac{t}{\sigma}\{1 - (t/\sigma)^{-2} + 3(t/\sigma)^{-4}\}^{-1}(1 + o(1))$$
$$\approx \frac{t}{\sigma}\{1 + (t/\sigma)^{-2} - 3(t/\sigma)^{-4}\}, t \to \infty \qquad (3.1)$$

The leading term in the approximation (3.1) for t large is $\frac{t}{\sigma} = g(t)$, say. Error in approximation for folded normal is small, as this is of order t^{-2}.

Since the hazard rate *characterises* a distribution, the only distribution with hazard rate g has the distribution function

$F(x) = 1 - e^{-\int_0^x g(y)dy} = 1 - e^{-x^2/(2\sigma^2)}$.

This is a Weibull model of the form $F(t) = 1 - \exp(-(t/\beta)^\alpha), \beta > 0$, $\alpha > 0, t > 0$, having hazard rate $g_1(t) = \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}$.

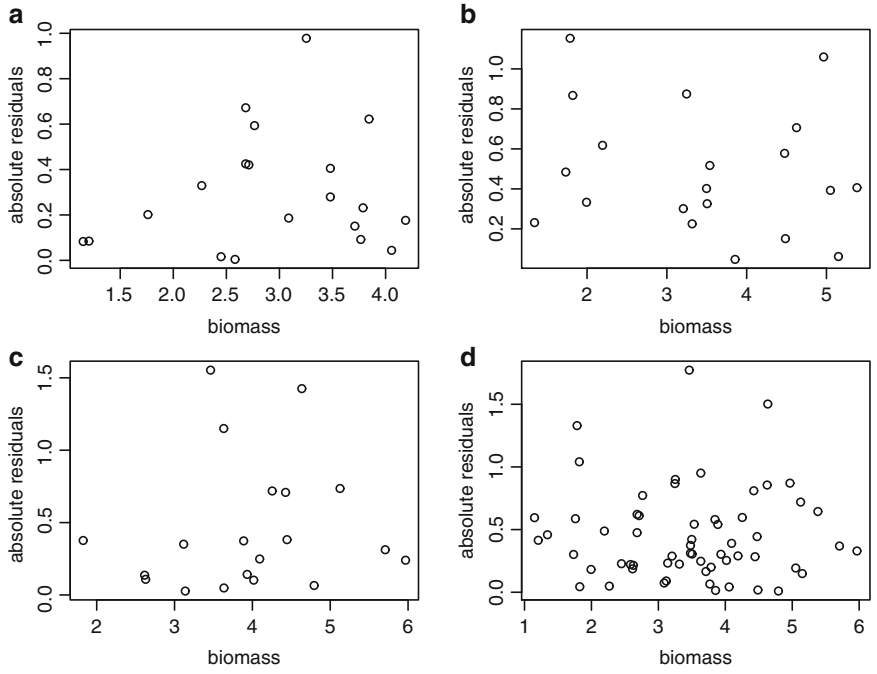


Fig. 12 Absolute residual plot of yam wt. in above ground biomass regression. (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

With $\alpha = 2, \beta = \sqrt{2\sigma}$, F matches G in terms of the limiting form of hazard rate. Hence the following.

Theorem 3.1. *Weibull model* $F(t) = 1 - \exp(-(t/\beta)^{\alpha}), \beta > 0, \alpha > 0, t > 0$, *with hazard rate* $h(t) = \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}$ *is the only distribution for* $\alpha = 2, \beta = \sqrt{2\sigma}$, *that has the limiting hazard rate of a folded normal variable* $Y = |X|$, *where* $X \sim N(0, \sigma^2)$. *The error in hazard rate approximation is* $O(t^{-2})$, *as* $t \to \infty$.

(ii) Discrete variable: Dasgupta (1993) characterised discrete version of the normal distribution by considering the m-dimensional random variable $(X_1, \cdots, X_m)$, where the X_i are independent, radially symmetric, and discrete, and found that if their joint distribution depends only on the displacement $r^2 = X_1^2 + \cdots + X_m^2$ then for $m \geq 4$ the marginal distributions have p.m.f. $P(X_i = x) = c \exp(-\beta x^2)$, $i = 1, \cdots, m$, where the support is $Z = \{0, \pm 1, \pm 2, \pm 3, \cdots\}, \beta > 0$, and c is a normalising constant involving Theta function of order 3. The characterisation does not hold for $m = 2, 3$ is also shown therein.

For a relationship between the sum of series involving the terms $\exp(-\beta x^2)$ and expansion of Theta functions, see, e.g., Whittaker and Watson (1927).

Characterisation results are possible with similar assumptions for discrete random variables on the set of nonnegative integers $Z^* = \{0, 1, 2, 3, \cdots\}$, so as to

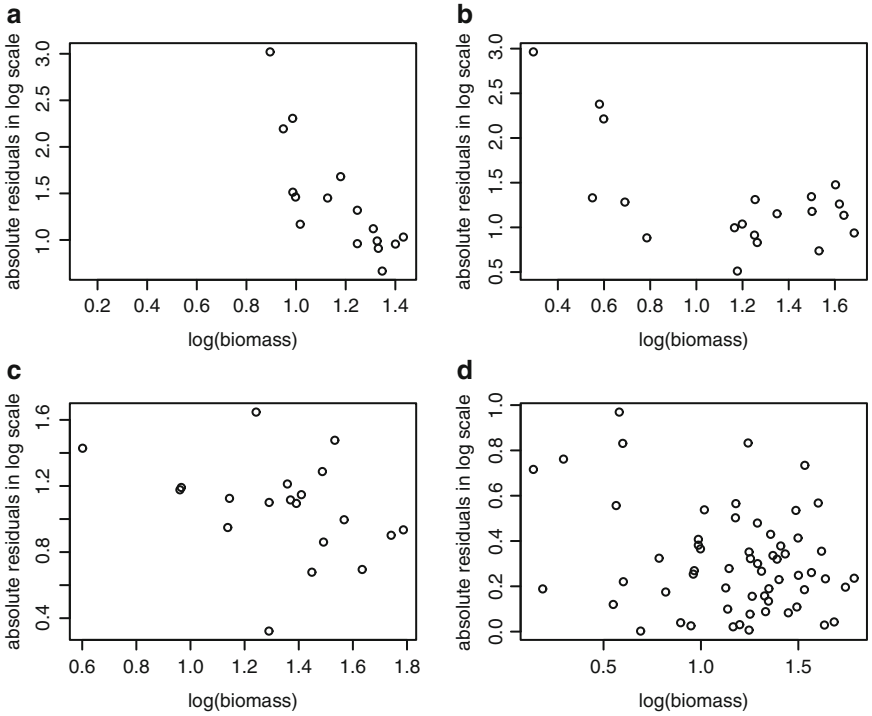


Fig. 13 Absolute residual plot of yam wt. in above ground biomass regression (log scale). (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

accommodate nonnegative discrete random variables like plant lifetime measured in days. Characterisation theorems help us to focus on appropriate choice of models. Proceeding in a similar fashion like Theorem 1 of Dasgupta (1993), it is possible to obtain the following.

Proposition 3.1. *Let* $(X_1, \cdots, X_m)$ *be the m-dimensional random variable where the* X_i *are independent with support as* Z^*. *If the joint distribution depends only on the displacement* $r^2 = X_1^2 + \cdots + X_m^2$, *then for* $m \geq 4$ *the marginal distributions have p.m.f.* $P(X_i = x) = c \exp(-\beta x^2), i = 1, \cdots, m, x \in Z^*, \beta > 0$, *and c is a normalising constant. This characterisation does not hold for* $m = 2, 3$.

