JOURNAL
OF THE
INDIAN SOCIETY
OF
AGRICULTURAL STATISTICS

भारतीय कृषि सांख्यिकी संस्था की पत्रिका

ISAS

April 2009 **VOL. 63** No. 1

INDIAN SOCIETY OF AGRICULTURAL STATISTICS

OFFICE BEARERS, 2009

President : Dr. Mangala Rai

Executive President : Prof. Prem Narain

Sessional President : Dr. V.K. Bhatia

Patrons

Prof. J.S. Rustagi
Dr. K.C. Seal
Dr. K.C. Raut*
Dr. M.N. Das
Dr. B.N. Singh
Shri J.S. Sarma
Prof. Bal B.P.S. Goel*

Vice Presidents

Prof. Aloke Dey
Dr. S.M. Jharwal
Dr. V.K. Gupta
Dr. S.D. Sharma

Members

Dr. H.V.L. Bathla
Shri S.C. Rai
Dr. B.V.S. Sisodia
Dr. R.C. Agrawal
Dr. S.N. Megeri
Shri V.K. Singh
Dr. D.K. Jain
Dr. Hukum Chandra
Dr. S. Ravichandran
Dr. B.S. Kulkarni
Dr. V.K. Mahajan
Dr. Kishore Sinha

Honorary Secretary : Dr. V.K. Bhatia

Honorary Joint Secretaries : Shri R.S. Khatri
Dr. Rajender Parsad
Dr. P.K. Malhotra

Honorary Treasurer : Dr. A.K. Vishandass

Ex-Officio Members (By Designation) :

1. Director, Indian Agricultural Statistics Research Institute, New Delhi.
2. Economic & Statistical Adviser, Ministry of Agriculture, Krishi Bhavan, New Delhi.
3. Director General, Central Statistical Organisation, Government of India, New Delhi.
4. Director General, National Sample Survey Organisation, Government of India, New Delhi.
5. Chairman, Editorial Board, JISAS, New Delhi.

* Nominated to attend the Executive Council Meetings during 2009.

CONTENTS

62nd Annual Conference

Agricultural Statistics: Theory and Applications

Computer Applications

Book/Monograph Review

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 1-23

Some Aspects of Estimating Poverty at Small Area Level[*]

A.K. Srivastava
Former Joint Director, Indian Agricultural Statistics Research Institute, New Delhi

SUMMARY

Issues relating to poverty have been at the core of development process in all developing countries. Measurement of poverty has been extremely important for evaluation of development strategies. Disparities do exist in the income as well as consumption and expenditure levels in different groups of society as also there are spatial dispersions. There are indicators for measuring incidence, depth and severity of poverty. Most of these indicators are estimated at State level with the help of data as obtained from Consumption Expenditure Surveys. For poverty alleviation programmes, as well as for planning other development strategies at micro-level, small area level estimates for poverty indicators are necessary. In this paper, an attempt is made to review some of the existing procedures for poverty mapping and an application of a Small Area Estimation technique is made for estimating poverty indicators at district level in Uttar Pradesh. Data from Consumption Expenditure Survey of NSSO (61st round 2004-05) has been used for this application.

Key words: Poverty indicators, Small Area Estimation, Area level models.

1. INTRODUCTION

Measurement of poverty and its estimation has been at the center stage of the planning process in every developing country. Household surveys for consumption expenditure have been main instruments of poverty measurement. These surveys are facilitated in most of the developing countries by United Nations through the National Household Survey Capability Programmes. But, India has been in an advantageous position with availability of regular data flow through National Sample Survey Organization (NSSO). As a part of its national level household surveys, Consumption Expenditure surveys are conducted every five years on a larger sample and annually on a thinner sample.

Poverty is commonly visualized as a state of not having enough resources to take care of basic needs such as food, clothing, and housing. The criterion developed for measurement of poverty revolves around quantification of minimum (food and non-food) requirements of individuals for a healthy living. The monetary value for such a requirement is termed as poverty line. Poverty is also sometimes defined as the state of living in a family with income below the defined poverty line. Poverty lines are obtained at the state levels with rural-urban classifications.

One of the most commonly used indicators of poverty is the poverty ratio which is a Head Count Ratio (HCR) of poor people and it measures the incidence of poverty. There are other indicators as well, which measure the depth and severity of poverty. In the Indian context, state-wise poverty lines are defined for rural and urban areas and are updated to take care of price changes. Poverty ratios for the States are estimated periodically using the current poverty lines and the NSSO data for Household Consumer Expenditure Surveys.

[*] *Technical Address presented during the 62nd Annual Conference of the Indian Society of Agricultural Statistics held at S.V. Agricultural College (ANGRAU), Tirupati during 24-26 November 2008.*

Poverty estimation at small area levels is a practical necessity in view of growing needs for micro level planning. Presently, estimates for number of poor as well as for poverty ratios are provided only at state level. In many countries, poverty mapping is done, based on small area level estimates. In the Indian context, even district level precise estimates are not available. Direct estimates, based on NSSO data are likely to be less precise due to smaller sample sizes at district level.

Several other aspects of poverty estimation, such as measuring the incidence, depth and severity of poverty, inequalities as well as distribution of poverty to different groups of population at district level are of interest. In this paper, some of these aspects are addressed and illustrated with the help of NSSO, 61st round (2004-05) data for Consumer Expenditure for Uttar Pradesh.

2. MEASURING POVERTY

The process for measuring poverty in the country was initiated in early sixties (1962), when a working group from Planning Commission provided a quantification of minimum (food and non food) requirements of individuals for a healthy living. At that time a poverty line was set up for national level at Rs. 20 per month at 1960-61 prices. In the late 70’s and early 80’s, several methodological issues relating to poverty estimation were sorted out. A weighting diagram was developed, with due consideration to age, sex and the nature of work performed by individuals. Subsequently, a poverty line was specified by the Planning Commission, Government of India in 1979. Using the calorie norms for individual groups, the average requirements in rural and urban areas were obtained as 2435 and 2095 K cal respectively. These were further rounded off to 2400 and 2100 K cal respectively. Based on NSSO 28th round (1973-74) data and the corresponding prices in the same year for a consumption basket satisfying the above calorie norms, monthly per capita consumption expenditures were worked out as Rs. 49.09 and Rs 56.64 for rural and urban areas respectively. These were poverty lines at the national level at that time i.e. in 1979.

An Expert Group on Estimation of Proportion and Number of Poor was constituted by the Planning Commission in 1989 which recommended that the above poverty lines should be adopted as the base line and it should be uniformly adopted for all the States. The Expert Group also suggested an approach for disaggregating the national level poverty lines to State specific poverty lines on the basis of State specific prices and inter-state price differential. The group also suggested a mechanism for regularly updating the poverty lines for rural and urban areas on the basis of prices using the Consumer Price Index of Agricultural Laborers (CPIAL) for rural poverty line and Consumer Price Index for Industrial Workers (CPIIW) for urban poverty line. Presently, the Expert Group method is being used for estimating the number and percentage of poor at national and State level by the Planning Commission, which is the nodal agency for estimation of poverty.

The Planning Commission released poverty estimates for 1973-74, 1977-78, 1983, 1987-88, 1993-94, 1999-2000 and 2004-05. It has, however, been noted that 1999-2000 estimates are not strictly comparable with the estimates of previous years due to different reference periods used in 55th round (1999-2000) of the NSSO consumer expenditure survey. Choice of different reference periods for different group of items has been a cause of considerable debate and experimentation in some of the previous rounds of NSSO surveys. The major concerns have been the quality of data and comparability of results. In the 61st round data (2004-05), two different consumption distributions have been obtained. The first one relates to 30-day recall period for all the items. The other one relates to data collected using 365-day recall period for five infrequently purchased non-food items i.e. clothing, footwear, durable goods, education and institutional medical expenses and 30-day recall period for remaining items. The two consumption distributions have been termed as Uniform Recall Period (URP) and Mixed Recall Period (MRP) consumption distributions respectively. For 2004-05 data, poverty estimates for both the distributions have been obtained. However, it is the URP approach which is comparable to previous results.

The debate on methodological issues for measurement of poverty has passed through many critical stages. In the early stages, national poverty line was uniformly followed for all the States. Also there were significant differences in the consumption expenditure data as obtained from NSSO, and the other from National Accounts Statistics (NAS). For quite some time the NSSO results used to be adjusted pro-rata with the NAS results. However, with the recommendations of the Expert group and with the introduction of State- specific poverty lines, the practice of adjustment in urban areas was discontinued. Sudden decline in the poverty estimates in 1999-2000 from 37.24 per cent to 27.1 per cent in the rural areas and from 32.4 per cent to 23.6 per cent in urban areas was partially attributed to the change in the reference periods. Some of these issues are discussed in detail in a collection of papers in *The Great Indian Poverty Debate, eds. Deaton and Kozel.*

3. INDICATORS OF POVERTY AND ECONOMIC INEQUALITY

Some of the commonly used indices for studying poverty and income inequality are as follows:

3.1 Poverty Ratio

As already mentioned earlier, most commonly used indicator of poverty is the poverty ratio which is a head count ratio (HCR) of poor people and it measures the incidence of poverty. Define

x^* = poverty line

x_i = monthly per capital consumption expenditure of i^{th} individual

N = total number of persons

P = number of persons with consumption expenditure less than x^*.

Poverty Ratio (PR) is defined as

$$\text{PR} = P/N$$

Poverty ratio is, thus, simply a head count ratio and it only measures the incidence of poverty. It is most commonly used measure of poverty globally. However, a major limitation of this index is that it does not take into account the level of poverty within poor people. Poverty ratio is not affected by upward or downward movement of poor people unless they cross the poverty line.

3.2 Income Gap Ratio

Income Gap Ratio (IGR) is defined as

$$\text{IGR} = \frac{1}{P}\sum_{i=1}^{P}\left(1 - x_i / x^*\right) \text{ for all } x_{\text{i}} < x^*$$

$$= 1 - \bar{x}_p / x^*$$

where $\bar{x}_p$ is the average consumption of the persons below poverty line. IGR provides information on the depth of poverty. It captures the average expenditure shortfall, or gap, for the poor in a given area to reach the poverty line. Another indicator for measuring the depth of poverty is Poverty Gap Ratio, which is defined as follows.

3.3 Poverty Gap Ratio

Poverty Gap Ratio (PGR) is defined as

$$\text{PGR} = \frac{1}{N}\sum_{i=1}^{P}(1 - x_i / x^*)$$

$$= \frac{P}{N}\frac{1}{P}\sum_{i=1}^{P}\left(1 - x_i / x^*\right)$$

$$= (\text{PR}) \cdot (\text{IGR})$$

This is an improved indicator for measuring the depth of poverty. It is more consistent with poverty ratio regarding persons crossing the poverty line.

3.4 Squared Poverty Gap Ratio

The Squared Poverty Gap Ratio (SPGR) is a measure for severity of poverty and is defined as

$$\text{SPGR} = \frac{1}{N}\sum_{i=1}^{P}\left(1 - x_i / x^*\right)^2$$

3.5 Foster-Greer-Thorbecke (FGT) Index

A generalized version of poverty indices was considered by Foster *et al*. (1984) as follows:

$$\text{FGT} = P_\alpha\left(x, x^*\right) = \frac{1}{N}\sum_{i=1}^{P}\left(1 - x_i / x^*\right)^\alpha$$

$$= \text{PR} \qquad \text{when } \alpha = 0$$
$$= \text{PGR} \qquad \text{when } \alpha = 1$$
$$= \text{SPGR} \qquad \text{when } \alpha = 2$$

This measure becomes more and more sensitive to poorer persons with higher values of α. Thus, it becomes a good indicator for more vulnerable poorest of the poor classes of society. An interesting feature of FGT is that it is decomposable to different components of the population. If the population is divided into g groups of households with ordered income vectors $x^{(j)}$ and population sizes N_j, then for $\alpha > 1$

$$P_\alpha(x, x^*) = \sum_{j=1}^{g} \frac{N_j}{N} P_\alpha\left(x^{(j)}, x^*\right)$$

If we define $T_j = \frac{N_j}{N} P_\alpha\left(x^{(j)}, x^*\right)$ and $T = \sum_j T_j$, then percentage contribution of j^{th} group to overall poverty is

$$\text{CTP}_j = (T_j / T) \times 100$$

Thus, FGT is an important tool for measuring the contribution to total poverty for various subgroups of the population.

3.6 A Measure of Income Inequality – Gini Coefficient

One of the most widely used measures for the extent of inequality is the Gini Coefficient. An important feature of this measure is its association with Lorenz curve in which the proportion of the population arranged from the poorest to the richest are represented on the horizontal X-axis and the proportion of income enjoyed by the bottom x proportion of the population is depicted on the vertical Y-axis. The mathematical formulation is as follows:

Let the income $y(\geq 0)$ has a continuous distribution with density function $f(y)$ with mean $\mu = \int_0^\infty yf(y)\,\mathrm{d}y$ Define

$$\text{F}(x) = \int_0^x f(y)dy \text{ and } \text{F}_1(x) = \frac{1}{\mu}\int_0^x yf(y)\mathrm{d}y$$

F(x) is the proportion of persons with income less than or equal to x and F$_1$(x) is the proportionate share of these persons in the aggregate income of all persons. Clearly, F(x) and F$_1$(x) both lie between 0 and 1 for x ranging from 0 to ∞ and F$_1$ is a monotone increasing function of F. The graph of F$_1$ against F is called Lorenz curve or the Concentration curve of the given distribution of income. The area between the Lorenz curve and the egalitarian line is called the area of concentration. Lorenz ratio, also known as the Gini coefficient is defined as

$$\text{G} = 2 \times \text{area of concentration}$$

$$= 1 - 2\int_0^1 \text{F}_1 \mathrm{d}\,\text{F}$$

G may also be represented in several alternative ways. Some of the representations and corresponding interpretations in terms of welfare economics are available in literature (Sen 1973).

4. POVERTY ESTIMATION

As mentioned earlier, poverty ratios are estimated for each State (rural and urban) as percentage of persons below respective poverty lines and then a pooled poverty ratio is obtained for the State by combining rural and urban estimates. The all India poverty ratio is obtained as a weighted average of state-wise poverty ratios. These estimates are available since 1973-74 (base year) and then for the years corresponding to five yearly larger samples for Consumption Expenditure Surveys of the NSSO. Since a time series data on poverty is available in the country, comparisons of poverty over space and time makes an interesting study. Debates and controversies on poverty estimates have also been a part of the entire process. Some of the issues have been discussed in *The Great Indian Poverty Debate eds. Deaton and Kozel (2005)*.

To get an idea about, how the poverty lines and poverty ratios have moved over the years, State specific poverty lines and poverty ratios for 16 States at different period of times are given in Table 1 and Table 2 respectively.