The p.m.f. stated in Proposition 3.1 refers to a discrete version V of folded normal distribution, folded at origin. To avoid confusion with notations let us write the form as

$$P(V = k) = p_k = ce^{-k^2/(2\sigma^2)},\ \sigma^2 \geq 0,\ k \in Z^* \tag{3.2}$$

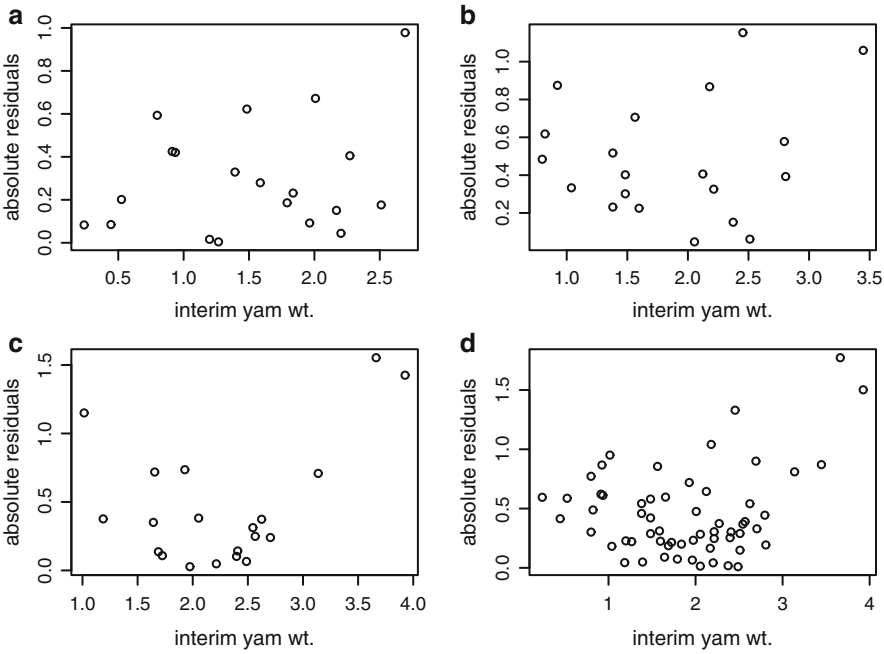


Fig. 14 Absolute residual vs. yam wt. in above ground biomass regression. (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

with hazard rate

$$h_k = \frac{p_k}{\sum_{j=k}^{\infty} p_j} = e^{-k^2/(2\sigma^2)} / \sum_{j=k}^{\infty} e^{-j^2/(2\sigma^2)} = \phi(k/\sigma) / \sum_{j=k}^{\infty} \phi(j/\sigma) \tag{3.3}$$

Next, write the denominator in (3.3) as

$$\sum_{j=k}^{\infty} e^{-\frac{j^2}{2\sigma^2}} = e^{-\frac{k^2}{2\sigma^2}} \{1 + e^{-\frac{(2k+1)}{2\sigma^2}} + e^{-\frac{(4k+4)}{2\sigma^2}} + e^{-\frac{(6k+9)}{2\sigma^2}} + e^{-\frac{(8k+16)}{2\sigma^2}} + \cdots\}$$

$$= e^{-k^2/(2\sigma^2)} \{1 + e^{-(2k+1)/2\sigma^2} (1 + O(e^{-\frac{k}{\sigma^2}}))\} \tag{3.4}$$

Thus from (3.3) and (3.4), one gets

$$h_k = 1 - \gamma e^{-\beta k} (1 + O(e^{-\frac{k}{\sigma^2}})) \sim 1 - \gamma e^{-\beta k}, \quad k \to \infty \tag{3.5}$$

where $\beta = \sigma^{-2}, \gamma = e^{-\beta/2}$.

If U is Weibull with survival function $\overline{F}(u) = \exp(-\lambda u^{\alpha})$, $\lambda > 0, \alpha > 0$, then the integer part $V^* = [U] \in Z^*$ is a discrete Weibull variable with survival function

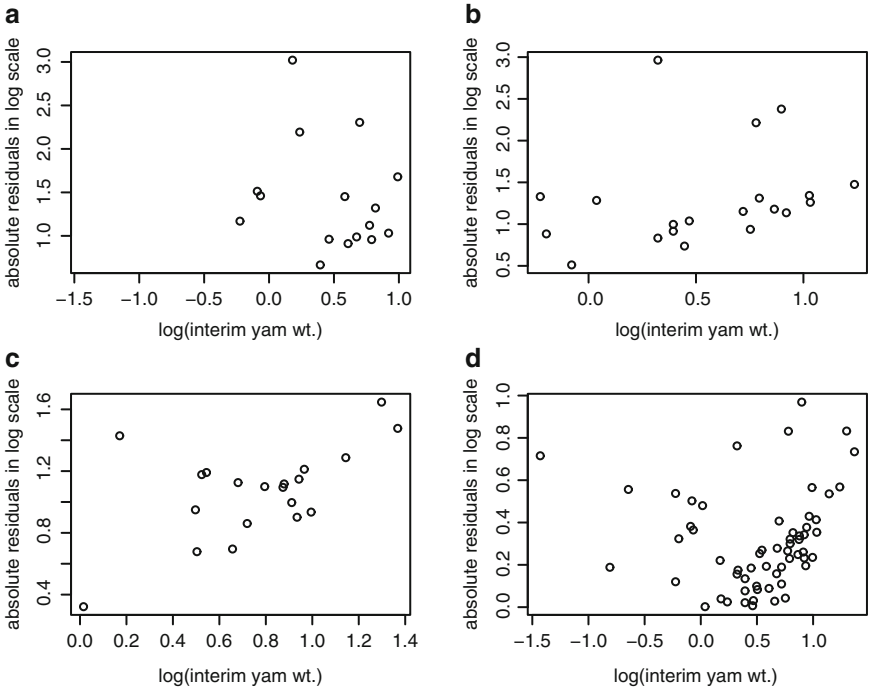


Fig. 15 Absolute residual vs. yam wt. in above ground biomass regression (log scale). (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

$$P(V^* \geq k) = P(U \geq k) = \exp(-\lambda k^\alpha) = q^{k^\alpha}, q = \exp(-\lambda);\ k \in Z^*.$$

The p.m.f. of V^* is,

$$p_k^* = P(V^* = k) = q^{k^\alpha} - q^{(k+1)^\alpha};\ k \in Z^*.$$

Hazard rate of discrete Weibull variable V^* for $k \in Z^*$ has the following representation.

$$r_k^*(q, \alpha) = p_k^* / \sum_{j=k}^{\infty} p_j^* = 1 - (q)^{(k+1)^\alpha - k^\alpha} = 1 - (q)^{\alpha \kappa_1^{\alpha-1}}, \kappa_1 \in (k, k+1).$$

The r.h.s. of (3.5) is of the form (8) of Dasgupta (2014) with $\alpha = 2$, as seen from the above expression of r^*. Hence the following,

Proposition 3.2. *Consider the discrete version of folded normal distribution with p.m.f. given in* (3.2)*. The hazard rate of distribution* (3.2) *is* $h_k \sim 1 - \gamma e^{-\beta k}$, $\beta = \sigma^{-2}, \gamma = e^{-\beta/2}$, $k \to \infty$, *resembling the form (8) with* $\alpha = 2$ *given in Dasgupta (2014) for hazard rate of a discrete Weibull variable.*

Proximity of these distributions with Weibull model is of interest in survival analysis including plant lifetime modeling.