From these tables, it is possible to study the spread of poverty over the States and also the trends and changes

Table 1. State-specific poverty lines (Rupees per capita per month)

S. No.	State	Rural			Urban		
		1973-74	1993-94	2004-05	1973-74	1993-94	2004-05
1	Andhra Pradesh	41.71	163.02	292.95	53.96	278.14	542.89
2	Assam	49.82	232.05	387.64	50.26	212.42	378.84
3	Bihar	57.68	212.16	354.36	61.27	238.49	435.00
4	Gujarat	47.10	202.11	353.93	62.17	297.22	541.16
5	Haryana	49.95	233.79	414.76	52.42	258.23	504.49
6	Himachal Pradesh	49.95	233.79	394.28	51.93	253.61	504.49
7	Karnataka	47.24	186.63	324.17	58.22	302.89	599.66
8	Kerala	51.68	243.84	430.12	62.78	280.54	559.39
9	Madhya Pradesh	50.20	193.10	327.78	63.02	317.16	570.15
10	Maharashtra	50.47	194.94	362.25	59.48	328.56	665.90
11	Orissa	46.87	194.03	325.79	59.34	298.22	528.49
12	Punjab	49.95	233.79	410.38	51.93	253.61	466.16
13	Rajasthan	50.96	215.89	374.57	59.99	280.85	559.63
14	Tamil Nadu	45.09	196.53	351.86	51.54	296.63	547.42
15	**Uttar Pradesh**	**48.92**	**213.01**	**365.84**	**57.37**	**258.65**	**483.26**
16	West Bengal	54.49	220.74	382.82	54.81	247.53	449.32
	All India	49.63	205.84	356.30	56.76	281.35	538.60

Source: Planning Commission, Government of India

Table 2. State-specific poverty ratios (Percentage)

S. No.	State	Rural			Urban			Combined		
		1973-74	1993-94	2004-05	1973-74	1993-94	2004-05	1973-74	1993-94	2004-05
1	Andhra Pradesh	48.41	15.92	11.2	50.61	38.33	28.0	48.86	22.19	15.8
2	Assam	52.67	45.01	22.3	36.92	47.73	3.3	51.21	40.86	19.7
3	Bihar	62.99	58.21	42.1	52.95	34.50	34.6	61.91	54.96	41.1
4	Gujarat	46.35	22.18	19.1	52.57	27.89	13.0	48.15	24.21	16.8
5	Haryana	34.23	28.02	13.6	40.18	16.38	15.1	35.36	25.05	14.0
6	Himachal Pradesh	27.42	30.34	10.7	13.17	9.18	3.4	26.39	28.44	10.0
7	Karnataka	55.14	29.88	20.8	52.53	40.14	32.6	54.47	33.16	25.0
8	Kerala	59.19	25.76	13.2	62.74	24.55	20.2	59.79	25.43	15.0
9	Madhya Pradesh	62.66	40.64	36.9	57.65	48.38	42.1	61.78	42.52	38.3
10	Maharashtra	57.71	37.93	29.6	43.87	35.15	32.2	53.24	36.86	30.7
11	Orissa	67.28	49.72	46.8	55.62	41.64	44.3	66.18	48.56	46.4
12	Punjab	28.21	11.95	9.1	27.96	11.35	7.1	28.15	11.77	8.4
13	Rajasthan	44.76	26.46	18.7	52.13	30.49	32.9	46.14	27.41	22.1
14	Tamil Nadu	57.43	32.48	22.8	49.40	39.77	22.2	54.94	35.03	22.5
15	**Uttar Pradesh**	**56.53**	**42.28**	**33.4**	**60.09**	**35.39**	**30.6**	**57.07**	**40.85**	**32.8**
16	West Bengal	73.16	40.80	28.6	34.67	22.41	14.8	63.43	35.66	24.7
	All India	56.44	37.27	28.3	49.01	32.36	25.7	54.88	35.97	27.5

Source: Planning Commission, Government of India over time. However, for micro level planning, estimates at district and even smaller levels are more meaningful.

5. ESTIMATION OF POVERTY AT DISTRICT LEVEL

For poverty estimation at micro-level, various approaches are available. One such approach, which is commonly used for poverty mapping, is essentially based on unit level data from surveys and corresponding census approximately of the same period. The approach is briefly described as follows.

5.1 Poverty Mapping

For poverty mapping, micro-level estimates of poverty parameters are needed. One of the methods, which is based on unit level small area model approach was developed by a group at the World Bank. (Hentschel *et al.* 2000 and Elbers *et al.* 2001).

The method requires data from a household survey which includes household consumption expenditure data (y). To calculate more specific poverty measures linked to a poverty line, log normal regressions are estimated to model per capita expenditure using a set of explanatory variables (x) that are common to both the household survey and the census (*e.g.* household size, education, housing and infrastructure characteristics and demographic variables).These first stage regression models are modeled at the lowest geographical level for which the household survey data is representative and a different first stage model is estimated for each stratum (*e.g.* region, rural and urban). Next, the estimated coefficients from these regressions are used to predict log per capita expenditure for every household in the census. This household level predicted data are then aggregated to small areas such as sub-counties, counties etc. to obtain estimates of the percentage of households below poverty line. These poverty rates are used to produce poverty maps, showing the spatial distribution of poverty.

This method has been used by the World Bank in a number of developing countries for poverty mapping. To start with, an initial exercise is needed to select the set of common variables in the survey and the census records. In some cases, the distributions of individual concomitant variables are examined with respect to their moments and percentiles for both the survey and census data. The success of method heavily depends on the suitable choice of these variables. One of the limitations of this method is that it requires unit level data for the census. Davis (2003) observes that one virtue of this methodology is the relative ease of checking the reliability of estimates that are built into the programmes provided by the World Bank to national poverty mapping analysis and the other virtues of this approach is that it has the institutional backing of the World Bank and a team of researchers concerned with developing methodology and training.

The method has got its own merits for poverty mapping, provided unit level data for census are available. Since unit level household expenditure values are predicted for each and every household, it is possible to estimate poverty parameters at reasonably smaller levels. In many of the developing countries, estimates at sub county level, which are much smaller than districts, have been obtained for poverty mapping purposes. In the Indian context, however, its application for poverty mapping has not been attempted.

5.2 Small Area Estimation Approach

Small area typically refers to the part of a population for which reliable statistics of interest cannot be produced due to small sample sizes. The topic of small areas has gained importance in view of growing needs of micro level planning. Demands for reliable small area statistics (SAS) are increasing with growing concerns of governments relating to issues of distribution, equity and disparity. The traditional sampling theory fails to provide reliable and valid estimates for small areas. Many Small Area Estimation (SAE) techniques have been developed which make use of information from other sources. They also borrow strength from related or similar areas through explicit and implicit models that connects the small areas via supplementary data. The need for statistics at lower levels has been felt for a long time and efforts have been made to meet the requirements through some traditional approaches. The initial SAE methods were invariably based on certain assumptions in the form of implicit models. These models were, however, subsequently explicitly modeled and a number of model based SAE techniques are now available. We consider some explicit model-based methods which are essentially mixed models and are used in specific situations based on data availability on the response variables of interest. These are (i) area level models where information on response variable is available only at the small area level; and

(ii) unit level models where information on the response variable is available at the unit level.

5.2.1 Area level models

An area level mixed model is represented as

$$\hat{\theta}_d = z_d^T \beta + v_d + e_d , \quad d = 1, ..., D$$

where $\hat{\theta}_d$ is the direct survey estimate of the parameter θ_d, as obtained from the sample survey data, z_d^T is the vector of concomitant variates, the model errors v_d are assumed to be independent and identically distributed with mean zero, variance σ_d^2 and e_d are the sampling errors which are assumed to be independent across small areas with mean zero and known variances χ_d. Here e_d and v_d are design-based and model-based random variables respectively. The model variance σ_d^2 is a measure of homogeneity of the areas after accounting for the covariates z_d.

Empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods have played an important role in the estimation of small area means. EBLUP method has been used in many practical applications.

The other methods EB and HB are applicable under specific distributional assumptions. The inferences in HB methods are obtained through posterior distributions. EBLUP and EB are identical under normality assumptions. For EBLUP and EB, an estimate of Mean Square Error, $\text{MSE}(\tilde{\theta}_d) = \text{E}(\tilde{\theta}_d - \theta_d)^2$ is used as a measure of variability of $\tilde{\theta}_d$, where the expectation is with respect to the model as considered above.

One of the early applications of this method was due to Fay and Herriot (1979). In fact, this method was adopted by the U.S. Bureau of Census in 1974 to form Per Capita Income (PCI) estimates for small places. An excellent example of recent application of this method is in a study on "Small Area Estimates of School-Age Children in Poverty" (Constance and Graham eds. 2000). In this application, estimates for number of school-age children, belonging to households in poverty has been estimated at the county level in USA. These estimates are used for distribution of Title-I funds of the Elementary and Secondary Education Act to Counties for onward distribution to School Districts.

We intend to apply this method for obtaining district level poverty estimates for the area level models and provide the estimation procedure, along with the method of obtaining the estimates of MSEs of estimated parameters.

Procedure for estimation

For the mixed model considered above, an Empirical Best Linear Unbiased Predictor (EBLUP) of θ_d is given by

$$\theta_d^* = \gamma_d \hat{\theta}_d + (1 - \gamma_d) z_d^T \hat{\beta}$$

This estimator is a linear combination of direct estimator $\hat{\theta}_d$ and the regression synthetic estimator $z_d^T \hat{\beta}$. Further γ_d and $\hat{\beta}$ are defined as follows:

$$\gamma_d = \frac{\hat{\sigma}_v^2}{\psi_d + \hat{\sigma}_v^2}$$

and

$$\hat{\beta} = \left[\sum_d^D \frac{z_d z_d^T}{\psi_d + \hat{\sigma}_v^2} \right]^{-1} \left[\sum_d^D \frac{z_d \hat{\theta}_d}{\psi_d + \hat{\sigma}_v^2} \right]$$

Calculation of $\hat{\sigma}_v^2$

One of the methods for obtaining $\hat{\sigma}_v^2$, which is the method of moments and was suggested by Fay and Herriot (1979), is an iterative procedure and is described as follows:

Define

$$h(\sigma_v^2) = \sum_d \frac{(\hat{\theta}_d - z_d^T \hat{\beta})^2}{\psi_d + \sigma_v^2}$$

and

$$h'(\sigma_v^2) = -\sum_d \frac{(\hat{\theta}_d - z_d^T \hat{\beta})^2}{(\psi_d + \sigma_v^2)^2}$$

This is an approximation to the first derivative of $h(\sigma_v^2)$.

The iterative equation is

$$\sigma_v^{2(\alpha+1)} = \sigma_v^{2(\alpha)} + \frac{1}{h'(\sigma_v^{2(\alpha)})}\left[m - p - h(\sigma_v^{2(\alpha)})\right]$$

Constraining $\sigma_v^{2(\alpha+1)} \geq 0$ and taking $\sigma_v^{2(0)} = 0$ when no solution exists for any a.

For iterations, we have to start with $\alpha = 0$ taking $\sigma_v^{2(0)} = 0$ and continue to α = 1, 2, 3, 4, . . . and so on, till the value of σ_v^2 stabilizes. Normally less than ten iterations are needed.

It may be noted that only some of the woredas are represented in the sample. For the woredas not represented in the sample, it would not be possible to develop direct estimators and regression synthetic estimator will be used.

Estimation of MSEs

The procedure for estimating MSEs may be given in two steps.

Step 1. Estimation of MSE (θ_d^*) for small areas which are in the sample

Estimate for MSE for sampled small areas is given by

$$mse(\theta_d^*) = g_{1d}(\hat{\sigma}_v^2) - b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2)\nabla_{g_{1d}}(\hat{\sigma}_v^2) + g_{2d}(\hat{\sigma}_v^2) + 2g_{3d}(\hat{\sigma}_v^2)$$

where

$$g_{1d}(\hat{\sigma}_v^2) = \gamma_d \psi_d$$

$$g_{2d}(\hat{\sigma}_v^2) = (1-\gamma_d)^2 X_d^T \left[\sum_d \frac{z_d z_d^T}{\psi_d + \hat{\sigma}_v^2}\right]^{-1} z_d$$

$$g_{3d}(\hat{\sigma}_v^2) = \psi_d^2(\psi_d + \hat{\sigma}_v^2)^{-3}\bar{\mathrm{V}}(\hat{\sigma}_v^2)$$

where

$$\bar{\mathrm{V}}(\hat{\sigma}_v^2) = 2m\left[\sum_i \frac{1}{\psi_d + \hat{\sigma}_v^2}\right]^{-2}$$

$$b_{\hat{\sigma}_v^2}(\hat{\sigma}_v^2) = \frac{2\left[m\sum_d(\psi_d + \hat{\sigma}_v^2)^{-2} - \left\{\sum_d(\psi_d + \hat{\sigma}_v^2)^{-1}\right\}^2\right]}{\left[\sum_d(\psi_d + \hat{\sigma}_v^2)^{-1}\right]^3}$$

and

$$\nabla_{g_{1d}}(\hat{\sigma}_v^2) = (1-\gamma_d)^2$$

Step 2. Estimation of MSE (A_i^*) for small areas which are not in the sample

For non-sampled small areas, the EBLUP estimator reduces to regression synthetic estimator $z_{d'}^T\hat{\beta}$, where $\hat{\beta}$ is the weighted least square (WLS) estimator computed from the sampled small areas $d \in s$.

A nearly unbiased estimator of MSE for regression synthetic estimator for d'-th non-sampled small area is given by

$$mse(\theta_{d'}^*) = z_{d'}^T\left[\sum_i \frac{z_d z_d^T}{\psi_d + \hat{\sigma}_v^2}\right]^{-1} z_{d'} + \hat{\sigma}_v^2$$

Here the subscripted notation *d* stands for sampled small areas whereas *d′* stands for the non-sampled small areas.

It may be observed that the leading term of $\text{MSE}(\hat{\theta}_d)$ is given by $\gamma_d\chi_d$ which shows that the EBLUP estimate can lead to large gains in efficiency over the direct estimate with variance χ_d , when γ_d is small i.e. the model variance σ_v^2 is small relative to the sampling variance χ_d . Choice of good auxiliary data to provide a good model fit is, therefore the key to successful application of the small area technique.

5.2.2 Unit level models

Consider a population of *N* units with *d*-th small areas consisting of N_d units. Let y_{dj} and x_{dj} be the unit level *y*-value and correlated covariate *x*-value for *j*-th unit in the *d*-th small area. It is assumed that the domain means $\bar{X}_d$ is known. Consider the following one-folded nested error linear regression model

$$y_{dj} = x_{dj}^T\beta + \mathrm{v}_d + e_{dj},\ j = 1,\ldots,N_d;\ d = 1,\ldots,D$$

where the random small area effects v_d have mean zero and common variance σ_v^2 and are independently distributed. Also, e_{dj} are assumed to be independently distributed with mean zero and variance σ_e^2 and are also independent of area effects v_d. This model was initially considered by Battese *et al.* (1988).

If N_d is large, the population mean $\bar{Y}_d$ is approximately equal to $x_d^T\beta + v_d$. The sample data $\{y_{dj}, x_{dj}, j = 1, ... n_d; d = 1, ..., D\}$ is assumed to satisfy the above population model. This happens in equal probability sampling. This will also follow in probability proportional to size sampling when the size measure is taken as the covariate in the model. Assuming $\bar{Y}_d = \bar{X}_d^T\beta + v_d$, the EBLUP estimate of $\bar{Y}_d$ is of the form

$$\bar{y}_d^* = \hat{\gamma}_d[\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T\hat{\beta}] + (1-\hat{\gamma}_d)\bar{X}_d^T\hat{\beta}, d = 1,...,D$$

Here, $\hat{\gamma}_d = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_d^{-1})$ with estimated variance components $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$, and $\hat{\beta}$ is the weighted least square estimate of β. It may be noted that the EBLUP estimator is a composite estimator combining the survey regression estimator with the regression synthetic estimator.

For the sampled data, in this model also, the leading term of MSE ($\bar{y}_d^*$) is given by $\gamma_d(\sigma_e^2 / n_d)$, which shows that EBLUP estimate can lead to large gains in efficiency over the survey regression estimate when γ_d is small. Battese *et al.* (1988) applied the nested error regression model to estimate area under corn and soybeans at county level in North-central Iowa using farm interview data in conjunction with LANDSAT satellite data.

It is seen that although the method used by World Bank is a unit level regression based approach, it is very much different than the mixed model approach described here.

For details of an exhaustive and thorough presentation of small area estimation, an excellent reference is the book by Rao (2003).