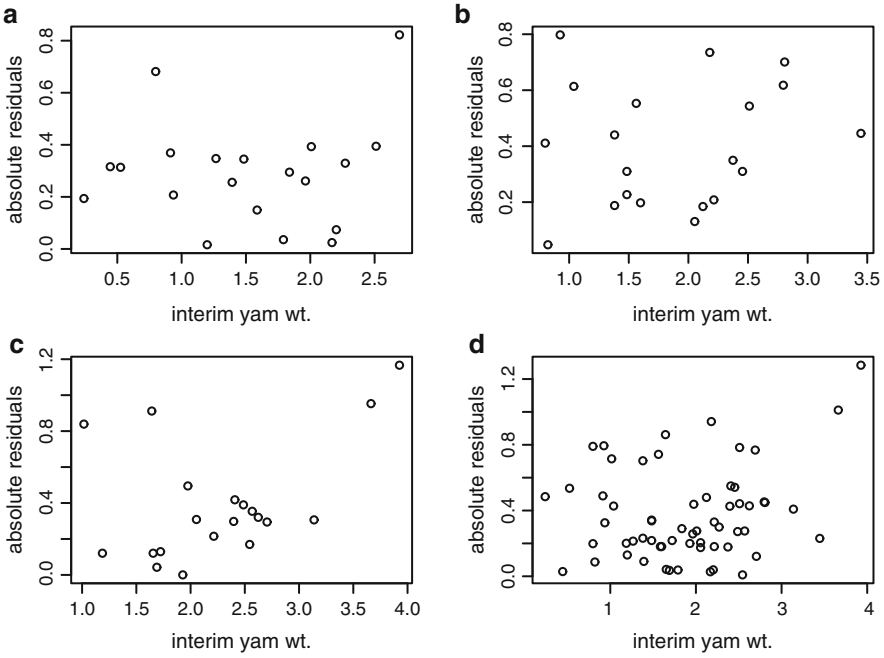


Fig. 16 Absolute residual vs. yam wt. in multiple regression. (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

4 Discussion and Conclusions

Growth curve of Elephant foot yam with seed weight 650 g shows superior yield in longitudinal study. As such this seed weight may be recommended to the farmers of Giridih, Jharkhand. The curve has a spike towards end of the plant lifetime. Yield will be much higher, to the tune of five times, if the crop is harvested at the end of season. Farmers sometime prefer early harvesting for monetary reasons. The yield is approximately double, if the crop is harvested after 75 days from sprouting of the plant; young and tender stems also have a market value. Above ground biomass of yam is significantly higher for seed weight 800 g than that for 500 g, in the middle of crop season. Growth slope and difference of growth slopes over different time segments for individual plants are linear combinations of yield observations, and are seen to follow normal distribution. Residual variations of the growth curve are modeled by the asymptotic variance of Ornstein–Uhlenbeck process. These are also compared with Wilcoxon 2 sample U statistics. Curve corresponding to 650 g of seed weight seems superior for less variation. Prediction of underground yam weight based on observable above ground biomass reveals a nonlinear relationship; the coefficient of determination r^2 is high in logarithmic scale. The same holds

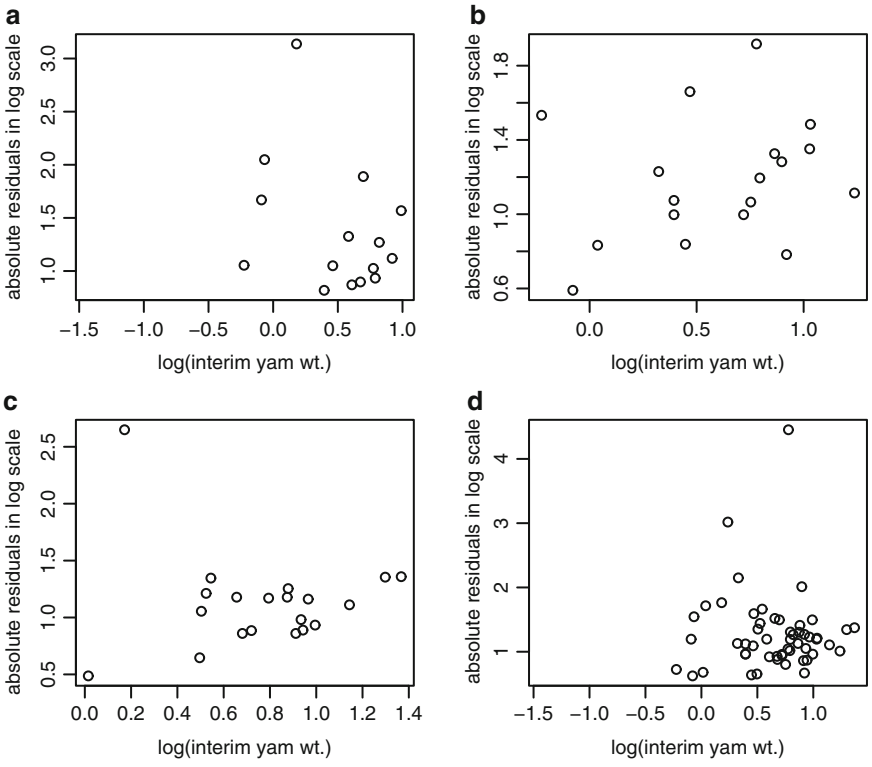


Fig. 17 Absolute residual vs. yam wt. in multiple regression (log scale). (**a**) Seed wt. 500 g; (**b**) seed wt. 650 g; (**c**) seed wt. 800 g; (**d**) all seed wt.

true for R^2 of multiple regression with additional predictor variable t of plant lifetime while recording interim yield. Multiple regression on final yield based on above ground biomass at the middle of the season along with plant lifetime till the end is high in logarithmic scale. At each stage of analysis, above ground biomass at the middle of season remains significant for yield prediction, indicating that biomass is a good predictor for crop yield.

Modeling of plant lifetime is made by Weibull distribution. For some specific choice of parameters Weibull is the only distribution that has the limiting hazard rate of folded normal variable. Similar results hold for discrete version of variables.

With a sample size 60, the present study reconfirms some of the results stated in Dasgupta (2015) based on a sample size 6, which is 10 % of the present size. This conforms to the general assertion that data contains a lot of information, even in small sample sizes.

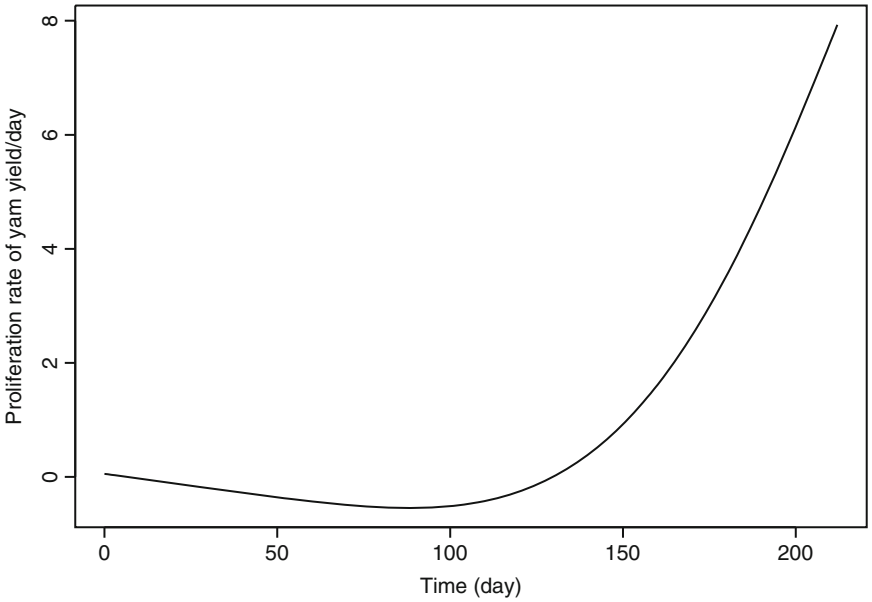


Fig. 18 Proliferation rate of yam with median, wt. exp(−.01 x); seed wt. 650 g

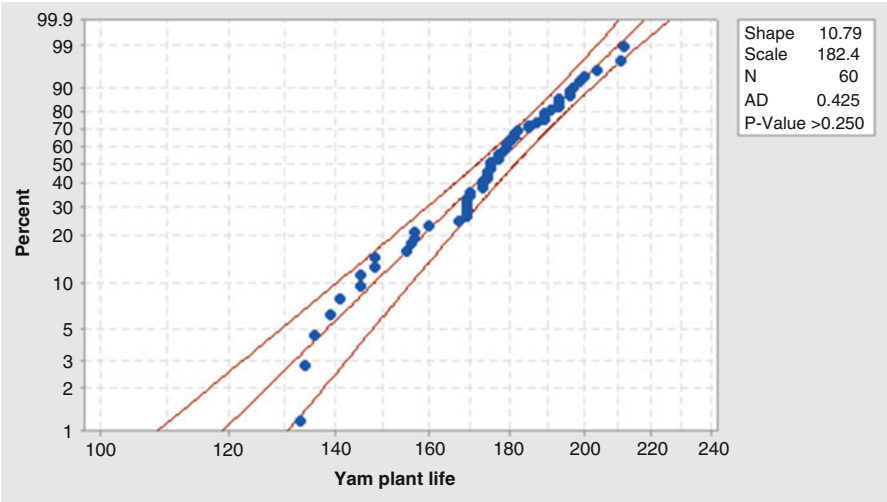


Fig. 19 Weibull probability plot of 60 yam plant life, 2014

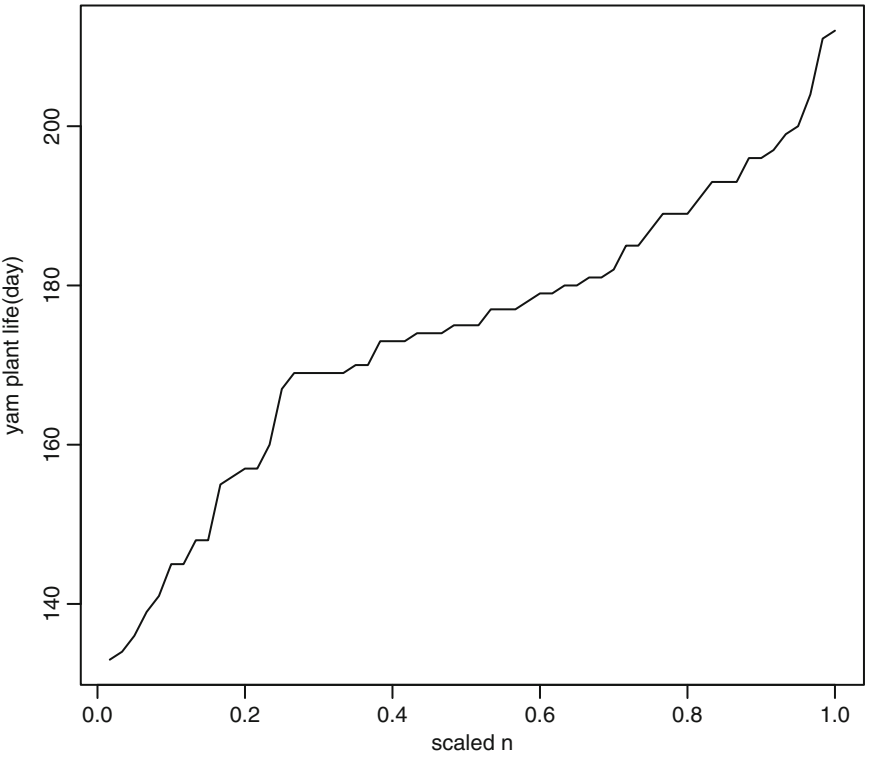


Fig. 20 Growth curve of 60 yam plant life time resembling Weibull growth

References

Cleveland WS (1981) LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat 35:54

Dasgupta R (1993) Cauchy equation on discrete domain and some characterization theorems. Theory Probab. Appl 38(3):520–524

Dasgupta R (2005) On the distribution of straightness. In: Proceedings of the national conference on tools & techniques for quality and productivity improvement. ISI, New Delhi, pp 49–55

Dasgupta R (2011) On the distribution of burr with applications. Sankhyā B 73:1–19

Dasgupta R (2013a) Yam growth experiment and above-ground biomass as possible predictor, Chap 1. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 1–33

Dasgupta R (2013b) Non uniform rates of convergence to normality for two sample U-statistics in non IID case with applications, Chap 4. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 61–88

Dasgupta R (2013c) Characterization theorems based on conditional quantiles with applications. J Environ Stat 4(6):1–25

Dasgupta R (2014) Characterization theorems for Weibull distribution with applications. J Environ Stat 6(4):1–25

Dasgupta R (2015) Plant sensitivity and growth curve analysis of elephant foot yam, Chap 1. In: Dasgupta R (ed) Growth curve and structural equation modeling, 1st edn. Springer proceedings in mathematics & statistics. Springer, New York

Leone FC, Nottingham RB, Nelson LS (1961) The folded normal distribution. Technometrics 3:543–550

Whittaker ET, Watson GN (1927) Modern analysis. Cambridge University Press; 4th edition

Optimal Choice of Small Regular Shapes for Accidentally Damaged Tessellation

Ratan Dasgupta

Abstract Objects of regular shapes like triangles, squares, pentagons, etc., are assembled to construct differently designed tessellation. We study the optimal choice of small regular shapes (SRS) so as to minimise replacement cost due to accidental damage in an assembled structure. The SRS are often used to cover regions vulnerable to damage from a sudden hit. Efficiency and "equivalent edge number" of an SRS ensemble are defined and applications indicated. Monotonicity of "equivalent edge number" in tessellations is proved under certain assumptions. Growth in efficiency of SRS with respect to increase in the number of sides of SRS is studied. An example discussed, modeling of the crack lengths in damaged shield by Ornstein–Uhlenbeck process is made and severities of damage in two occasions are compared. A data set arising from growth experiments having a structured layout like tessellations of seedling is analysed in a similar technique to study possible propagation of damage due to pest infection in plants from afflicted pits to adjacent pits.

Keywords Small regular shape • Regular polygon • Equivalent edge number • Tessellation • Fractal dimension • Ornstein–Uhlenbeck process

MS subject classification: Primary: 60D05, secondary: 62P30.

1 Introduction

Consider a floor or window glass pane that has to be covered with small regular shaped tiles. We consider tessellation by small regular shapes (SRS) to minimise the repair cost of accidental damage/crack caused due to sudden hit by some external body e.g., random hit by a hard object like bullet. Subsequently those broken SRS have to be replaced by new tiles to repair the damaged panel.

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.), *Growth Curve and Structural Equation Modeling*, Springer
Proceedings in Mathematics & Statistics 132, DOI 10.1007/978-3-319-17329-0_15

The regular shapes are seen to arise in nature, e.g., hexagonal structure of a honeycomb cell, pentagonal cross-section of okra, sea star, star-fruit with cross section having fivefold radial symmetry, etc.

To cover an area by objects of SRS i.e., small regular polygons of *same type*, we need to restrict the number of edges in SRS. Only three regular polygons tessellate in the Euclidean plane resulting in uniform pattern viz., triangles, squares or hexagons, see Branko and Shephard (1987). Since the regular polygons in a tessellation must fill the plane at each vertex, the interior angle must be an exact divisor of 360°. This works for the triangle, square and hexagon. For all the other SRS, the interior angles are not exact divisors of 360°, and therefore those figures cannot tile the plane.

Hirschhorn and Hunt (1985) considered tessellation of plane with convex pentagon having different internal angles at cojoined edges. It is possible to cover areas in a plane by adjusting edges of different SRS giving rise to different designs in architecture, e.g., see Steinhaus (1999).

Some broken/incomplete shapes may be required near the edges of a large square panel to fill the gaps.

To start with, we shall make the following simplifying assumption (to be relaxed later): the size of the hitting object is small enough compared to the size of an SRS, so that if the hit is made completely within the periphery of a particular SRS and cause damage, then a single SRS is affected, and only that SRS has to be replaced for damage repair. In such a situation inside crack may extend up to the periphery of that SRS, and we assume that it will not cross the boundary of that SRS. If the hit occurs on an edge of SRS affecting the nearby SRS as well, replacement cost is more. The scenario may be generalised to a higher dimension. SRS of dimension 3 are of interest when a given volume has to be filled by objects of smaller volume; replacement cost will be less when only one such SRS is accidentally damaged. Cost will increase if accidental damage crosses the boundary of originating SRS and affects adjacent SRS as well. We show that SRS that are more near to circular (spherical) shapes are superior for damage control and cost reduction.