6. APPLICATION OF AREA LEVEL MODEL BASED SAE APPROACH FOR ESTIMATING POVERTY INDICATORS

As described earlier, in the Indian context, the main source of data for studying poverty is the Consumption Expenditure Surveys of NSSO. There have been some attempts to apply SAE techniques. Singh *et al.* (2005) used NSS data for application of Spatio-Temporal Models in Small Area Estimation. Sastry (2003) explored the feasibility of using NSS Household Consumer Expenditure Survey Data for estimation of district poverty estimates. The study was, however, confined to examining the distribution of relative standard errors (RSE) of direct estimates for Monthly Per Capita Expenditure (MPCE) and those of the sample sizes at district level as obtained from 55th round of NSS data. This was not an application of any SAE technique, but it was an attempt to obtain the district level estimates, following the usual approach of estimating domain parameters. It was observed that in rural areas, 451 out of 490 districts (92%) are having RSEs less than 5% only. It was also observed that only 2% districts had RSEs of 10% or more. Although this study shows a promise for estimating average MPCE at district level from the sample data, it has little bearing on estimation of poverty indicators. Most of the poverty indicators depend on the estimates of number of persons below poverty line, for which RSEs do not necessarily behave similar to those of estimated MPCEs.

Here, we are trying to illustrate an application of Area Level model based Small Area Estimation approach for estimating some of the poverty indicators at district level for Uttar Pradesh using 61st round of NSSO data (2004-05).

6.1 Some Features of the Data Used

As mentioned earlier, data from 61st round of the NSSO survey (2004-05) for Consumption Expenditure Survey has been used for estimation of district level poverty. Uttar Pradesh is one of the most important States in the country, with population at approximately 1.66 million (2001 Census) and population density as 690 per sq. km. It is also one of the poorest states with poverty ratio at 32.8, whereas national level is 27.5 (2004-05). A review of the nature and evolution of poverty in Uttar Pradesh (Kozel and Parker 2005) makes an interesting study. The state has experienced quite a bit of reorganization of districts in the recent past. Between 1991 and 2001 sixteen new districts have been carved out. A glimpse of reorganized districts is given in Table 3.

Table 3. Reorganization of districts between 1991 and 2001

Name of new district(s), 2001	Districts as in 1991 from which, new districts carved out
Jyotiba Phule Nagar	Moradabad
Baghpat	Meerut
Gautam Buddha Nagar	Ghaziabad, Bulandshahr
Hatharas	Mathura, Aligarh
Kannauj	Farrukhabad
Auraiya	Etawah
Mahoba	Hamirpur
Chitrakoot	Banda
Kaushambi	Allahabad
Ambedkar Nagar	Faizabad, Azamgarh
Shravasti	Bahraich
Balrampur	Gonda
Sant Kabir Nagar	Basti, Siddharthnagar
Kushinagar	Deoria
Chandauli	Varanasi
Sant Ravidas Nagar (Bhadohi)	Varanasi

At present, there are 70 districts in the State which are organized in four zones according to NSSO classification. The regions are as follows.

Table 4. Zone-wise distribution of districts in UP

Zones	Names of districts
Western	Saharanpur, Muzaffarnagar, Bijnor, Moradabad, Rampur, J Phule Nagar, Meerut, Baghpat, Ghaziabad, G. Buddha Nagar, Bulandshahr, Aligarh, Hathras, Mathura, Agra, Firozabad, Etah, Mainpuri, Budaun, Bareilly, Pilibhit, Shahjahanpur, Farrukhabad, Kannauj, Etawah, Auraiya
Central	Kheri, Sitapur, Hardoi, Unnao, Lucknow, Rae Bareli, Kanpur Dehat, Kanpur Nagar, Fatehpur, Barabanki
Eastern	Pratapgarh, Kaushambi, Allahabad, Faizabad, Ambedkar Nagar, Sultanpur, Bahraich, Shrawasti, Balrampur, Gonda, Siddharthnagar, Basti, S. Kabir Nagar, Maharajganj, Gorakhpur, Kushinagar, Deoria, Azamgarh, Mau, Ballia, Jaunpur, Ghazipur, Chandauli, Varanasi, S.R. Nagar (Bhadohi), Mirzapur, Sonbhadra
Southern	Jalaun, Jhansi, Lalitpur, Hamirpur, Mahoba, Banda, Chitrakoot

Sampling design: The sampling design of 61st round was broadly similar to the standard sampling designs used in NSSO. It was a stratified multi-stage design with first stage units (fsu) as the census villages in rural sector and Urban Frame Survey (UFS) blocks in the urban sector. The ultimate stage units, in both sectors, were households. Within each district, two basic strata were formed consisting of rural and urban sectors. However, in the urban sectors, larger cities with population of 10 lakhs or more were considered as separate basic strata. Further, sub-stratification was done in both rural and urban sectors. After determining the overall sample size, further downward allocation was done in proportion to population sizes. Selection of fsu's in rural areas was done by probability proportional to size with replacement (PPSWR) while in urban areas it was done with simple random sampling without replacement (SRSWOR). At the second stage further stratification was done with respect to affluence related criteria.

With the availability of unit level data of NSSO surveys along with the unit weights, further analysis at disaggregated levels has become quite convenient. It is simple to obtain direct estimates not only for districts but also for further disaggregated levels like different groups of the population. Although, sample sizes for these groups become smaller and smaller and estimates so obtained lack adequate precision levels.

Sample size: The sample sizes allocated were as follows:

Table 5. Sample sizes

Sample sizes	Uttar Pradesh		All India	
	Rural	Urban	Rural	Urban
No. of villages/ UFS blocks surveyed	792	336	7999	4602
No. of sample households	7868	3345	79298	45346
No. of sample persons	47067	18387	403207	206529

Source: Level and Pattern of Consumer Expenditure, 2004-05, NSSO 61st round, Report No. 508.

6.2 Predictor Variables Used in the Models

For different poverty related indicators and corresponding variables district level models were fitted.

A number of district-level variables were attempted but we present only those, which were ultimately chosen for the model on the basis of maximum R^2 value.

1) Total number of persons

In this case, for rural as well as urban sectors, as expected, the best predictor variable was the number of persons in the district as per 2001 Population Census with R^2 value as 0.955 and 0.934 in rural and urban sectors respectively.

2) Number of persons below poverty line

Rural: The variables were number of persons corresponding to female literacy, SC/ST population, persons having no specified assets, marginal and small landholding, households living in dilapidated houses and number of agricultural labour accounting for $R^2 = 0.649$.

Urban: Six predictor variables from census 2001 were taken as female literacy, SC/ST population, no assets, household size greater than equal to 5 and dwelling rooms less than equal to one, dilapidated houses and number of industrial labour. Corresponding R^2 was 0.671.

3) Poverty ratio

Rural: Initially seven predictor variables from Population Census 2001 were considered as proportion of female literacy, SC/ST population, no assets, marginal and small landholding, dilapidated houses, houses with no latrine and agricultural labour. Backward regression resulted in the improved set of variables as SC/ST population, no latrine, and agricultural labour as regressors with $R^2 = 0.414$. This model was further improved by including two more variables, female workers and infant mortality rate. Also, outliers were detected and removed and backward regression was applied to form the final model. It comprised of SC/ST population, no latrine, small & marginal landholding and agricultural labour as regressors with $R^2 = 0.573$.

Urban: Final set of variables comprised proportion of SC/ST population, dilapidated houses, infant mortality ratio and no specified assets with $R^2 = 0.357$.

4) Poverty gap ratio

Rural: The model consisted of proportion of SC/ST population, households with no latrine, small & marginal landholding, female literacy and agricultural labour as regressors with $R^2 = 0.487$.

Urban: The model finally consisted of proportion of SC/ST population, dilapidated houses, household size greater than equal to 5 and dwelling rooms less than equal to one and no assets as regressors with $R^2 = 0.286$.

6.3 Some Results on Different Poverty Indicators

Results on different poverty indicators were obtained based on direct estimators as well as model based Small Area Estimators. District wise results for Direct and Small Area Estimate (both rural and urban) along with the percent CVs are presented in Appendix-I. Also, Small Area Estimates for different poverty indicators such as poverty ratio, poverty density, poverty gap ratio, squared poverty gap ratio and Gini coefficient are given in Appendix-II.

A relative distribution of districts according to per cent CV (RSEs) is presented in Table 6. It may be seen that number of persons is being estimated quite precisely even with direct estimates. This number is being estimated with less than 10 per cent CV in nearly 75 per cent districts in rural areas. This is not the case with poverty related indicators. Direct estimates do not provide good estimates for number of persons below poverty line as well as for poverty ratio and poverty gap ratio. It is also seen that SAE estimates show considerable gains over direct estimates.

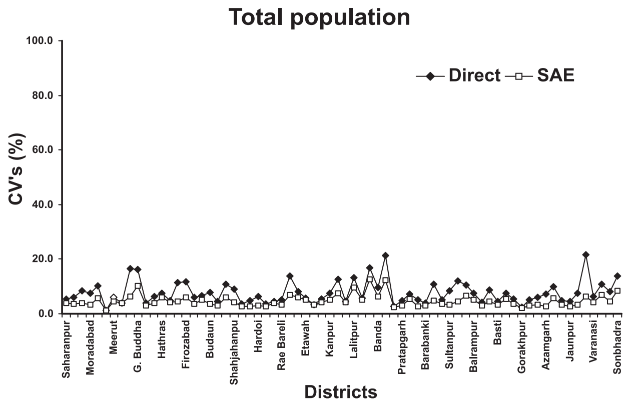
Table 6. Relative distribution of districts according to CV classes for different poverty related characteristics - Rural

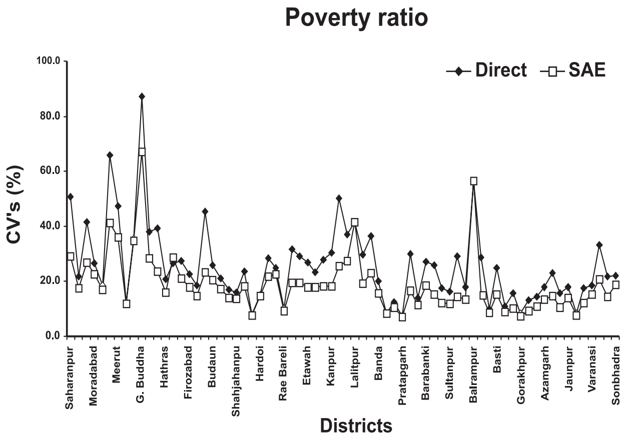
CV classes (%)	Number of persons		Number of persons below poverty line		Poverty ratio		Poverty gap ratio	
	Direct	SAE	Direct	SAE	Direct	SAE	Direct	SAE
0-5	31.4	68.6			8.6	12.9	2.9	2.9
5-10	44.3	27.1	4.3	4.3	31.4	57.1	18.6	37.1
10-20	21.4	4.3	27.1	45.7	35.7	21.4	37.1	42.9
20-30	2.9		35.7	31.4	11.4	2.9	20.0	11.4
30-40			11.4	14.3	5.7	2.9	12.9	1.4
40-50			12.9		7.1	2.9	8.6	4.3
>=50			8.6	4.3				
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

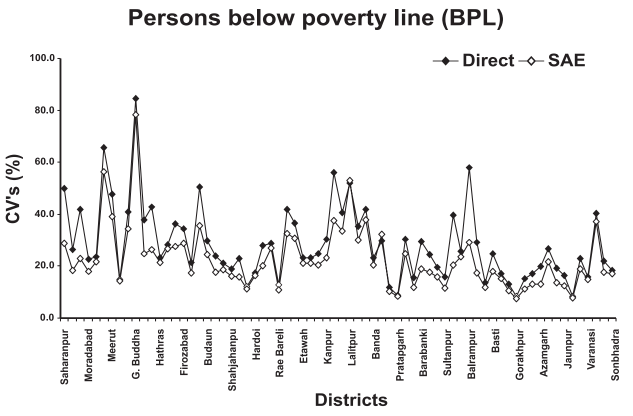
Table 7. Relative distribution of districts according to CV classes for different poverty related characteristics – Urban

CV classes (%)	Number of persons		Number of persons below poverty line		Poverty ratio		Poverty gap ratio	
	Direct	SAE	Direct	SAE	Direct	SAE	Direct	SAE
0-5	2.9	4.3	1.4	2.9	8.6	7.1	1.4	1.4
5-10	15.7	34.3	2.9	2.9	17.1	18.6	15.7	17.1
10-20	25.7	32.9	12.9	30.0	44.3	47.1	42.9	47.1
20-30	51.4	28.6	50.0	57.1	22.9	21.4	24.3	30.0
30-40	4.3		27.1	7.1	5.7	1.4	12.9	2.9
40-50			5.7		1.4	4.3	2.9	1.4
>=50								
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

In the following figures the CVs for Direct Estimates and SAEs for total population, persons below poverty line, poverty ratio and poverty gap ratio are presented for rural areas (Fig.1) and urban areas (Fig. 2). Gains due to application of SAE technique are clearly evident.







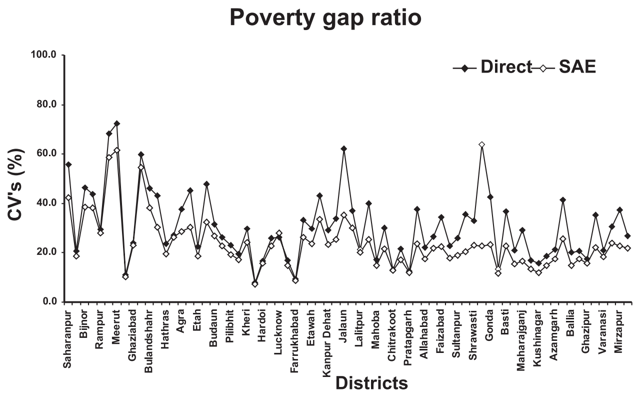
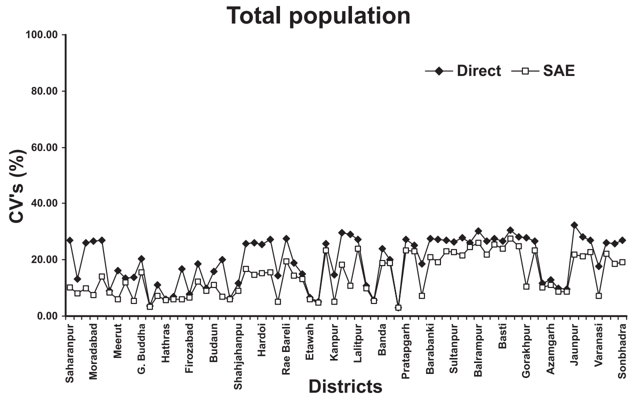
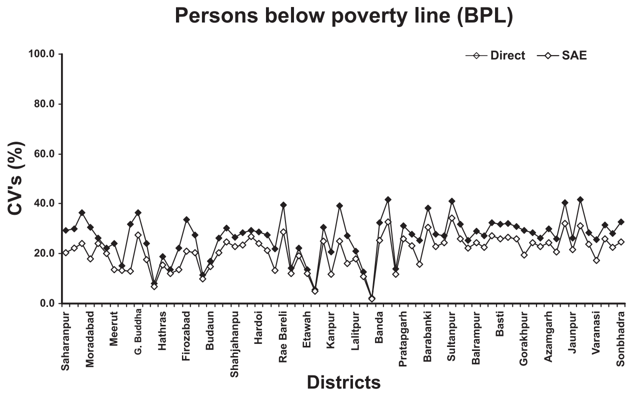


Fig. 1. CV's for Direct and SAE estimates - Rural





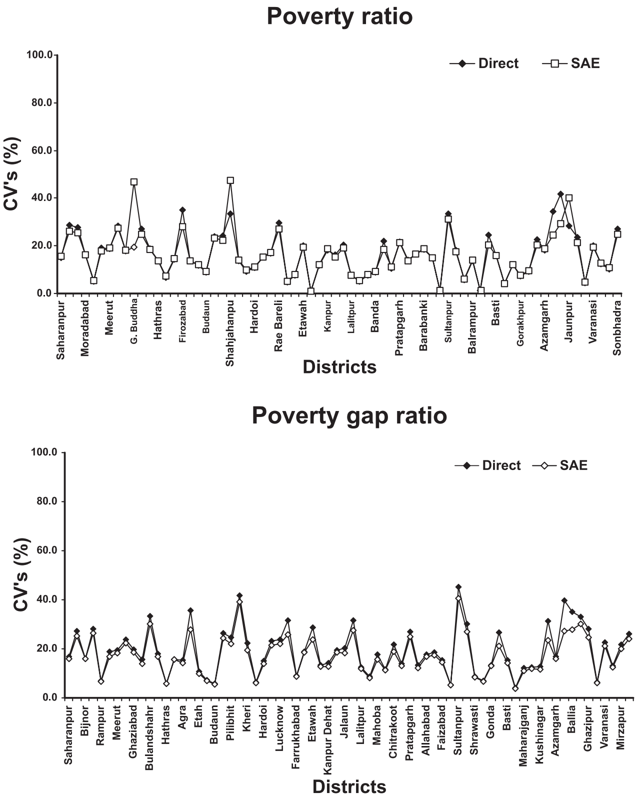


Fig. 2. CV's for Direct and SAE estimates - Urban

Following maps indicate the spatial distribution of poverty ratio, poverty density, poverty gap ratio, squared poverty gap ratio and Gini coefficients over the districts in UP. The map is, however, based on districts as existed in 1991 population census. Heavy concentration of high and medium range of poverty indicators in eastern districts is on expected lines. Also, inequalities as

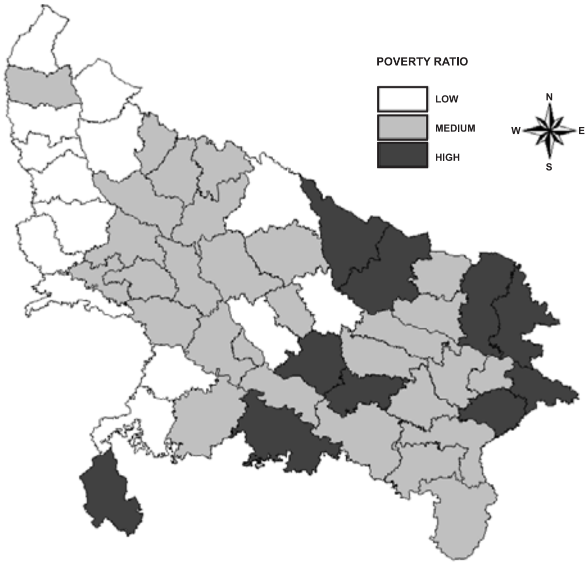


Fig. Spatial distribution of poverty ratio in UP

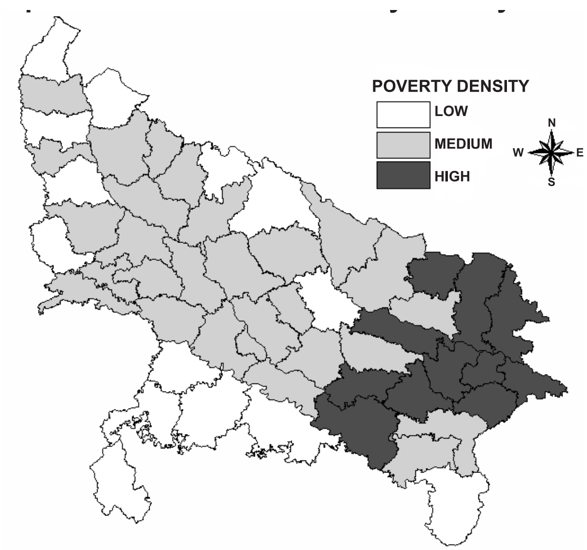


Fig. Spatial distribution of poverty density in UP

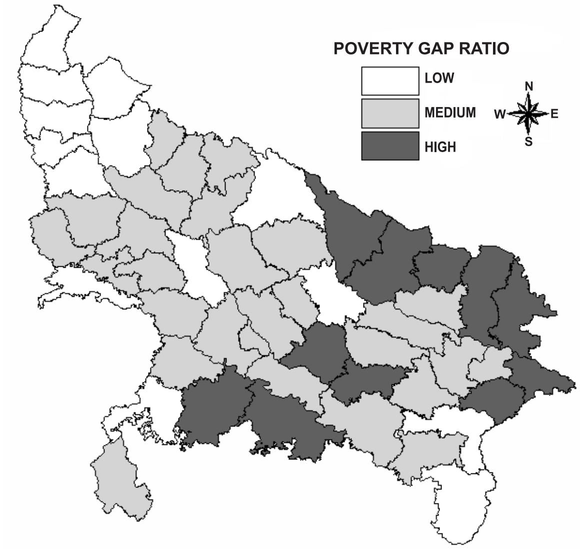


Fig. Spatial distribution of poverty gap ratio in UP

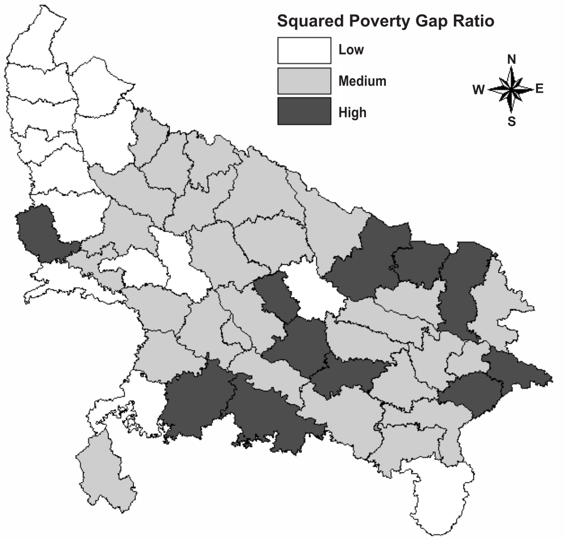


Fig. Spatial distribution of squared poverty gap ratio in UP

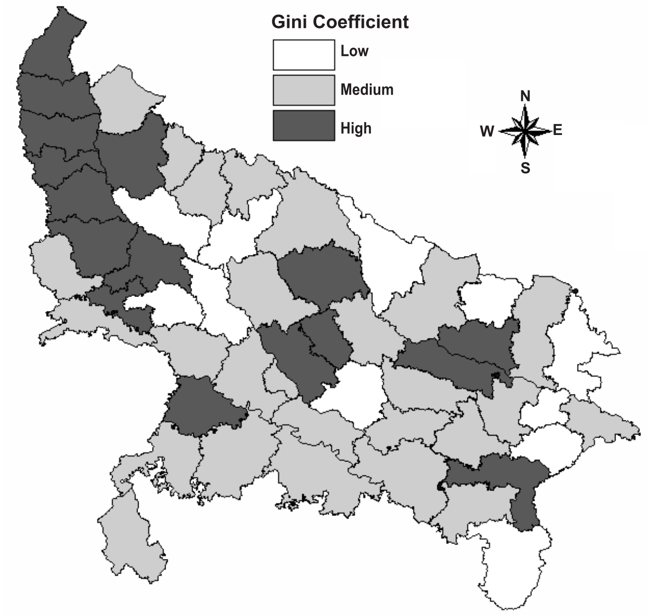


Fig. Spatial distribution of Gini eoefficient in UP

measured by Gini coefficients are more prominent areas having lesser levels of poverty.

As mentioned earlier, poverty ratio measures the incidence of poverty, poverty gap measures the depth of poverty and squared poverty gap provides an idea about severity of poverty. Poverty gap ratios are presented in the previous tables and graphs. As mentioned in Section 3.5, squired poverty gap ratio provides a measure for contribution to poverty (CTP). In the present illustration, share of poor (SP) and CTP has been calculated for each district. Wherever CTP is higher than SP, it indicates that those districts are under severe poverty conditions. Table 8 provides a list of districts with CTP more than SP.

7. CONCLUDING REMARKS

The analysis carried out here indicates that it is feasible to estimate poverty indicators at district level by scaling down the State level poverty estimates utilizing small area estimation techniques. The choice of SAE model and corresponding variables is crucial for successful application of the SAE method. In the process of application of SAE method, it was realized that still there is enough scope for the choice of variables. Efforts for improving the estimates and to apply it for other States, is in process. If unit level data from census may be available, then other methods for poverty mapping may also be attempted.

Table 8. Districts with CTP greater than SP

Districts	Rural	Urban
Bulandshahr		✓
Mathura	✓	✓
Agra		✓
Firozabad		✓
Etah	✓	
Mainpuri		✓
Kheri		✓
Sitapur		✓
Hardoi		✓
Unnao	✓	✓
Lucknow	✓	
Rae Bareli	✓	✓
Kannauj		✓
Etawah	✓	✓
Auraiya	✓	✓
Kanpur Dehat		✓
Jalaun	✓	✓
Lalitpur		✓
Hamirpur	✓	✓
Mahoba	✓	
Banda	✓	✓
Fatehpur		✓
Pratapgarh	✓	✓
Kaushambi	✓	
Faizabad	✓	
Ambedkar Nagar	✓	✓
Bahraich		✓
Shrawasti	✓	
Balrampur	✓	✓
Gonda	✓	
Siddharthnagar	✓	✓
Basti	✓	
S. Kabir Nagar	✓	
Maharajganj	✓	✓
Gorakhpur		
Kushinagar	✓	✓
Deoria		✓
Jaunpur		
Ghazipur	✓	✓

Estimates for poverty indicators are based on consumption expenditure survey data. It may be worthwhile to examine the poverty estimates based on income data, if reliable information on income may be obtained from household surveys. The distributions of expenditure and income are likely to differ and the differences should depend on the income expenditure levels of households. The socio-economic and spatial factors may also contribute towards the variability in the income expenditure patterns. One of the sources for household level income data is the surveys conducted by NCAER. SAE methods have got an important role to play in disaggregated estimates at small area levels.

Estimation of trends and changes are important in poverty studies. Comparability of results sometimes poses serious problems. One of the consumption expenditure surveys (55th round) of NSSO is an example, in which an attempt was made to try different reference periods for different items. The idea was to take care of recall lapse and improve the quality of data, but there are problems in comparability of results with other rounds.

REFERENCES

Angus, Deaton and Valerie, Kozel (eds) (2005). *The Great Indian Poverty Debate*. Macmillan India Limited.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.*, **83**, 28-36.

Benjamin, Davis (2003). Choosing a method for poverty mapping. Food and Agricultural Organization of the United Nations, Rome.

Citro, C.F. and Kalton, Graham (eds) (2002). Small area estimates of school-age children in poverty. National Academy of Sciences.

Datta, K.L. and Sharma, Savita (2002). *Facets of Indian Poverty*. Concept Publishing Company.

Dubey, Amaresh (2007). Poverty and Hunger - Chapter III. In : *District Level Deprivation in the New Millenium* (ed.) Bibek Debroy and Laveesh Bhandari.

Elbers, C., Lanjouw, J.O. and Lanjouw, P.V. (2001). Welfare in Villages and Towns: Micro-level Estimation of Poverty and Inequality. Vrije Universiteit, Yale University and the World Bank (mimeo).

Elbers, C., Lanjouw, J.O. and Lanjouw, P.V. (2003). Micro-level estimation of poverty and inequality. *Econometrica,* **71(1)**, 355-364.

Fay, R.E. and Herriot, R.A. (1979). Estimation of income from small places: An application of James Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761-765.

Hentschel, J., Lanjouw, P. and Poggi, J. (2000). Combining census and survey data to trace spatial dimensions of poverty: A case study of equador. *The World Bank Eco. Rev.*, **14(1)**, 147-165.

Kulshreshtha, A.C. and Kar, Alok (2005). *Estimates of Food Consumption Expenditure from Household Surveys and National Accounts: The Great Indian Poverty Debate.* Macmillan India Limited.

Kozel, V. and Parker, B. (2005). A *Profile and Diagnostic of Poverty in Uttar Pradesh: The Great Indian Poverty Debate*. Macmillan India Limited, 533- 569.

Meenakshi, J.V. and Vishwanathan, Brinda (2005). *Calorie Deprivation in Rural India between 1983 and 1999/2000: Evidence from Unit Record Data: The Great Indian Poverty Debate.* Macmillan India Limited.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley-Interscience, Wiley Series in Survey Methodology.

Sastry, N.S. (2003). District level poverty estimates - Feasibility of using NSS household consumer expenditure survey data. *Eco. Pol. Weekly,* January 25, 2003.

Sen, Amartya (1973). *On Economic Inequality*. Clarendon Press, Oxford.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Meth.*, **31**, 183-195.

APPENDIX I

District wise estimates of poverty ratio - Rural

S. No.	Districts	Number of households selected	Direct estimates		SAE estimates	
			Estimate (%)	C.V. (%)	Estimate (%)	C.V. (%)
1	Saharanpur	120	14.6	50.6	18.6	29.1
2	Muzaffarnagar	160	30.6	21.7	28.9	17.7
3	Bijnor	150	17.8	41.5	20.6	26.8
4	Moradabad	160	17.1	26.7	17.7	22.8
5	Rampur	80	31.7	18.5	28.8	17.0
6	J Phule Nagar	80	4.7	65.7	7.0	41.1
7	Meerut	80	6.5	47.1	7.9	36.2
8	Baghpat	80	28.2	12.5	27.6	11.7
9	Ghaziabad	70	14.9	35.3	12.8	34.9
10	G. Buddha Nagar	40	2.6	87.2	3.2	66.9
11	Bulandshahr	119	14.9	37.9	16.0	28.5
12	Aligarh	118	19.8	39.4	23.0	23.7
13	Hathras	79	31.5	20.8	30.9	16.0
14	Mathura	80	41.0	26.6	23.8	28.9
15	Agra	120	22.1	27.6	23.0	21.1
16	Firozabad	79	26.5	22.7	27.0	17.8
17	Etah	159	30.8	18.5	31.2	14.7
18	Mainpuri	80	22.9	45.5	26.5	23.4
19	Budaun	160	28.8	25.7	25.9	20.6
20	Bareilly	160	30.2	20.9	29.0	17.1
21	Pilibhit	80	27.3	17.0	28.6	14.0
22	Shahjahanpur	120	37.4	15.9	34.0	13.8
23	Kheri	160	21.5	23.6	23.1	18.3
24	Sitapur	199	27.6	8.1	27.5	7.8
25	Hardoi	160	34.2	15.1	29.4	14.8
26	Unnao	160	24.1	28.4	23.9	21.7
27	Lucknow	80	35.6	25.0	27.0	22.7
28	Rae Bareli	160	54.4	10.4	48.8	9.4
29	Farrukhabad	80	28.5	31.6	29.6	19.6
30	Kannauj	80	25.4	29.1	27.6	19.6
31	Etawah	79	32.3	26.8	31.3	18.0
32	Auraiya	80	28.8	23.4	28.0	18.0
33	Kanpur Dehat	80	35.6	27.7	32.6	18.2
34	Kanpur Nagar	80	28.6	30.3	30.9	18.3
35	Jalaun	80	15.3	50.1	22.0	25.5
36	Jhansi	80	19.8	37.2	19.8	27.6

APPENDIX I (Contd..)