In Sect. 2 we state the main results and develop a measure of efficiency to compare two tessellations with different combinations of SRS. For an SRS ensemble we introduce a notion of "equivalent" edge number that may not be an integer, this is much like a fractal dimension in self-similar patterns, and show that "equivalent" edge number is monotone. Assumption that a single SRS be affected is relaxed. Modeling the crack lengths by Ornstein–Uhlenbeck process and comparison of damage severity in two occasions are made in terms of process parameters. Some examples are discussed. The results may be extended to higher dimensions. In Sect. 3 we analyse a data set on agricultural growth experiment taking into account propagation of damage due to infection in adjacent pits percolated from damaged pits, where the experimental data arise from a structured layout like tessellation. In Sect. 4, three dimensional honeycomb structure and choice of SRS are examined.

2 Efficiency in SRS Ensemble and Main Results

Denote the perimeter and area of a planar shape by L and A, respectively. Then

$$4\pi A \leq L^2 \tag{1}$$

This is known as the Isoperimetric Inequality, e.g., see Osserman (1986). The equality holds only for a *circle.* Isoperimetric Inequality can be extended to higher dimensional spaces. For example, if S is a surface area while V a volume of a three dimensional body, then

$$36\pi V^2 \leq S^3 \tag{2}$$

The inequality states that among all three dimensional solid bodies with a given surface area the *sphere* has the largest volume.

Inequality (1) provides an indication that an SRS that is nearer in shape to circle of same area has lower perimeter and may have advantage over other SRS in two dimensions for replacement cost reduction due to damage. Lower perimeter of SRS with area fixed makes it less likely for a bullet to hit on edges.

It may be mentioned that the terminology "isoperimetric" is also commonly used in a metric space associated with Borel probability measure, and *should not be confused* with the problem considered in the present context. Isoperimetric (literal meaning "having the same perimeter") inequalities in the context of probability and associated concentration of Borel probability measure on a compact metric space lead to exponential bounds for concentration function. Under certain conditions, complement of the neighbourhood of radius r of a set with probability larger than $1/2$ decreases exponentially fast when $r \uparrow \infty$; for example, see Ledoux and Talagrand (2002).

Since the SRS are adjusted side-by-side, an edge of the SRS is cojoined with the edge of adjacent SRS. As a result, total length of small edges inside the large panel, ignoring possibly those incomplete edges near the panel sides, is approximately equal to the total length of cojoined edges, and that equals half of the total perimeter of these SRS.

2.1 Computation of Efficiency

Assuming that the bullet hits the panel at random, the probability of damage to more than one SRS is proportional to its perimeter to a first degree of approximation, when we consider SRS with different number of edges having equal area. To have an optimal choice, we compare perimeters of SRS. The restriction of equal area is made from the viewpoint of production-cost of SRS. To a first approximation, cost is proportional to amount of production material required, which in turn is proportional to area.

It is possible to cojoin SRS of different types to have different type of patterns. We consider broader class of shapes of the type squares (four sides), pentagon (5), hexagon (6), heptagon (7), octagon (8), nonagon (9), etc., to compare relative advantage in minimising the perimeter. We prove the following.

Proposition 1. *Let the probability of hitting the edge of an individual SRS in an ensemble of SRS in a tessellation be proportional to the perimeter of SRS. Let the efficiency $e_{i,4}$ of an i-edged SRS, compared to a square SRS with same area, be defined as the ratio of the corresponding probabilities, $i \geq 3$.*

Then,

$e_{3,4} = 0.8773827,\ e_{5,4} = 1.049336,\ e_{6,4} = 1.07457,\ e_{7,4} = 1.089304,$
$e_{8,4} = 1.098684,\ e_{9,4} = 1.105034, \cdots, e_{\infty,4} = 1.128379.$

Proof. Let N be the number of sides and r be length from centre to a corner. Then area of a regular polygon with N sides is $(1/2)N \sin(2\pi/N)r^2$.

Area of a pentagonal SRS with side length t and perimeter $5t$ is $A_5 = \frac{t^2\sqrt{25+10\sqrt{5}}}{4} \approx 1.720477t^2$. Area of a square shaped SRS with side length d and perimeter $4d$ is $A_4 = d^2$. Equating the areas $A_4 = A_5$, one gets advantage of pentagonal SRS over square SRS as the ratio of the probabilities (or, ratio of perimeters) expressed approximately as
$e_{5,4} = (4d)/(5t) = 1.049336.$

Similarly, area of a hexagonal SRS with side length t and perimeter $6t$ is $A_6 = \frac{3\sqrt{3}}{2}t^2 \approx 2.5981t^2$. Equating the areas $A_4 = A_6$, one gets advantage of hexagonal SRS over square SRS as the ratio of the probabilities expressed approximately as
$e_{6,4} = (4d)/(6t) = 1.07457.$

For heptagonal SRS with side length t and perimeter $7t$, the area is
$A_7 = \frac{7}{4}t^2 \cot\frac{\pi}{7} \approx 3.633912444t^2$, and $e_{7,4} = (4d)/(7t) = 1.089304$.

For octagonal SRS with side length t and perimeter $8t$, the area is
$A_8 = 2(1+\sqrt{2})t^2 \approx 4.828427t^2$, and $e_{8,4} = (4d)/(8t) = 1.098684$.

For nonagonal SRS with side length t and perimeter $9t$, the area is
$A_9 = \frac{9}{4}t^2 \cot\frac{\pi}{9} \approx 6.18182t^2$ and efficiency under the restriction of same area $A_9 = A_4$ is $e_{9,4} = (4d)/(9t) = 1.105034$.

For a triangular SRS with side length t and perimeter $3t$, the area is
$A_3 = \frac{\sqrt{3}}{4}t^2$ and $e_{3,4} = (4d)/(3t) = 2(3^{1/4})/3 = 0.8773827$.

It is evident that as the number of sides in SRS increases, efficiency also increases and the shapes gradually tend to be circular.

The limiting efficiency can be computed in terms of ratio of approximate probabilities as follows:
$\pi r^2 = d^2, e_{\infty,4} = \frac{4d}{2\pi r} = 2/\sqrt{\pi} = 1.128379.$

Note that triangular shaped SRS are inferior to square shaped SRS. Gain in efficiency due to the variation of SRS is about 25 %. Tessellations of plane by two or more convex regular polygons such that the same polygons in the same order

surround each polygon vertex are called semiregular tessellations, or sometimes Archimedean tessellations. Demiregular (or polymorph) tessellations are orderly compositions of regular and semiregular tessellations.

Assume that the cost of replacing an SRS is proportional to its area. Further assume that the number of SRS is large enough so that the edge perimeter of the panel to be covered is negligible compared to the combined perimeters of the SRS required for tessellation of a large area. Broken, incomplete shapes to fill the gaps near periphery may also be ignored in the approximation. We have the following.

Proposition 2. *Let the number of regular polygon with i-edge be $n_i, i = 3, 4, 5 \cdots$ used for a general polymorph tessellation A of a two dimensional area. The efficiency of such a tessellation against damage by a random bullet-hit with respect to tessellation made by squares alone, having same respective area with those SRS used in tessellation A is approximately,*

$$e_{A,4} = \frac{1}{n} \sum_i n_i e_{i,4} + o(1), \text{ as } n \to \infty \tag{3}$$

where $n = \sum n_i$.

Note: The measure $e_{A,4}$ represents ratio of the two probabilities of damage for more than one SRS in the ensemble A, and that for a hypothetical ensemble made up with squares only, having respective areas with SRS in A.

Proof. The proof follows from the fact that probability of a striking bullet or small object damaging more than one SRS is proportional to the sum of perimeters of SRS to a first degree of approximation, from the assumptions made.

Since $e_{i,4}$ is increasing in the first coordinate, it is clear that regular polygons with more number of edges are more desirable for less damage, if the area of the polygons remain same to the respective squares.

Proposition 1 gives rise to the possibility of comparing two different tessellations, A with respect to B for the same area with the following measure of efficiency

$$e_{A,B} = e_{A,4}/e_{B,4} \tag{4}$$

Tessellations A and B for a region may be compared by Eq. (4), even though the same region may not be tessellated by square SRS of different areas, as those used in A and B.