District wise estimates of poverty ratio - Rural

S. No.	Districts	Number of households selected	Direct estimates		SAE estimates	
			Estimate (%)	C.V. (%)	Estimate (%)	C.V. (%)
37	Lalitpur	40	42.7	41.4	42.7	-
38	Hamirpur	40	44.1	29.6	36.9	19.3
39	Mahoba	40	23.2	36.4	26.1	23.2
40	Banda	79	52.8	20.0	41.0	15.8
41	Chitrakoot	40	81.5	8.2	81.5	-
42	Fatehpur	120	31.1	12.4	33.1	10.4
43	Pratapgarh	158	65.2	7.4	58.4	7.0
44	Kaushambi	80	45.5	30.0	42.2	16.5
45	Allahabad	200	34.5	13.9	35.8	11.4
46	Barabanki	160	14.2	27.2	18.6	18.5
47	Faizabad	80	25.0	25.9	32.1	15.3
48	Ambedkar Nagar	120	50.4	17.5	47.4	12.2
49	Sultanpur	160	28.5	16.3	33.7	11.8
50	Bahraich	120	43.7	29.2	45.2	14.5
51	Shrawasti	80	56.1	18.0	45.6	13.3
52	Balrampur	80	18.6	56.5	18.6	-
53	Gonda	160	39.0	28.7	41.0	15.6
54	Siddharthnagar	120	66.3	10.0	60.3	8.5
55	Basti	120	23.2	24.8	30.1	15.3
56	S. Kabir Nagar	80	58.0	10.8	55.1	8.8
57	Maharajganj	120	53.4	15.5	56.5	10.2
58	Gorakhpur	160	56.5	8.3	55.2	7.3
59	Kushinagar	160	54.8	13.2	58.6	9.3
60	Deoria	160	41.9	14.3	44.2	10.7
61	Azamgarh	190	29.5	17.7	32.3	13.4
62	Mau	80	39.5	22.9	39.3	14.7
63	Ballia	160	51.5	15.5	53.4	10.6
64	Jaunpur	200	27.9	17.8	29.9	14.1
65	Ghazipur	159	53.7	8.4	49.9	7.8
66	Chandauli	70	36.0	17.6	40.2	12.2
67	Varanasi	120	33.0	18.7	31.8	15.4
68	S.R. Nagar (Bhadohi)	80	30.6	33.3	30.1	20.8
69	Mirzapur	120	28.6	21.8	34.0	14.2
70	Sonbhadra	80	24.8	22.1	25.7	18.8
	Total rural	7868	33.3	2.6	33.3	0.0

APPENDIX I (Contd....)

District wise estimates of poverty ratio - Urban

S. No.	Districts	Number of households selected	Direct estimates		SAE estimates	
			Estimate (%)	C.V. (%)	Estimate (%)	C.V. (%)
1	Saharanpur	40	29.0	15.0	27.1	15.5
2	Muzaffarnagar	40	21.8	28.8	22.3	26.0
3	Bijnor	40	12.7	27.6	13.5	25.3
4	Moradabad	40	25.9	16.3	24.9	16.3
5	Rampur	40	42.2	5.2	38.8	5.6
6	J Phule Nagar	40	39.8	19.1	38.1	18.0
7	Meerut	119	16.0	18.9	15.6	19.1
8	Baghpat	40	13.2	28.5	13.4	27.3
9	Ghaziabad	40	33.9	17.9	31.3	18.0
10	G. Buddha Nagar	40	4.5	19.3	31.4	46.8
11	Bulandshahr	39	24.7	27.1	24.8	24.7
12	Aligarh	39	28.4	18.6	27.0	18.6
13	Hathras	39	28.0	13.7	27.1	13.8
14	Mathura	39	60.9	6.8	55.1	7.3
15	Agra	120	29.6	14.3	28.3	14.5
16	Firozabad	38	34.1	35.2	33.8	27.9
17	Etah	40	41.9	13.7	38.9	13.8
18	Mainpuri	40	28.7	11.9	27.5	12.1
19	Budaun	40	45.8	9.0	42.6	9.4
20	Bareilly	80	24.2	23.9	23.3	23.2
21	Pilibhit	40	46.8	24.2	40.6	22.4
22	Shahjahanpur	40	3.3	33.5	31.0	47.6
23	Kheri	39	34.0	13.8	31.8	14.1
24	Sitapur	38	53.4	9.2	47.5	9.8
25	Hardoi	40	42.1	10.9	38.8	11.3
26	Unnao	40	50.3	15.2	44.4	15.4
27	Lucknow	160	14.7	17.0	14.2	17.3
28	Rae Bareli	39	40.5	29.7	34.7	27.2
29	Farrukhabad	40	43.7	4.9	40.9	5.2
30	Kannauj	40	73.3	7.5	64.6	8.0
31	Etawah	40	17.7	20.2	18.0	19.3
32	Auraiya	40	62.8	0.8	58.6	0.9
33	Kanpur Dehat	40	61.5	11.8	53.9	12.2
34	Kanpur Nagar	160	15.0	18.2	14.3	18.8
35	Jalaun	40	68.1	16.2	59.1	15.3
36	Jhansi	40	24.1	20.3	24.6	19.0

APPENDIX I (Contd....)

District wise estimates of poverty ratio - Urban

S. No.	Districts	Number of households selected	Direct estimates		SAE estimates	
			Estimate (%)	C.V. (%)	Estimate (%)	C.V. (%)
37	Lalitpur	40	34.9	7.5	33.0	7.8
38	Hamirpur	40	54.5	5.2	50.7	5.5
39	Mahoba	40	49.1	7.5	46.2	7.8
40	Banda	40	71.6	9.1	64.3	9.3
41	Chitrakoot	40	54.0	21.9	50.9	18.5
42	Fatehpur	39	49.2	10.7	44.4	11.2
43	Pratapgarh	40	23.3	21.6	22.5	21.3
44	Kaushambi	40	53.2	14.0	49.0	13.8
45	Allahabad	79	35.6	16.5	33.0	16.7
46	Barabanki	40	30.3	18.4	28.1	18.7
47	Faizabad	40	37.9	14.8	35.0	15.1
48	Ambedkar Nagar	40	70.6	1.3	66.0	1.4
49	Sultanpur	40	13.2	33.5	13.5	31.4
50	Bahraich	40	36.8	18.2	35.5	17.4
51	Shrawasti	40	48.7	5.7	45.5	6.1
52	Balrampur	40	28.1	13.8	27.0	14.0
53	Gonda	40	43.8	1.2	41.0	1.3
54	Siddharthnagar	40	36.7	24.4	37.9	20.5
55	Basti	40	36.3	15.8	33.7	16.0
56	S. Kabir Nagar	40	69.3	4.0	63.6	4.3
57	Maharajganj	40	67.5	12.0	58.5	12.2
58	Gorakhpur	40	54.8	7.3	49.9	7.8
59	Kushinagar	40	57.1	9.3	51.9	9.7
60	Deoria	40	59.7	22.6	49.2	20.4
61	Azamgarh	40	12.3	18.2	11.8	18.8
62	Mau	40	36.2	34.4	39.6	24.5
63	Ballia	40	19.6	41.6	24.6	29.2
64	Jaunpur	40	7.7	28.3	36.8	40.2
65	Ghazipur	40	46.5	23.7	42.0	21.2
66	Chandauli	40	74.4	4.6	67.5	4.9
67	Varanasi	119	23.6	20.0	23.2	19.5
68	S.R.Nagar (Bhadohi)	39	45.5	12.6	41.6	12.9
69	Mirzapur	40	53.0	10.2	47.8	10.7
70	Sonbhadra	40	33.3	27.1	31.5	24.8
	Total urban	3345	30.1	2.7	30.1	

APPENDIX II

District wise poverty indicators - Rural

S.No.	Districts	Poverty ratio	Poverty density	Poverty gap ratio	Squared poverty gap ratio	Gini coefficient
1	Saharanpur	18.6	117.6	2.9	0.6	0.30
2	Muzaffarnagar	28.9	207.4	3.9	0.7	0.30
3	Bijnor	20.6	100.6	3.2	0.7	0.25
4	Moradabad	17.7	132.3	2.4	0.5	0.34
5	Rampur	28.8	217.3	5.3	1.4	0.28
6	J Phule Nagar	7.0	36.4	0.4	0.0	0.24
7	Meerut	7.9	50.9	1.1	0.3	0.31
8	Baghpat	27.6	216.0	3.9	0.7	0.29
9	Ghaziabad	12.8	207.6	1.5	0.2	0.30
10	G. Buddha Nagar	3.2	20.9	0.3	0.0	0.23
11	Bulandshahr	16.0	87.7	2.3	0.4	0.36
12	Aligarh	23.0	154.0	3.9	0.7	0.34
13	Hathras	30.9	177.5	3.5	0.5	0.25
14	Mathura	23.8	110.7	6.7	2.7	0.28
15	Agra	23.0	118.8	3.2	0.5	0.25
16	Firozabad	27.0	157.7	5.0	1.2	0.30
17	Etah	31.2	166.5	7.1	2.1	0.30
18	Mainpuri	26.5	160.4	4.2	0.7	0.18
19	Budaun	25.9	149.3	5.1	1.5	0.20
20	Bareilly	29.0	186.6	5.6	1.5	0.26
21	Pilibhit	28.6	101.5	4.5	0.8	0.25
22	Shahjahanpur	34.0	154.9	6.8	1.6	0.19
23	Kheri	23.1	85.7	3.8	0.8	0.24
24	Sitapur	27.5	174.6	5.2	1.1	0.36
25	Hardoi	29.4	166.6	6.0	1.5	0.25
26	Unnao	23.9	129.8	4.9	1.6	0.30
27	Lucknow	27.0	184.6	6.4	3.5	0.38
28	Rae Bareli	48.8	258.8	10.0	3.2	0.19
29	Farrukhabad	29.6	199.9	3.8	0.7	0.19
30	Kannauj	27.6	178.7	3.4	0.8	0.15
31	Etawah	31.3	152.4	5.9	1.8	0.27
32	Auraiya	28.0	140.8	5.5	2.1	0.29
33	Kanpur Dehat	32.6	170.5	6.2	1.8	0.24
34	Kanpur Nagar	30.9	160.2	5.2	1.0	0.28
35	Jalaun	22.0	53.7	5.0	1.6	0.44
36	Jhansi	19.8	46.3	2.4	0.3	0.28
37	Lalitpur	42.7	81.1	6.1	1.5	0.24
38	Hamirpur	36.9	73.3	8.6	3.5	0.27

APPENDIX II (Contd....)

District wise poverty indicators - Rural

S.No.	Districts	Poverty ratio	Poverty density	Poverty gap ratio	Squared poverty gap ratio	Gini coefficient
39	Mahoba	26.1	59.4	6.5	2.1	0.23
40	Banda	41.0	100.6	9.3	3.2	0.24
41	Chitrakoot	81.5	206.1	10.7	2.5	0.13
42	Fatehpur	33.1	159.4	6.5	1.6	0.25
43	Pratapgarh	58.4	363.2	13.1	5.6	0.24
44	Kaushambi	42.2	263.4	9.1	3.7	0.37
45	Allahabad	35.8	262.1	6.9	1.7	0.27
46	Barabanki	18.6	107.8	2.7	0.5	0.26
47	Faizabad	32.1	300.9	6.4	1.6	0.50
48	Ambedkar Nagar	47.4	411.9	10.0	3.7	0.26
49	Sultanpur	33.7	203.1	4.8	0.8	0.23
50	Bahraich	45.2	206.1	9.0	1.9	0.22
51	Shrawasti	45.6	192.8	10.1	7.3	0.26
52	Balrampur	18.6	68.2	8.5	1.5	0.19
53	Gonda	41.0	257.1	9.2	5.0	0.26
54	Siddharthnagar	60.3	390.8	14.3	5.3	0.22
55	Basti	30.1	219.9	6.4	1.5	0.36
56	S. Kabir Nagar	55.1	462.7	11.4	4.0	0.18
57	Maharajganj	56.5	380.6	12.1	3.7	0.21
58	Gorakhpur	55.2	514.2	10.6	2.7	0.23
59	Kushinagar	58.6	597.1	12.2	3.0	0.24
60	Deoria	44.2	491.1	8.6	2.0	0.22
61	Azamgarh	32.3	284.1	5.9	1.4	0.25
62	Mau	39.3	326.8	6.7	1.4	0.22
63	Ballia	53.4	409.7	10.2	2.3	0.24
64	Jaunpur	29.9	271.9	5.0	1.1	0.26
65	Ghazipur	49.9	409.8	10.5	4.9	0.21
66	Chandauli	40.2	234.4	5.5	0.8	0.24
67	Varanasi	31.8	424.7	5.8	1.4	0.23
68	S.R. Nagar (Bhadohi)	30.1	328.3	5.2	1.0	0.19
69	Mirzapur	34.0	143.4	6.2	1.2	0.21
70	Sonbhadra	25.7	32.3	3.7	0.7	0.14
	Total rural	33.3	188.4	6.3	1.8	0.29

Appendix II (Contd...)

District wise poverty indicators - Urban

S.No.	Districts	Poverty ratio	Poverty density	Poverty gap ratio	Squared poverty gap ratio	Gini coefficient
1	Saharanpur	27.1	2019.9	5.9	1.4	0.29
2	Muzaffarnagar	22.3	1639.7	5.1	1.5	0.24
3	Bijnor	13.5	793.9	1.1	0.1	0.23
4	Moradabad	24.9	1360.7	4.0	0.9	0.31
5	Rampur	38.8	2123.7	7.3	1.7	0.21
6	J Phule Nagar	38.1	3010.2	8.2	1.9	0.23
7	Meerut	15.6	1011.2	2.9	0.8	0.28
8	Baghpat	13.4	522.0	2.8	0.5	0.22
9	Ghaziabad	31.3	1864.4	6.2	1.1	0.23
10	G. Buddha Nagar	31.4	350.7	1.0	0.2	0.24
11	Bulandshahr	24.8	1460.9	7.5	3.0	0.37
12	Aligarh	27.0	1367.1	6.5	2.2	0.28
13	Hathras	27.1	2153.6	4.9	1.1	0.22
14	Mathura	55.1	3180.3	16.2	6.8	0.30
15	Agra	28.3	2614.0	8.5	2.8	0.51
16	Firozabad	33.8	2537.7	9.8	3.5	0.36
17	Etah	38.9	2496.0	9.2	2.6	0.36
18	Mainpuri	27.5	635.9	7.3	2.3	0.22
19	Budaun	42.6	822.6	9.4	2.2	0.29
20	Bareilly	23.3	922.8	4.6	1.3	0.39
21	Pilibhit	40.6	2637.0	10.6	3.2	0.21
22	Shahjahanpur	31.0	208.0	0.8	0.2	0.14
23	Kheri	31.8	1529.6	9.1	3.2	0.28
24	Sitapur	47.5	1649.8	17.1	6.6	0.31
25	Hardoi	38.8	1674.1	10.6	3.4	0.24
26	Unnao	44.4	1987.3	14.8	8.2	0.35
27	Lucknow	14.2	1126.7	3.0	0.9	0.44
28	Rae Bareli	34.7	1645.3	10.5	4.0	0.31
29	Farrukhabad	40.9	2865.8	10.1	3.0	0.26
30	Kannauj	64.6	1430.4	18.4	10.9	0.36
31	Etawah	18.0	490.3	6.0	1.8	0.32
32	Auraiya	58.6	3379.2	18.5	8.0	0.31
33	Kanpur Dehat	53.9	1351.1	16.9	6.1	0.34
34	Kanpur Nagar	14.3	1537.0	3.4	0.9	0.40
35	Jalaun	59.1	1125.4	19.2	10.5	0.31
36	Jhansi	24.6	1406.8	5.8	1.8	0.25

Appendix II (Contd...)

District wise poverty indicators - Urban

S.No.	Districts	Poverty ratio	Poverty density	Poverty gap ratio	Squared poverty gap ratio	Gini coefficient
37	Lalitpur	33.0	2878.7	9.6	3.3	0.31
38	Hamirpur	50.7	2328.7	15.5	4.8	0.29
39	Mahoba	46.2	728.0	10.4	2.6	0.27
40	Banda	64.3	2574.7	21.9	12.2	0.29
41	Chitrakoot	50.9	5380.6	7.6	1.6	0.33
42	Fatehpur	44.4	1198.0	13.5	5.6	0.32
43	Pratapgarh	22.5	790.4	8.1	3.5	0.36
44	Kaushambi	49.0	1483.3	11.0	2.5	0.19
45	Allahabad	33.0	2313.5	6.7	1.9	0.32
46	Barabanki	28.1	861.5	5.1	0.8	0.32
47	Faizabad	35.0	1098.0	9.2	2.4	0.42
48	Ambedkar Nagar	66.0	4823.5	21.8	7.7	0.24
49	Sultanpur	13.5	397.6	2.2	0.4	0.22
50	Bahraich	35.5	1462.4	9.5	3.6	0.28
51	Shrawasti	45.5	1560.5	10.6	2.8	0.25
52	Balrampur	27.0	723.7	8.4	3.7	0.35
53	Gonda	41.0	1882.5	8.8	2.9	0.28
54	Siddharthnagar	37.9	1199.9	15.0	7.0	0.33
55	Basti	33.7	1066.3	4.8	0.7	0.38
56	S. Kabir Nagar	63.6	3513.2	17.0	4.6	0.26
57	Maharajganj	58.5	1485.0	18.1	5.3	0.27
58	Gorakhpur	49.9	1118.2	10.0	2.4	0.27
59	Kushinagar	51.9	1287.2	14.9	4.7	0.29
60	Deoria	49.2	1431.2	16.1	7.5	0.28
61	Azamgarh	11.8	633.5	3.3	0.9	0.26
62	Mau	39.6	1508.1	9.7	1.9	0.19
63	Ballia	24.6	747.7	4.1	0.8	0.23
64	Jaunpur	36.8	975.9	1.8	0.4	0.25
65	Ghazipur	42.0	2016.1	14.4	8.6	0.35
66	Chandauli	67.5	3234.4	17.3	5.3	0.28
67	Varanasi	23.2	1997.1	5.1	1.6	0.32
68	S.R.Nagar (Bhadohi)	41.6	1127.0	8.0	1.8	0.29
69	Mirzapur	47.8	2347.1	11.0	3.0	0.21
70	Sonbhadra	31.5	1154.7	4.5	0.8	0.21
	Total urban	30.1	1489.4	7.1	2.3	0.37

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 25-34

Thoughts on Some Applied Statistical Techniques Requiring Attention in Indian Agriculture

K.C. Seal
Former Director General, CSO and Adviser, Planning Commission, Government of India, New Delhi

PROLOGUE

I feel greatly honoured for being invited by the Indian Society of Agricultural Statistics to deliver Dr. V.G. Panse Memorial Lecture at its 62nd Annual Conference being held at S.V. Agricultural College (ANGRAU), near the famous temple of Lord Venkateswara. As I am not an agricultural scientist and currently not very physically fit, I was initially rather hesitant to accept this invitation. I finally accepted it primarily due to the fact that Dr. V. G. Panse had played a vital role in my career in the past to turn my interest from academic research to applied statistics in general in my later career. I met Dr. Panse way back in 1957 for the first time in the UPSC Selection Board just prior to my migration from Calcutta to Delhi for joining Planning Commission, Government of India. After coming to Delhi, I had a number of occasions to listen to his erudite Lectures. I had a great personal regard for his keen intellect and wide field of interest dealing with pragmatic applied research problems especially relating to agricultural research and overall economic development in the country. To commemorate his memories, I thought it appropriate to deliberate here on a few known but relatively recent statistical techniques which in my view could have wider applications in dealing with the diverse types of agricultural research problems in our country. I am not going into their technicalities but elaborating only the basic ideas of these applied techniques with which, I believe, most of you are already familiar.

The following Statistical Techniques are outlined in this lecture:

- Network Analysis and Critical Path Method
- Lorenz Curve and Gini Coefficient
- Fractile Graphical Analysis and Fractile Regression
- Cross Validation and Revalidation
- Data Mining and Spatial Data Analysis
- Fuzzy Data Analysis, Fuzzy Linear Regression and Clustering
- Cost Benefit Analysis
- Meta Analysis and Cochrane Collaboration
- Small Area Statistics
- Use of Ensemble Confidence Limit for Management Action

1. NETWORK ANALYSIS AND CRITICAL PATH METHOD

Network Analysis is the general name given to certain specific techniques which can be used for the planning, management and control of projects. It is a vital technique in project management. It enables us to take a systematic quantitative structured approach to the problem of managing a project before its formulation to its successful completion. It is generally linked with the Critical Path Method (CPM) which is a mathematically based algorithm for scheduling a set of project activities. Critical Path Analysis (CPA) is used to organize and plan projects so that they are completed on time and within budget. The project is structured so that tasks which are dependent on each other are identified at first i.e. to find out the implicit network and thereafter identify critical tasks which need special attention. CPM

[1] *Dr. V.G. Panse Memorial Lecture delivered at 62nd Annual Conference of the Indian Society of Agricultural Statistics at S.V. Agricultural College (ANGRAU), Tirupati on 24 November 2008.*

calculates the longest path of planned activities till the end of the project; it tries to determine the earliest and latest that each activity can start and finish without making the project longer. This process determines which activities are 'critical' (i.e. on the longest path) and which have 'total float' (i.e. can be delayed without making the project longer). A critical path is the sequence of project network activities which add up to the longest overall duration. This determines the shortest time possible to complete the project. Any delay of an activity on the critical path directly impacts the planned project completion date.

2. LORENZ CURVE AND GINI COEFFICIENT

The Lorenz Curve is a graphical representation of the cumulative distribution function of a probability distribution; it is a graph showing the proportion of the distribution assumed by the bottom *y*% of the values. It is often used to represent income distribution, where it shows for the bottom *x*% of households, what percentage *y*% of the total income they have. The percentage of households is plotted on the *x*-axis and the percentage of income on the *y*-axis. It can also be used to show distribution of assets and for representing income distribution. In such use, many economists consider it to be a measure of social inequality.

The Gini Coefficient

The Gini Coefficient is the area between the line of perfect equality and the observed Lorenz curve, as a percentage of the area between the line of perfect equality and the line of perfect inequality. This equals two times the area between the line of perfect equality and the observed Lorenz curve. It is defined as a ratio with values between 0 and 1; the numerator is the area between the Lorenz curve of the distribution and the uniform distribution line; the denominator is the area under the uniform distribution line. A low Gini coefficient indicates more equal income or wealth distribution, while a high Gini coefficient indicates more unequal distribution. 0 corresponds to perfect equality (e.g. everyone has the same income) and 1 corresponds to perfect inequality (e.g. one person has all the income, while everyone else has zero income). The Gini coefficient requires that no one have a negative net income or wealth.

The Gini coefficient is also commonly used for the measurement of the discriminatory power of rating systems in credit risk management.

The Gini index is the Gini coefficient expressed as a percentage, and is equal to the Gini coefficient multiplied by 100. The Gini coefficient is equal to half of the relative mean difference.

3. FRACTILE GRAPHICAL ANALYSIS AND FRACTILE REGRESSION

Fractile Graphical Analysis (FGA) is a useful method to compare economic data related to different populations in India over time as well as to populations differing in respect of geographical regions or in other ways. It is of great importance to policy makers of a country like India to understand the economic condition of the rural community. They would also like to ascertain whether their policies have been able to improve the economic condition especially of the rural population over a period of time. As a measure of the economic well-being of the rural community, we generally consider the proportion of expenditure on food articles to the total expenditure incurred. It is expected that lower this proportion, the greater is the possibility of the rural community being better off.

Let *X* be the total expenditure per capita per 30 days in a household and *Y* be the proportion of total expenditure on food articles per capita per 30 days in the household. Mahalanobis wanted to perform a regression analysis of *Y* on *X* and was interested in comparing the regression functions at two different time points. But due to inflation, the total expenditure (per capita per 30 days) for the two time points become incompatible and may cease to be comparable. Just comparing the regression functions for the two populations did not make much sense. Mahalanobis thought it appropriate to compare the means of the Y-variable in different fractile groups corresponding to the X-variable. This approach leads to a novel way of standardizing the covariate *X* so that comparison of the two different time periods can be done in a more meaningful way. More precisely, FGA does require standardization by considering *F(X)* instead of *X* as the regressor, where F is the distribution function of *X*. While comparing two regression functions, it is sometimes more important to understand the behaviour of the functions over a fractile interval of *X* and not on the entire range of *X*, e.g. in the example cited at the beginning, we would be more concerned with the economic condition of the bottom 5% or 10% of the rural/urban population. Such localized comparison of the regression functions can be done by restricting our attention only to the corresponding fractile intervals under FGA.

Fractile graphs are a more general version of the Lorenz concentration curve and more specific concentration curves where we look at the cumulative relative sums of the levels of the variable of interest (for example expenditure or income) in place of the actual values.

FGA was used by Mahalanobis as an instrument for evaluation of standard of living over different periods of time (for example, total consumption of households between different rounds of National Sample Survey separately for urban and rural populations).

Fractile Regression

We now consider the problem of the effect of the covariates on distributions. Linear regression has been the usual method for investigating the effects of the *x*-variables or covariates on the response variable-*y*. A very simple example of that could be the effect of educational qualification measured in years of education on income or future income. It could be argued that educational qualification is a proxy for ability; hence higher educational qualification would lead to higher earning. However, performing simple linear regression on this somewhat naive model of "Returns to Education" misses some major parts of the story. First, the story of endogeneity, that is to say that it is very rare that education is randomly assigned, so individuals choose education based on their ability and opportunity cost. Hence, it would be wrong to assign the credit of higher income solely to education; there could be quite a few omitted variables. In fact, the error term u in the population linear regression model, i.e.

$$y = b_0 + b_1x + u$$

where y is, say, log of income and x is the number of years of education, b_0 and b_1 are the partial regression coefficients, might be correlated with the independent variable x-problem often times referred to as "endogeneity" in Econometrics.

Apart from the problem of endogeneity, there is another aspect missed by simple linear regression. It is very likely that people with high ability or high educational qualification might command a much higher salary for one extra year of education compared with someone with low ability or education. Linear regression fails to capture this 'differential' treatment of the covariates or in particular 'fractiles of the covariates'. So instead of looking at regression of y on x emphasis is laid on the regression of Y grouped according to fractiles of X; we can then answer the question: for the bottom 10% of educational qualification in the society what is the effect of one more year of education, all else remaining the same.

Fractile Graphical Analysis techniques and in particular, Fractile Regression methods are useful for comparing distributions. For instance, to study male-female or younger-older workers wage gap with respect to returns to education; productivity gap between large and small farm productivity with respect to farm size; difference on returns to equity with farm size, etc.

4. CROSS VALIDATION AND REVALIDATION

Cross validation and revalidation are very important tools in statistical analysis. When a large data set, say S cases, is available, we can divide it into subsets with S_1 and S_2 cases which are also sufficiently large. We can use the subset S_1 to formulate a certain decision rule R based on the discovery of patterns through a search engine. The second set S_2 is thereafter used to evaluate the performance of R using some loss function. In view of the largeness of S_2 we expect to get a reasonably precise estimate of the average loss. This procedure is known as cross validation and is well known in statistical literature. With large data base cross validation is a very useful tool to check how a model will generalize to new data.

Use of interpenetrating sub-samples in large scale sample surveys such as National Sample Surveys in India was emphasized by Indian Statistical Institute, Calcutta as a very useful practical check of the quality of sampled data. This is a simplified version of cross-validation.

There are other possibilities when a large sample is available, especially when the search engine suggests several possible rules R_1, R_2….. based on the subset S_1 of cases. We then divide the S_2 into two subsets S_{21} and S_{22}, and use cross validation of rules R_1, R_2, …. on S_{21} and choose the rule R^* with the minimum loss. We can then compute the loss in applying R^* on the second subset S_{22}. We, thus, have an unbiased estimate of loss in using the rule R^*. This method is described as revalidation.

5. DATA MINING AND SPATIAL DATA ANALYSIS

Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the question that needs to be answered cannot be addressed using existing data analysis techniques; new methods need to be developed.

Data Mining

Data mining is the most important step in the process of knowledge discovery in data base. It is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It is usually described as 'the nontrivial extraction of implicit, previously unknown, and potentially useful information from data' and 'the science of extracting useful information from large data sets or databases'. It has become an indispensable technology for businesses and researchers in many fields. Drawing on work in such areas as statistics, machine learning, pattern recognition, databases, and high performance computing, data mining extracts useful information from the large data sets now-a-days available to industry and science.

Tasks in Classical Data Mining

Data mining tasks are generally divided into two major categories:

Predictive tasks

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Descriptive tasks

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often explanatory in nature and frequently require post processing techniques to validate and explain end results.

Data Mining in Agriculture

One of the most important fields of data mining applications is Agriculture. In agriculture data related to production, consumption, agricultural marketing, fertilizer consumption, seeds, prices (wholesales and retail), technology, agricultural census, marketing region(s), livestock, crops, agricultural credit, plant protection, watershed, area under productions yields, land use statistics, finance and budget etc. can be mined to reach to important information and then to take decisions based on the information. Data mining attempts to bridge the analytical gap by giving knowledge workers the tools to navigate the complex analytical space consisting of rapidly growing data warehouses. There have been very important applications of data mining in agriculture, few of them are enlisted below:

- Mushroom grading
- Apple pest management (PICO)
- Apple proliferation disease
- Soil salinity
- Integrated production in agriculture
- Pesticide abuse
- Precision agriculture
- Drought risk management
- Cow culling studies
- Apple bruising
- Cows in heat

Spatial Data Analysis

Spatial Data Analysis is to detect spatial properties of data. It categorically emphasizes three aspects of spatial data

- Detecting spatial patterns in data
- Formulating hypotheses based on the geography of the data
- Assessing spatial models

In case of spatial data, it is important to be able to link numerical and graphical procedures with the map-need to be able to answer such question as: Where are those cases on the map? With modern graphical interfaces this is often done by ‘brushing’ – for example cases are identified by brushing the relevant part of a boxplot, and the related regions are identified on the map. But with the latest systems like GIS, it is easy to focus on the range of analytical tools required for spatial data analysis.

Spatial Data Mining

Spatial data mining i.e. mining knowledge from large amounts of spatial data, is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing to geographical information system (GIS), computer cartography, environmental assessment and planning etc. The collected data far exceeded human’s ability to analyze. Recent studies on data mining have extended the scope of data mining from relational and transactional databases to spatial databases. It shows that spatial data mining is a promising field, with fruitful research results and many challenging issues.

The application areas of spatial data mining have the dominance in various fields ranging from Geographic Information System (GIS), Resource and Environmental Management, Geology etc. One major and most important field of spatial data mining applications is site-specific Agriculture. Large investments in technology and data collection are being made in the area of precision agriculture/variable rate application practices. Relatively little analysis on the utility of data is currently being performed. A richer set of analytical tools are needed to examine interactions between spatial and temporal characteristics of exogenous effects (weather), inputs (fertilizer, seed variety) and in-situ resources (soil characteristics, physiographic properties). To evaluate and ultimately provide spatial data analysis tools for agricultural practices to reduce in-field variability in an effort to maintain or improve crop yields is a challenging field. The initial work should focus on assessing the feasibility of predicting the spatial dry yield map based on soil sample data (soil chemistry, physical properties), weather information (degree days, precipitation), planting date, cropping and management history and other available information. An expected by-product of the analysis is an initial assessment of the potential benefits of variable rate application. Subsequent work in this area should incorporate web-based data serving and analysis tools under development to prototype an agricultural spatial analysis sub-system. This combination is expected to greatly extend most spatial information systems from data repositories to an active analytical and modeling tool.