Simple to complex patterns are studied as time progressed, see, e.g., Devlin (2001). While studying patterns of regular shapes in geometry, the simplest figure is an equilateral triangle, where the sides are all equal and the angle of each vertex is 60°. Then comes a square, followed by a regular pentagon (a 108° angle between touching edges), a regular hexagon, and so on. One may plot the gain in above-mentioned efficiency when geometrical shapes are sequentially considered from edge 3 onwards, see Fig. 1; the curve is drawn in SPLUS using cubic spline (with smoothing parameter $n = 300$). *The boundary of the convex hull of points $e_{i,4}$ considered in* (3) *is obtained by joining these points by straight lines.* The spline curve mimics this boundary of convex hull.

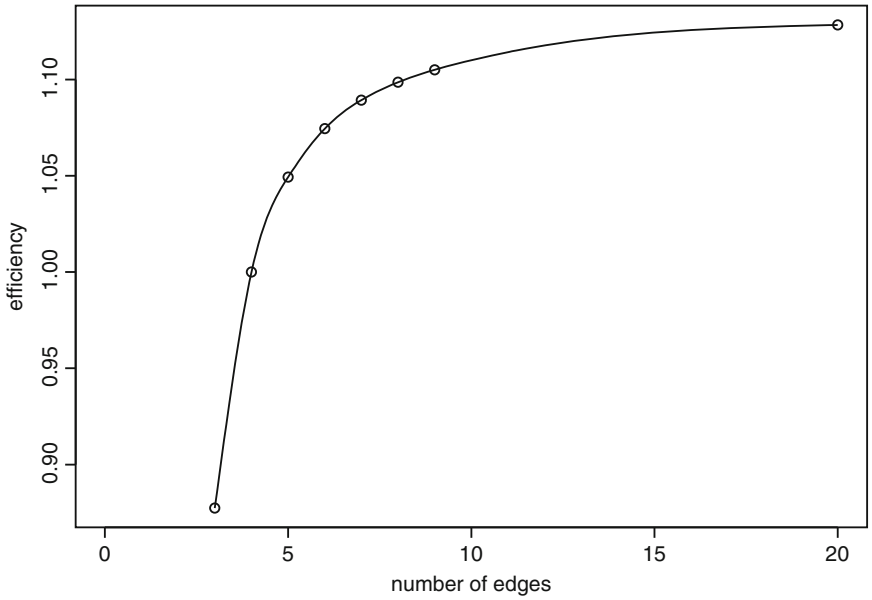


Fig. 1 Efficiency of SRS with respect to square. The above growth curve explains that efficiency of SRS increases with increase in equivalent number of edges. The resultant SRS approaches to a circular shape to achieve the limiting efficiency, as stated in Proposition 1

The y coordinate on the curve in Fig. 1 represents the collection of efficiency $e \in [0.8773827, 1.128379]$ for different *assembly of SRS* as mentioned on the r.h.s. of Eq. (3). The x coordinate corresponding to this value of y on the curve may be interpreted as the "equivalent" number of edges for such an assembly of SRS giving rise to this particular efficiency for the least value of the "number of edges". This "equivalent" edge number may contain fraction, much like a fractal dimension. This summary characteristic represents the closeness of the assembly (in terms of the efficiency) to a single number for the edge of SRS. The curve of Fig. 1 explains growth in efficiency in terms of equivalent edge number.

Denote $\mathbf{n} = (n_3, n_4, n_5, \cdots)$; we show that for an assembly of SRS the equivalent edge number $m = m(\mathbf{n})$ is monotone in the following sense.

Proposition 3. *Let the number of regular polygon with i-edge be $n_i, i = 3, 4, 5\cdots$ used for a general polymorph tessellation A of a two dimensional area. Let there be another tessellation B with SRS of similar areas and there exists an integer $p \geq 3$, such that the number of regular polygons with i-edge $n_i^{'}, i = 3, 4, 5\cdots$ satisfies $n_i^{'} \leq n_i$ for $3 \leq i \leq p$ and $n_i^{'} \geq n_i$ for $i > p$. Then considering only the main terms of (3), the corresponding equivalent edge numbers for the assembly A and B satisfy*

$$m \leq m^{'} \tag{5}$$

where $m = m(\mathbf{n}), m' = m(\mathbf{n}')$. The inequality in (5) is strict unless $\mathbf{n} = \mathbf{n}^{'}$.

Proof. This follows from the fact that from Eq. (3), $e_{A,4} \leq e_{B,4}$. The curve in Fig. 1 is monotone and strictly increasing as the quantities $e_{i,4} \uparrow$, as $i \uparrow$. Note that the x coordinate corresponding to the value of y on the boundary of the convex hull is defined as the "equivalent" number of edges for an assembly of SRS giving rise to this particular efficiency for the least value of the "number of edges".

In three dimensions, a polyhedron which is capable of tessellating space is called a space-filling polyhedron. Examples include the cube, rhombic dodecahedron and truncated octahedron. There is also a 16-sided space-filler and a convex polyhedron known as the Schmitt–Conway biprism which fills space only aperiodically, making a nonperiodic structure. In three dimensions the optimal solution is sphere, as seen from inequality (2). Thus the SRS that is nearer to spherical shape has advantage over other SRS of same volume.

2.2 Generalisations

Assumption made that the cracks are contained in a single SRS when bullet hits inside an SRS may be relaxed. Even then SRS with nearly circular shape has advantage. A circular region is severely affected surrounding the point of bullet hit from which lateral cracks propagate. Consider the maximum length of propagating lateral cracks in a hit. The distribution of the maximum of a random number of random variables is an extreme value distribution under mild assumptions, e.g., see Galambos (1973). The angle θ of direction of the maximum length may be measured w.r.t. a fixed coordinate system, θ may be taken to be uniform $(0, 2\pi]$. It therefore follows that to contain a crack of maximum-length inside an SRS assembled, the SRS has to be of nearly circular shape; as the direction is uniform.

Consider $W_{\nu_n} = \max\{X_1, X_2, \cdots, X_{\nu_n}\}$ to be the maximum of different crack lengths originating from a single hit, where the number of cracks ν_n is random. From Theorem 1 of Galambos (1973), under certain regularity condition, one may write $\lim_{n\to\infty} P[(W_{\nu_n} - b_{\nu_n})/a_{\nu_n} < x] = e^{-w(x)}$, where $w(x)$ can have three specific forms corresponding to three extreme value distributions.

2.3 An Example

In Fig. 2, we plot in X axis the length of 12 linear cracks on glass due to a sudden hit in a car shield, versus normal quantile plot in Y axis. Approximate linear relationship is observed with a high value of $r = 0.980125$, where r^2 is coefficient of determination, indicating that crack length may be approximated by a normal random variable, leading to $w(x) = e^{-x}$ in the limit law of maximum length, when total number of cracks ν_n is random.

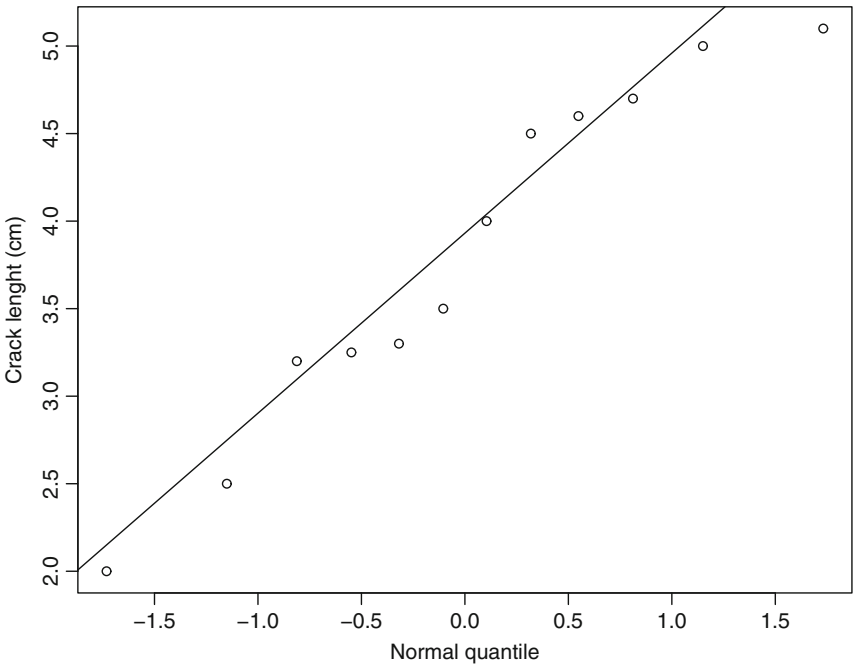


Fig. 2 Normal quantile vs. crack length quantile. Length of 12 glass cracks due to sudden hit versus normal quantile plot indicates that the distribution of crack length may be approximated by a normal random variable

2.4 Superiority of Nearly Circular SRS

One may probabilistically approximate number of affected SRS from the constants involved in the model and radius of SRS, r. Assume that the variables X of crack lengths are standardised: $X \to \frac{X-c}{d}$. Let $a_n = \frac{1}{b_n} - \frac{1}{2} b_n (\log\log n + \log 4\pi)$, $b_n = (2\log n)^{-1/2}$.