6. FUZZY DATA ANALYSIS, FUZZY LINEAR REGRESSION AND CLUSTERING

Mathematical inference needs a model for its useful application. If there is not yet a reasonable model or if the assumed model is not adequate then the results of statistical inference can become useless or even misleading. Hence, the investigation is usually preceded by methods from data analysis to find an appropriate model (pattern recognition). If these data are rather uncertain (e.g. measurements and observations with coarse scales or from greytone pictures) or vague (e.g. verbal statements of experts’, answers in questionnaires, descriptions of contours and colours), then it seems reasonable to use methods from fuzzy theory, see e.g. (Zadeh 1987) including his epoch-making paper from 1965, (Dubois and Prade 1980), (Bandemer and Gottwald 1995), especially for modeling and use of such uncertain or vague data methods of a fuzzy data analysis are established (Bandemer and Nather 1992). The concept of Fuzzy Logic was conceived by Zadeh in 1965 and presented not as a control methodology but as a way of processing data by allowing partial set membership rather than crisp set membership or non-membership. This approach to set theory was not applied to control systems until the 70’s due to insufficient small-computer capability prior to that time. Professor Zadeh argued that in many situations people do not require precise, numerical information input, and yet they are capable of highly adaptive control. If feedback controllers could be programmed to accept noisy, imprecise input, they would be much more cost effective and perhaps easier to implement in practice. As a concrete illustration instead of dealing with temperature control in precise terms such as “T < 1000 F” or “210 C < TEMP < 220C”, we may consider imprecise terms like “IF (process is too cool) AND (the process is getting colder) THEN (add heat to the process)” or IF (process is too hot) AND (process is heating rapidly) THEN (cool the process quickly) – a problem often encountered while using shower in the bath room. Fuzzy Logic is capable of mimicking this type of behaviour but at a very high rate. For Fuzzy Data Analysis each given datum is modeled by an appropriate fuzzy set in specifying its membership function which represents, for every element of a given universe, the degree to which this element belongs to this fuzzy set. Then these fuzzy sets are used for an inference either according to principles of mathematical statistics (Viertl 1995) or according to intrinsic lines of fuzzy set theory by a transfer of the uncertainty and vagueness to another environment, e.g. a parameter space, in which the inference problem can be solved or, at least, a reasonable model can be found. In the above mentioned paper three examples from real world application are sketched, where the statistical approach led to only weak results, whereas by fuzzy data analysis, using additional information from the uncertainty of the data, the problems are solved with total satisfaction.

Fuzzy Linear Regression

In ordinary regression analysis we estimate the best mathematical expression (model) describing the functional relationship between one response variable and a set of independent or explanatory variables. The parameters in the regression are determined based on the well-known principle of least squares. But when there is ‘vagueness’ or ‘impreciseness’ in the measurement of either the response variable or the independent variable(s) or both, the classical regression cannot be applied. It may also be the case that some observations can be described only in linguistic or qualitative terms (such as fair, good and excellent). Although symbolic numbers like 1, 2 and 4 can be assigned to such attributes, and used in regression, this may lead to loss of useful information for regression models. For such data, fuzzy set theory provides a means to model the linguistic or qualitative variables utilizing fuzzy membership functions. Fuzzy regression can be used to fit both fuzzy data and crisp data into a regression model whereas ordinary regression can fit only crisp data. The conventional procedure also cannot deal with interval response variable and non-linear models. Another point to note is that even in the absence of imprecision, if the available data is small, one has to be cautious in the use of probabilistic regression. Fuzzy regression is also a plausible alternative when some of the basic assumptions of classical regression are not fulfilled (as for example, the coefficient of regression must be constant) or the underlying model is vague. So the situations favouring fuzzy regression are vagueness or fuzziness in underlying model and/or the data, presence of linguistic or qualitative variables, interval response variable, and non-linearity, small sample size and violation of distributional or model assumptions.

In conventional regression analysis deviations between observed and estimated values are assumed to be due to random factors whereas in fuzzy regression analysis they are viewed as the fuzziness of the model structure as considered in Tanaka *et al.* (1982). Since then other methods using different optimality criteria were proposed for fitting fuzzy regression. Fuzzy Linear Regression Analysis (FLRA) can be broadly classified into two alternative groups: i) Proposals based on the use of possibilistic concepts (Dubois and Prade 1988), involving the use of Linear Programming (Tanaka *et al*. 1982) and (ii) proposals based on minimum central values, mainly through the use of least squares method as enunciated in the works of Diamond (1988). In the former case minimum fuzziness criterion is employed whereas in the latter the least squares principle is used in combination with either maximum compatibility criterion or minimum fuzziness criterion.

Fuzzy Clustering

One way of doing data analysis and image analysis is with fuzzy clustering methods. A cluster analysis is a method of data reduction that tries to group given data into clusters. Data of the same cluster should be similar or homogenous; data of disjunct clusters should be maximally different. Assigning each data point to exactly one cluster often causes problems, because in real world problems a crisp separation of clusters is rarely possible due to overlapping of classes. Also there are usually exceptions which cannot be suitably assigned to any cluster. For this reason a fuzzy cluster analysis specifies a membership degree between 0 and 1 for each data sample to each cluster.

Most fuzzy cluster analysis methods optimize a subjective function that evaluates a given fuzzy assignment of data to clusters. By suitable selection of parameters of the subjective function it is possible to search for clusters of different forms : on the one side solid clusters in form of (hyper-dimensional) solid spheres, elliptoids or planes, and on the other side shells of geometrical contures like circles, lines, or hyperboles (shell cluster). Latter are especially suitable for image analysis. From the result of a fuzzy cluster analysis a set of fuzzy rules can be obtained to describe the underlying data.

Advantages of Fuzzy Logic

Fuzzy logic has an advantage over many statistical methods in that the performance of a fuzzy expert system is not dependent on the volume of historical data available. Since these expert systems produce a result based on logical linguistic rules, extreme data points in a small data set do not unduly influence these models. Because of these characteristics, fuzzy logic may be a more suitable method for water supply forecasting than current regression modeling techniques.

The use of fuzzy regression enables the specification of decision makers' preferences to the adopted procedure and renders the parameter estimation to be more robust in the presence of extreme values. The methodology is used to estimate groundwater availability.

Potential Area of Application

Fuzzy logic based modeling techniques are applicable for forecasting water supply. Currently, the potential basin runoff is modeled through classical regression, relating the natural runoff to various combinations of data from these sites. Several regression models are developed for each site and, operationally, the forecasts from these various models are compared and a potential range of runoff is selected as the forecast. Water management is planned based on the forecasted range of values and adjusted as the year progresses. Therefore, absolute numerical prediction of the regression models is not as important as correctly forecasting a potential range of runoff volume.

7. COST BENEFIT ANALYSIS

The Cost Benefit Analysis (CBA) is a technique for determining the feasibility and profitability of the outsourcing by quantifying its costs and benefits. For example, a company should ensure that the benefits gained from employing outsourcing services are greater than the costs involved in obtaining the same. Such a decision should include both qualitative and quantitative measures, and must be fully documented. Again, the outsourcing may prove to be more costly or require more time, but ultimately may still be the best solution to meet the growth requirements and economic progress of the company.

Cost Benefit Analysis is typically used by governments to evaluate the desirability of a given intervention in markets. The aim is to gauge the efficiency of the intervention relative to the status quo. The costs and benefits of the impacts of an intervention are evaluated in terms of the public's willingness to pay for them (benefits) or willingness to pay to avoid them (costs). Inputs are typically measured in terms of opportunity costs – the value in their best alternative use. For assessing value of benefits the guiding principle is to list all of the parties affected by an intervention, and place a monetary value of the effect it has on their welfare as it would be valued by them.

The process involves study of monetary value of initial and ongoing expenses vs. expected return. Constructing plausible measures of the costs and benefits of specific actions is often very difficult. In practice, analysts try to estimate costs and benefits either by using survey methods or by drawing inferences from market behaviour.

Cost benefit analysis is mainly, but not exclusively, used to assess the value for money of very large private and public sector projects. This is because such projects tend to include costs and benefits that are less amenable to being expressed in financial or monetary terms (e.g. environmental damage), as well as those that can be expressed in monetary terms. Private sector organizations tend to make much more use of other project appraisal techniques, such as rate of return, where feasible.

The practice of cost-benefit analysis differs between countries and between sectors (e.g. transport, health) within countries. Some of the main differences include the types of impacts that are included as costs and benefits within appraisals, the extent to which impacts are expressed in monetary terms and differences in discount rate between countries.

Accuracy Problems

The accuracy of the outcome of a cost-benefit analysis is dependent on how accurately costs and benefits have been estimated. It will be desirable to indicate the margin of uncertainty of a cost-benefit ratio using plausible estimates from alternative acceptable sources. This is particularly important in social cost-benefit analysis.

8. META ANALYSIS AND COCHRANE COLLABORATION

Meta Analysis

Meta Analysis is a statistical technique for combining the results of several studies that address a set of related research hypotheses. It helps the research workers in reviewing past research work on specified research topics. It is widely used in epidemiology and evidence-based medicine. It has both advantages and disadvantages. An advantage is its objectivity, and yet like any research, ultimately its value depends on making some qualitative-type contextualizations and understandings of the objective data. Another weakness of the method is the heavy reliance on published studies, which may increase the effect as it is very hard to publish studies that show no significant results. This publication bias or 'file-drawer effect' (where non-significant studies end up in desk drawer instead of in the public domain) should be seriously considered while interpreting the outcomes of a meta-analysis. Because of the risk of publication bias many meta-analysis now include a 'failsafe *N*' statistic that calculates the number of studies with null results that would need to be taken into account for drawing meaningful conclusions. Good meta analysis aims for complete coverage of all relevant studies, look for the presence of heterogeneity and explore the robustness of the main findings using sensitivity analysis. Such a technique would prove very useful in agricultural research also, for instance to evaluate the advantages of mixed cultivation of several varieties of a single crop for sustainable agricultural production. Meta analysis leads to a shift of emphasis from single studies to multiple studies. It emphasizes the practical importance of the effect size instead of the statistical significance of individual studies. A weakness of the method is that the source of bias is not controlled by the method. A good meta-analysis of badly designed studies will still result in bad statistics.

Cochrane Collaboration

It is an international organization whose goal is to help people make well informed decisions about health care by preparing, maintaining and ensuring the accessibility of systematic reviews of the effects of health care interventions. A group of over 11500 volunteers spread over more than 90 countries collaborate to carry out the assigned task. It applies a rigorous, systematic process to carefully review the effects of interventions tested in biomedical randomized control trials. The results of these systematic reviews (including updated versions whenever available) are published in the Cochrane Library. Similar systematic reviews for agricultural research studies spread all over the country in specific areas would facilitate proper evaluation of any promising agricultural variety before its general acceptance. This is already being undertaken in many countries including India.

9. SMALL AREA STATISTICS

Small area estimation plays a prominent role in survey sampling due to growing demands for reliable small area statistics from both public and private sectors.

Sample surveys, whether they are conducted by government organizations or by private entities, aim to produce reasonably accurate direct estimators, not only for the characteristics of whole population but also for a variety of subpopulations or domains. These direct estimators are based on domain specific sample data. However, many policy makers and researchers also want to obtain statistics for small domains. A domain is regarded as 'small' if the domain-specific sample is not large enough to support a direct estimator of adequate precision. These small domains are also called small areas, so called because the sample size in the area or domain from the survey is small. Thus, we need special methods to estimate the characteristics of these small areas, referred to as the small area estimation (SAE) techniques.

Each small area typically denotes a subset of the population for which very little information is available from the sample survey. These subsets refer to a small geographic area (e.g. a country, a municipality, a census division, block, tehsil, gram panchayat etc.) or a demographic group (e.g. a specific age-sex-race group of people within a large geographical area) or a cross classification of both. A small area can be any part of the population defined by any method of stratification. The statistics related to these small areas are often termed as small area statistics. The term small area and small domain are interchangeably used in the literature.

In recent years, many countries in the world are transferring the responsibilities for many social and economic policies from national governments to the local governments. Policy planners want to make sure that resources are targeted effectively and efficiently at the areas most in need and for the evaluation of the success of this targeting at a local level, they need reliable small area statistics. The private sector also needs small area statistics for policy making since many businesses and industries rely on local socio-economic conditions. Feasibility studies, for example, require the use of small area statistics. Small area estimates can be made available from various censuses of population, businesses, housing and agriculture. However, the demand for small area estimate also exists for the intercensal period when data usually come from sample surveys.

Due to the increasing demand, survey organizations are faced with producing the small area estimates from existing sample surveys. Unfortunately, sample sizes in small areas tend to be too small, sometimes non-existent, to provide area specific reliable direct estimates for these small areas. That is for small areas, area specific direct estimates are too unstable to be used for planning and policy-making purposes as they are likely to produce unacceptably large standard errors due to the small sample size. Accurate direct estimates for small areas would require a substantial increase in the overall sample size which in turn could overwhelm an already constrained budget and which could further lengthen the data processing time.

The problem of SAE is two fold. First is the fundamental question of how to produce reliable estimates of characteristics of interest, (means, counts, quantiles etc.) for small areas, based on very small samples taken from these areas. The second related question is how to assess the estimation error. Budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. Also, it is often the case that small areas of interest are only specified after the survey has already been designed and carried out. Having only a small sample (and possibly an empty sample) in a given area, the only possible solution to the estimation problem is to borrow information from other related data sets. Potential data sources can be divided into two broad categories:

- Data measured for the characteristics of interest in other 'similar' areas,
- Data measured for the characteristics of interest on previous occasions.

Thus, the SAE methods look at producing estimates with adequate precision for such small areas, through an estimation procedure that 'borrows strength' from related areas or time periods (or both) and thus increase the overall (effective) sample size and precision. These estimation procedures are based on either implicit or explicit models that provide a link to related areas or time periods (or both) through the use of supplementary data (auxiliary information) such as recent census counts and current administrative records, see Rao (2003). Therefore, for estimation at the small areas, it is necessary to employ the estimation methods that 'borrow strength' from related areas. These estimators are often referred to as the indirect estimators since they use values of survey variables (and auxiliary variables) from other small areas or times, and possibly from both. The traditional indirect estimation techniques based on implicit linking models are synthetic and composite estimation. These estimators have advantage of being simple to implement. In addition, these estimation techniques provide a more efficient estimate than the corresponding direct estimator for each small area through the use of implicit models which 'borrow strength' across the small areas. These models assume that all the areas of interest behave similarly with respect to the variable of interest and do not take into account the area specific variability. However, we can find situations where validity of assumed model fails leading to a biased estimator. That is, it can lead to severe bias if the assumption of homogeneity is violated or the structure of the population changed since the previous census. Also, unless the grouping variables are highly correlated with the variable of interest, the synthetic estimators fail to account for local factors. The area specific variability typically remains even after accounting for the auxiliary information. This limitation is handled by an alternative estimation technique based on an explicit linking model, which provides a better approach to SAE by incorporating random area-specific effects that account for the between area variation beyond that is explained by auxiliary variables included in the model, referred as the mixed effect model. These random area effects in the mixed model capture the dissimilarities between the areas. In general, estimation methods based on explicit models are more efficient than methods based on an implicit model. The explicit models used in SAE are a special case of the linear mixed model and are very flexible in formulating and handling complex problems in SAE.

Several methods for SAE based on the nested error regression model, the random regression coefficients model and simple random effects model as special cases of the mixed model have been proposed in the literature. The estimators based on such models, include empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) estimators. Based on the level of auxiliary information available and utilized, two types of random effects model for SAE are described in the literature. The area level mixed effect model which uses area-specific auxiliary information and unit level mixed effect model which uses the unit level auxiliary information. These are special cases of the linear mixed model, usually referred as area level and unit level small area models.

10. USE OF ENSEMBLE CONFIDENCE LIMIT FOR MANAGEMENT ACTION

This is a new topic which I am suggesting as the last item of my lecture in order to get critical comments/ suggestions of this august audience.

The rapid advance in statistical theory quite frequently leads to generation of more than one plausible and valid statistical estimate of the same unknown parameter. Each of these is often claimed to be an improvement over commonly used Best estimate(s). The procedure for getting optimum estimates using maximum likelihood principle, relative cost and efficiency of derived estimates are fairly well-known. In many practical situations, however, where assumptions involved in estimation are considered not quite appropriate, it becomes sometimes difficult for the management and policy makers to arrive at an agreed statistical estimate which can be used by them to avoid future criticism to the extent possible. An innovative practical suggestion based primarily on rational considerations is being made below for criticism/ comments of agricultural statisticians present here.

The primary consideration for making the suggestion is to assist non-statistical policy makers to use a reasonably pragmatic 'point estimate' of the unknown parameter that is likely to be acceptable to majority of the statisticians. We should keep in mind that a 'point estimate' is generally an approximation to the true value of unknown parameter.