From (Galambos 1973), $(W_{\nu_n} - a_{\nu_n})/b_{\nu_n} \xrightarrow{d} \exp(-e^{-x}), -\infty < x < \infty$, i.e., $W_{\nu_n} = (2\log \nu_n)^{1/2}(1 + o_p(1))$.

Approximate length of maximum crack is $\ell = ((2\log \nu_n)^{1/2} + c)d(1 + o_p(1))$.

Approximate number of SRS to be replaced in the direction of maximum length is then $\frac{\ell}{2r} = ((2\log \nu_n)^{1/2} + c)\frac{d}{2r}(1 + o_p(1))$.

Since each directional angle is equally likely, the total number N of SRS to be replaced due to crack has an approximate bound,

$N \le (\ell/r)^2(1 + o_p(1))$.

In the case of exponential distribution for length of cracks, similar bounds hold with a different choice of standardising constants a_n, b_n.

Angle made by the crack having maximum length is equally likely in each direction, indicating superiority of nearly circular SRS to minimise replacement cost, even when cracks are likely to affect the adjacent SRS.

2.5 A Measure for Severity of Hit

Maximum length of crack originating from point of hit is a measure of impact severity. Bullet proof glass is usually constructed using polycarbonate, thermoplastic, and layers of laminated glass. A piercing bullet in glass forms a nearly circular hole with a series of outward cracks propagating towards periphery from the centre. High impact at the point of hit makes the cracks look like a continuous curve with spikes around a hole of large perimeter at centre. Dissecting the formed hole at a point on the rim and then by straightening the perimeter of length T along the X-axis, one may visualise the cracks at each point $t \in [0, T)$ on rim along Y-axis. Successive crack magnitudes may be (weakly) correlated. Figure 2 indicates these may be Gaussian. Justification of quantile plot to assess normal distribution for weakly correlated process is made in Dasgupta (2013). In the following we state a modification of relevant result. The proof is similar.

Theorem A. *Consider a Gaussian process* $X(t),\ 0 \le t \le T$ *with mean* $m(t)$ *and covariance kernel* $\sigma(t,u) = \sigma(t)\sigma(u)\rho(t,u)$, *where* $m(t) \to 0,\ \sigma(t) \to \sigma;\ t \to \infty$. *Assume* $X(t)$ *has the weak limit denoted by* $X(\infty)$ *and the correlation function* $|\rho(t,u)| < K|t-u|^{-\beta},\ K > 0,\ \beta > 0$. *Consider the empirical distribution function of the process based on the observations at time points* $t_1, t_2, \cdots, t_n$ *which are not necessarily equispaced. Let the time interval* $[0, T)$ *of recording the observations be subdivided into k subintervals and the length of all except finitely many subintervals and the number of observations in each subinterval, except finitely many increase to* ∞. *Also let the time gap between two consecutive observations within each subinterval be homogeneous and the number* n^* *of "isolated" observations which do not fall in any one of the homogeneous subintervals, be negligible compared to* n, *i.e.,* $n^* = o(n)$. *Then the empirical distribution function of the recorded observations from the process is a strongly consistent estimate for distribution function of the limiting variable* $X(\infty)$, *as* $n \to \infty$.

Bullet hit may cause extensive damage to a glass shield in which some region projected outward from centre is blown up due to excessive impulse with possible missing observations on crack length. Theorem A applies even in that case. The crack lengths, measured from point of hit may then be examined for modeling by Ornstein–Uhlenbeck process $V(s)$, a stationary continuous Gaussian process with exponentially decaying autocorrelation function.

The Ornstein–Uhlenbeck (O–U) process is continuous, strongly Markov, strictly stationary and Gaussian. Apart from some pathological examples, the above properties *characterise* O–U process. In the present case, the crack lengths are Gaussian, the variables may be taken stationary due to arbitrary choice of axis in

polar coordinates, with origin at the point of hit. Markov property is a simplifying assumption in view of the fact that an in-between crack may deter a crack to influence another crack. High pressure exerted from bullet hit causes material grains to displace and oscillate around its steady state resulting in crack formation.

Ornstein–Uhlenbeck process satisfies the following differential equation.

$$dV(s) = -\beta V(s)ds + \gamma dB(s),\ \beta > 0,\ \gamma > 0 \tag{6}$$

where $B(s)$ is the standard Brownian motion, γ is the spread parameter, β is the drift parameter; $\beta V(s)$ is a restoring force directed towards origin proportional to the distance $V(s)$.

The parameter γ in (6) reflects the magnitude of hit, parameter β is related to the resistance of glass to minimise crack due to bullet hit.

Using the relationship $V(s) = e^{-\beta s}B[\gamma^2(e^{2\beta s} - 1)/2\beta]$, see, e.g., Karlin and Taylor (1981), one may write

$$\overline{\lim}_{t\to\infty}[\frac{\gamma^2}{\beta}(1 + o(1))\log t]^{-1/2} \sup_{0\le s\le t} \mid V(s) \mid = 1,\ \text{a.s.} \tag{7}$$

From (7), the value of $(2\log t)^{-1/2}\sup_{0\le s\le t} \mid V(s) \mid$ may be taken as an estimate of $\gamma/(2\beta)^{1/2}$, which equals to the standard deviation of the Ornstein–Uhlenbeck process. In the present case, t represents the rim-perimeter of holes, on which the cracks are formed. The crack lengths, caused by bullet hit, over the rim of pierced hole at different points $s,\ 0 \le s < t$, are modeled by the Gaussian process $V(s)$. Intensity of hit may cause formation of several layers of ring-cracks propagating from the point of hit, see Fig. 3. These curves indicate propagation of impact as distance from centre increase.

Cracks formed in hit 1 and 2 may be compared by the ratio of maximum crack lengths occurring in two occasions, with associated index

$[\{\frac{\gamma_1^2}{\beta_1}\log t_1\}/\{\frac{\gamma_2^2}{\beta_2}\log t_2\}]^{1/2}$ for relative severity of damages, where β_i, γ_i, t_i represent the drift, spread and hole perimeters, respectively, in occasion $i = 1, 2$.

If the hole perimeters are same to a first approximation, $t_1 \approx t_2$, then the ratio of maximum crack lengths is an estimate of ratio of two asymptotic standard deviations $\gamma/(2\beta)^{1/2}$ for the associated processes. Applications of O-U process in industrial context are also made in Dasgupta (2006) and Dasgupta (2011).

In the following we consider damage assessment in an agricultural experiment on yam. The data arise from a structured layout of seed weight and seed skin texture.

3 Damage Assessment in a Designed Growth Experiment

Growth model experiments in Graeco–Latin square design with Elephant foot yam are conducted in Indian Statistical Institute, Giridih farm to examine inter alia the effect of seed weight and seed skin texture on yield. Two Latin squares are

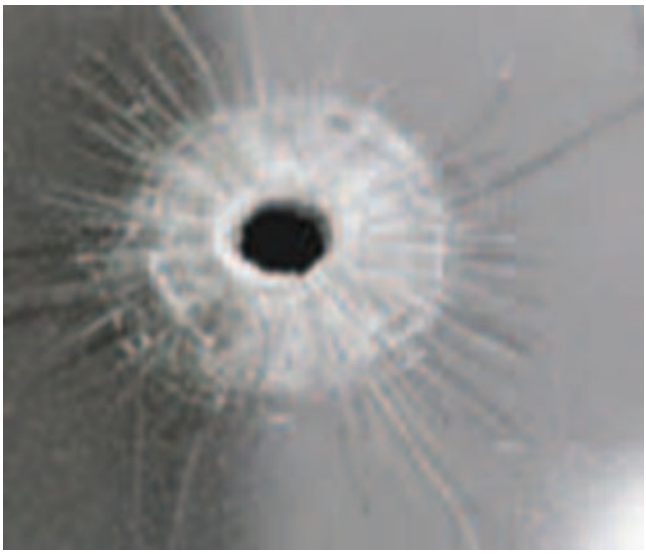


Fig. 3 Bullet hit glass shield. A bullet hit shield shows several layers of circular type damage rings propagating outwards from centre, the point of hit. The lateral cracks of high magnitudes are also seen

orthogonal if the two squares when superimposed have the property that each pair of letter appears only once. The superimposed square is called a Graeco–Latin square. A 5×5 Graeco–Latin square is shown below.

$$\begin{pmatrix} A\alpha & B\beta & C\gamma & D\delta & E\epsilon \\ B\gamma & C\delta & D\epsilon & E\alpha & A\beta \\ C\epsilon & D\alpha & E\beta & A\gamma & B\delta \\ D\beta & E\gamma & A\delta & B\epsilon & C\alpha \\ E\delta & A\epsilon & B\alpha & C\beta & D\gamma \end{pmatrix}$$

In the field experiment two characteristics of the planted cut seed of yam corm tested are represented by Latin and Greek letters; viz. weight and surface area, respectively. Here $A = 200, B = 350, C = 500, D = 650, E = 800$ are weights in grams of the seed, and the Greek letters α represents largest and smoothest surface area of the seed, β being the second largest & smoothest and so on; ϵ represents the smallest and roughest surface area.

The above Graeco–Latin square were repeated four times side by side in a block of two squares, and the pit-holes are numbered 1–5 for the first row in reverse direction i.e., as $E\epsilon, D\delta, C\gamma, B\beta, A\alpha$, in the first experiment. The numbering goes on in the same row of adjacent second design as 6 to 10, and so on. Finally, the pit holes are numbered 91–95 for the last row of the 2×2 block viz., $D\gamma, C\beta, B\alpha, A\epsilon, E\delta$, in the third experiment; and the numbering goes on in the same row i.e., the 10-th row with above combination of alphabets in adjacent design number 4, as 96–100.

White ant and fungal infection are two main causes that hampers the sprouting of seed corm, and these may percolate to geographically nearest neighbours from an infected pit.

The following data relates to weights in kilogram of 100 yams from a growth experiment conducted in the year 2010 at Indian Statistical Institute, Giridih farm in the above-mentioned serial order from 1 to 100 (from start, left to right; row wise). Data structure presented below mimics the actual layout adopted in the yam experiment on field. We wish to see the observed efficacy of the design for damage control with respect to random uniform weight and skin structure of the seed corms allotment.

4.50, 3.20, 2.60, 3.15, 2.05, 2.10, 2.65, 0.80, 1.70, 1.15,
2.90, 3.50, 4.35, 3.85, 3.60, 1.30, 2.20, 1.70, 3.70, 2.50,
3.40, 3.10, 4.45, 5.60, 4.15, 1.50, 1.90, 2.00, 3.10, 3.00,
3.10, 2.25, 2.65, 2.90, 3.60, 1.50, 1.20, 0.70, 2.80, 2.70,
3.75, 2.05, 1.60, 1.50, 3.60, 2.20, 1.40, 1.20, 0.00, 2.40,
2.50, 1.45, 1.05, 0.70, 0.00, 2.25, 2.00, 2.45, 1.55, 0.90,
0.75, 2.65, 2.25, 1.20, 2.25, 2.00, 3.80, 3.00, 3.00, 2.35,
1.05, 0.80, 3.80, 2.30, 3.80, 1.60, 0.00, 3.60, 1.60, 4.00,
3.00, 1.95, 2.00, 3.65, 3.60, 1.40, 1.40, 1.30, 3.90, 3.60,
5.50, 2.90, 2.60, 1.70, 2.80, 1.90, 1.70, 1.80, 1.10, 2.80.

Observe that the yield is nil for pit number 49 with combination $A\epsilon$, pit number 55 with combination $A\alpha$, and pit number 77 with combination $A\gamma$. For pit number 49 with nil yield, adjacent one-step pits surrounding it diagonally and sidewise are having yield as (0.70, 2.80, 2.70, 1.20, 2.40, 2.45, 1.55, 0.90). Simultaneous damage due to fungal & white ant infection from one pit to adjacent pits in semi-porous land stretch with heterogeneous soil structure at ISI Giridih farm land is possible. Such damages can be assessed by digging up the corms when no sprouting was observed after a sufficient time-gap. For pit number 49, 55 and 77 damage did not percolate to adjacent pits. This observation is similar to the assumption made for SRS in tessellation: only one SRS (pit) is affected when the hit (infection) is not on the periphery.

The average seed weights in these infected pits is A, i.e., 200 g, the average skin texture is the middle rank i.e., γ.

Expected damage putting uniform weight to all pits, given that in total three pits are infected, is the mean weight 500 g with middle rank of skin texture γ. This is because the average is taken over all possible combination in a symmetrical manner in the adopted design. Observed damage is less than the expected damage in the experiment conducted.

Similar case of damage propagation may arise in three dimensions e.g., in packing of fruits, when items may be infected by immediate neighbours, and assessment may be made in a similar fashion.

4 A Three Dimensional Structure and Different SRS Combinations

The three dimensional structure of honeycomb with partly elliptical cross section consists of two sided layers of hexagonal cells. These layers are separated by a thin membrane. The backside layer of hexagonal cells (i.e., cells having hexagonal cross section) are so arranged as to keep the honeycomb sturdy, when assembled with front side layer of hexagonal cells. Each cell on backside has edges passing through centre of the circumscribing circle of hexagonal cell in the front, the outward corner points of both sides do meet at the common edge points. Such structure of tessellation by hexagonal SRS makes honeycomb damage resistant on both sides.

For covering a plane area, regular hexagonal SRS are seen to be cost efficient compared to triangular and square SRS, when replacement due to accidental damage is of concern. Tessellation by SRS combinations of different sizes and shapes are common. While covering with different SRS, gain in efficiency with respect to square SRS of same sizes is the weighted average of efficiencies for individual SRS weights depending on the number of SRS of particular types. "Equivalent edge number" represents the property of an ensemble of different SRS in terms of a single edge number that may contain fraction, this index is much like a fractal dimension representing nearness of a self-similar pattern.

Acknowledgements Thanks are due to Dr. Avinash Dharmadhikari, Tata Motors for suggesting the problem and interesting discussions.

References

Branko G, Shephard GC (1987) Tilings and patterns. W.H. Freeman, New York

Dasgupta R (2006) Modeling of material wastage by Ornstein-Uhlenbeck process. Calcutta Stat Assoc Bull 58:15–35

Dasgupta R (2011) On the distribution of burr with applications. Sankhyā B 73:1–19

Dasgupta R (2013) South pole ozone profile and lower tolerance limit, Chap 8. In: Advances in growth curve models: topics from the Indian Statistical Institute. Springer proceedings in mathematics & statistics, vol 46. Springer, New York, pp 149–170

Devlin KJ (2001) The math gene: how mathematical thinking evolved and why numbers are like gossip. Basic Books, New York

Galambos J (1973) The distribution of the maximum of a random number of random variables with applications. J Appl Probab 10:122–129

Hirschhorn MD, Hunt DC (1985) Equilateral convex pentagons which tile the plane. J Comb Theory Ser A 39(1):1–18

Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic, London

Ledoux M, Talagrand M (2002) Probability in Banach spaces. Springer, New York

Osserman R (1986) A survey of minimal surfaces. Dover, New York

Steinhaus H (1999) Mathematical snapshots. Dover, New York