Let the number of plausible statistical estimates of the unknown parameter θ be '*m*' (which has to be as few as possible but above one). Let the confidence level for the derived point estimate to be used for management action be $(1 - \alpha)$ [usually α is taken as 5 per cent or 1 per cent]. Thus, both '*m*' and 'α' should be known a'priori. To arrive at an optimum estimate taking into account all the '*m*' plausible estimates, we first consider upper and lower confidence limits with level '$(1 - \beta)$', where β is derived from the equation $(1- \beta)^{m} = 1 - \alpha$ i.e. $\beta = 1- (1 - \alpha)^{1/m}$.

The intersection of the confidence intervals of '*m*' plausible estimates using the derived confidence level $(1-\beta)$ should then be recommended for further management consideration. This common intersection interval of the *m* confidence intervals should normally be fairly short. The upper and lower limits of this intersection interval should be considered thereafter to select an optimum point estimate of the unknown parameter *q* on the lines elaborated in the next paragraph.

Either upper/or lower limits (instead of taking any simple combination, such as a simple average of the two limits) should be preferred so as to ensure that the error to be committed by the management will be 'on the conservative side' so that the use of the limit does not leave out needy beneficiaries by the management action.

There could be some situations in which the common intersection interval of '*m*' intervals is a 'null' set. This would happen in practice only when one or more of the plausible *m* estimates are fairly wide apart. In such situations further careful re-examination of each *m* plausible estimate will be required in consultation with expert statisticians and subject specialists so as to exclude 'outlier estimates' among the *m* estimates for further examination by the management.

ACKNOWLEDGEMENT

Before I conclude this lecture, I gratefully acknowledge the vital support provided to me by several Faculty Members of IASRI (currently headed by Dr. V.K. Bhatia) in the preparation of this lecture. While preparing this lecture, I came across a large number of research papers pertaining to some of the specified topics. I have not purposely included most of these references at the end of this lecture for the sake of convenience.

REFERENCES

Bandemer, H. and Nather, W. (1992). *Fuzzy Data Analysis*. Kluwer Academic Publishers, Dordrecht.

Bandemer, H. and Gottwald, S. (1995). *Fuzzy Sets, Fuzzy Logit, Fuzzy Methods and Applications*. Wiley, Chichester.

Diamond, P. (1988). Fuzzy least squares. *Inform. Sci.*, **46,** 141-157.

Dubois, D. and Prade, H. (1988). *Fuzzy Sets and Systems, Theory and Applications*. Academic Press, New York.

Rao, J.N.K. (2003). *Small Area Estimation.* John Wiley & Sons, Inc. Hoboken, New Jersey, USA.

Tanaka, H., Uejima, S. and Asai, K. (1982). Linear regression analysis with fuzzy model. *IEEE Trans. Systems Man. Cybernet.*, **12,** 903-907.

Viertl, R. (1995). *Statistical Methods for Non-precise Data.* CRC Press, Bota Raton.

Zadeh, L.A. (1987). *Fuzzy Sets and Applications.* Selected Papers, R.R. Yager *et al*. (eds.), Wiley, New York.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 35-42

Inter-District Variation of Socio-economic Development in Andhra Pradesh

Prem Narain, S.D. Sharma, S.C. Rai and V.K. Bhatia
Indian Society of Agricultural Statistics, New Delhi

SUMMARY

The level of development of different districts of Andhra Pradesh was obtained with the help of composite index based on optimum combination of fifty socio-economic indicators. The district-wise data for the year 2001-02 in respect of these fifty indicators were utilized for 22 districts of the State. The level of development was estimated separately for agricultural sector, infrastructural facilities and overall socio-economic sector. The district of West Godavari was ranked first in overall socio-economic development and the district of Guntur was found on the first position in respect of agricultural development. Wide disparities were observed in the level of development among different districts. Infrastructural facilities were found to be positively associated with the level of developments in agricultural sector and overall socio-economic field. Agricultural development was influencing the overall socio-economic development in the positive direction.

Key words: Developmental indicators, Composite index, Potential targets, Model districts.

1. INTRODUCTION

Developmental programmes have been taken up in the country in a planned way through various Five Years Plans for enhancing the quality of life of people by providing basic necessities as well as effecting improvement in their social and economic well being. The green revolution in agricultural sector has enhanced the crop productivities and commendable progress in the industrial front has increased the quantum of manufactured goods but there is no indication that these achievements have been able to reduce substantially the level of regional disparities in terms of socio-economic development. For focussing the attention of scientists, planners, policy makers and administrators on the problems of estimation of level of development, a seminar was organized jointly by Planning Commission, Government of India and State Planning Institute, Government of Uttar Pradesh during April 1982. Realizing the seriousness and importance of estimation of level of development, the Indian Society of Agricultural Statistics conducted a series of research studies in this direction.

The present study is conducted in the State of Andhra Pradesh where the district level data on socio-economic variables for the year 2001-02 are analyzed for estimating the level of development.

2. DEVELOPMENTAL INDICATORS

Development is a multidimensional process and its impact cannot be fully captured by a single indicator. A number of indicators when analyzed individually do not provide an integrated and easily comprehensible picture of reality. Hence, there is a need for building up of a composite index of development based on optimum combination of all the indicators. Each district faces situational factors of development unique to it as well as common administrative and financial factors. Developmental indicators common to all the districts have been included in the analysis. Composite indices of development have been obtained for different districts by using the data on the following developmental indicators.

01. Percentage forest area
02. Percentage net area sown
03. Percentage of net area sown more than once
04. Percentage area irrigated
05. Fertilizer consumption (kg/ha)

06. Cropping intensity
07. Yield rate of rice (kg/ha)
08. Yield rate of groundnut
09. Yield rate of sugarcane
10. Yield rate of cotton
11. Yield rate of chillies
12. Yield rate of total foodgrains
13. Per capita area of operational holdings
14. Number of cattle (per lakh population)
15. Number of buffaloe (per lakh population)
16. Number of sheep (per lakh population)
17. Number of goat (per lakh population)
18. Number of poultry (per lakh population)
19. Production of milk (per lakh population)
20. Production of eggs (per lakh population)
21. Production of meat (per lakh population)
22. Hand operated implements (‘000 no.) (per lakh population)
23. Animal operated implements (‘000 no.) (per lakh population)
24. Percentage of cultivators
25. Percentage of agricultural labourers
26. Work participation rate
27. Percentage of workers engaged in the non-agricultural activities
28. Percentage of SC population
29. Percentage of ST population
30. Decennial growth rate of population (1991-2001)
31. Sex ratio
32. Population density (No. of persons per square km. of area)
33. Rural literacy rate
34. Total literacy rate (rural + urban)
35. Number of primary schools (per lakh population)
36. Teacher-pupil ratio
37. Drop out rates (Class I-V)
38. Percentage of urban population
39. Annual birth rate
40. Annual death rate
41. Number of PHC and medical dispensaries (per lakh population)
42. Number of doctors (per lakh population)
43. Number of factories (per lakh population)
44. Number of post offices (per lakh population)
45. Road length (per 1000 sq.km. of area)
46. Average population per bank (in ‘000)
47. Credit/Deposit ratio
48. Number of beneficiaries under WSHP (per lakh population)
49. GDP at current prices
50. GDP at constant prices

A total of 50 developmental indicators have been included in the analysis. These indicators may not form an all inclusive list but these are the major interacting components of development.

3. METHOD OF ANALYSIS

There are several statistical methods which are used for estimating the level of development but most of these methods are having their own limitations. The major limitation arises from the assumptions made about the developmental indicators themselves and their weightage in aggregate index. Keeping in view the limitations of different methods in estimating the level of development, the following statistical procedures are used in this study. Variables for different developmental indicators are taken from different population distributions and these are recorded in different units of measurement. The values of the variables are not quite suitable for combined analysis. Hence, the variables are transformed for the combined analysis as given below.

Let $[X_{ij}]$ be data matrix giving the values of the variables of i^{th} district, $i = 1, 2, \dots n$ (number of districts) and j^{th} indicator, $j = 1, 2, \dots k$ (number of indicators).

For combined analysis $[X_{ij}]$ is transformed to $[Z_{ij}]$ as follows:

$$\left[Z_{ij}\right] = \frac{X_{ij} - \overline{X}_j}{s_j}$$

where $\overline{X}_j$ = mean of the j^{th} indicator

s_j = standard deviation of j^{th} indicator

$[Z_{ij}]$ is the matrix of standardized indicators.

From $[Z_{ij}]$, identify the best value of each indicator. Let it be denoted as Z_{oj}. The best value will be either the maximum value or the minimum value of the indicator depending upon the direction of the impact of indicator on the level of development. For obtaining the pattern of development C_i of i^{th} district, first calculate P_{ij} as follows:

$$P_{ij} = (Z_{ij} - Z_{oj})^2$$

Pattern of Development is given by

$$C_i = \left[\sum_{j=1}^{k} P_{ij} / (\text{CV})_j\right]^{1/2}$$

$(\text{CV})_j$ = coefficient of variation in X_{ij} for j^{th} indicator

Composite index of development is given by

$$D_i = C_i / C$$

where

$$C = \overline{C} + 3\text{S}_{Di}$$

$\overline{C}$ = Mean of C_i

S_{Di} = Standard Deviation of C_i

Smaller value of D_i will indicate high level of development and higher value of D_i will indicate low level of development.

For identifying the model districts for low developed districts, the distance between different pairs of districts based on all the indicators is calculated.

The distance between two districts i and p is given by d_{ip} where

$$d_{ip} = \left[\sum_{j=1}^{k} (Z_{ij} - Z_{pj})^2\right]^{1/2}$$

$$i = 1, 2, \ldots n \text{ and } p = 1, 2, ..., n$$

Here $d_{ii} = 0$ and $d_{ip} = d_{pi}$

Now d_{ip} can be written as

$$d_{ip} = \begin{bmatrix} 0 & d_{12} & \square & d_{1n} \\ d_{21} & 0 & \square & d_{2n} \\ \square & \square & \square & \square \\ d_{n1} & d_{n2} & \square & 0 \end{bmatrix}$$

From the above distance matrix, find out the minimum distance for each row. Let the minimum distance for row i is given by d_i.

Obtain the Critical Distance (CD) as follows :

$$\text{CD} = \overline{d} + 2\text{S}_d$$

where $\overline{d}$ = Mean of d_i

and S_d = Standard Deviation of d_i

Model districts will be identified as follows:

Model districts for district A will be those districts whose composite index of development is less than that of district A and the developmental distance of these districts from district A is less than or equal to Critical Distance (CD). Thus, model districts will be better developed in comparison to district A.

The best value of each developmental indicator of the model districts will be taken up as the potential target of that indicator for district A.

The advantages and disadvantages of composite index of development are as follows:

Advantages

- It can summarize complex or multi-dimensional issues.
- It is easier to interpret.
- It facilitates the task of ranking states/districts/ regions etc. on complex issues.
- It can assess the progress of different regions over time.
- It reduces the size of a set of indicators or includes more information within the existing size limit.
- It places performance and progress of different regions at the centre of policy arena.
- It facilitates communication with general public (citizen, media etc.) and promotes accountability.

Disadvantages

- It may send misleading policy messages if it is poorly constructed.
- It may invite simplistic policy conclusions which may not be possible for adoption.
- It may be misused.
- The selection of indicators and weights for aggregating the composite index can change the final conclusions.
- It may lead to inappropriate conclusions if indicators that are difficult to measure, are ignored.

4. RESULTS AND DISCUSSIONS

4.1 The Level of Development

The composite indices of development have been worked out for different districts for agricultural sector, infrastructural facilities and overall socio-economic sector. The districts have been ranked on the basis of developmental indices. The composite indices of development along with the rank of the districts are given in Table 1.

In case of agricultural sector, Guntur was found to be the best developed district in the State whereas the district of Ranga Reddy was on the last place. The composite indices of development varied from 0.60 to 0.87. In case of infrastructural facilities, the district of West Godavari was on the first position and the district of Ranga Reddy was on the last position. The composite indices varied from 0.60 to 1.00. As regards overall socio-economic development, the district of West Godavari was on the first place and the district of Ranga Reddy was on the last place. The composite indices varied from 0.61 to 0.99. Four most developed districts are found to be West Godavari, Karimnagar, East

Table 1. Composite Indices of Development (CI) and Rank of District

S.No.	District	Agricultural Sector		Infrastructural Facilities		Socio-economic Sector	
		C.I.	Rank	C.I.	Rank	C.I.	Rank
1	Srikakulam	0.73	11	0.65	6	0.68	7
2	Vizianagaram	0.76	14	0.67	8	0.71	9
3	Visakhapatanam	0.87	21	0.72	12	0.78	18
4	East Godavari	0.65	5	0.64	5	0.66	3
5	West Godavari	0.61	2	0.60	1	0.61	1
6	Krishna	0.64	4	0.75	16	0.73	11
7	Guntur	0.60	1	0.68	10	0.68	6
8	Prakasam	0.71	10	0.75	17	0.75	14
9	Nellore	0.68	6	0.65	7	0.67	5
10	Chittoor	0.77	15	0.63	3	0.69	8
11	Cuddapah	0.80	16	0.73	13	0.77	16
12	Anantpur	0.83	17	0.82	20	0.84	20
13	Kurnool	0.83	18	0.81	19	0.83	19
14	Mahboobnagar	0.84	19	0.86	21	0.87	21
15	Ranga Reddy	0.87	22	1.00	22	0.99	22
16	Medak	0.76	13	0.68	9	0.72	10
17	Nizamabad	0.69	9	0.64	4	0.67	4
18	Adilabad	0.86	20	0.70	11	0.76	15
19	Karimnagar	0.64	3	0.61	2	0.63	2
20	Warangal	0.73	12	0.77	18	0.77	17
21	Khammam	0.69	7	0.74	15	0.74	13
22	Nalgonda	0.69	8	0.74	14	0.74	12

Godavari and Nizamabad and four least developed districts are Ranga Reddy, Mahboobnagar, Anantpur and Kurnool. During 1991-92, four most developed districts were found to be East Godavari, West Godavari, Guntur and Krishna. The districts of Guntur and Krishna have gone down in the relative ranking within a period of 10 years from 1991-92 to 2001-02 mostly due to shortfall in the infrastructural facilities. The districts of Ranga Reddy, Anantpur, Mahboobnagar and Nalgonda were found to be low developed during 1991-92. Most of these districts are still found to be among the low developed districts of the State.

4.2 Different Stages of Development

For classificatory purposes, a simple ranking of the districts on the basis of composite index of development is sufficient. However, a more meaningful characterization of different stages of development would be in terms of suitable fractile classification from the assumed distribution of the mean of the composite indices. For relative comparison, it appears quite valid to assume that the districts having the composite indices less than or equal to (Mean – SD) are in high developed category, the districts having the composite indices in between (Mean – SD) to (Mean) are in high middle level category, the districts having composite indices in between (Mean) to (Mean + SD) are in low middle level developed category and the districts having the composite indices greater than or equal to (Mean +SD) are in low level developed category.

On the basis of above classifications, the districts are put in four stages of development as high, high middle, low middle and low. Table 2 presents the number of districts along with the percentages of area and population lying in different stages of development.

It is observed from the table that in agricultural sector, five districts are in high developed category. These districts cover about 18 per cent area and 29 per cent population of the State. Seven districts having about 32 per cent area and 27 per cent population of the State are lying in high middle developed category. Six districts are found in low middle developed category. These districts are having about 30 per cent area and 26 per cent population of the State. Four districts covering about 19 per cent area and 19 per cent population of the State are lying in low developed category. Immediate actions are required to be taken in these districts for enhancing agricultural development.

Infrastructural facilities are quite important and these are extremely essential for enhancement of level of development of different sectors of the economy. Five districts of the State are found to have high category of these facilities. These districts cover about 20 per cent area and 25 per cent population of the State. Eight districts covering about 33 per cent area and 33 per cent population of the State are lying in high middle developed category. Seven districts are found in low middle developed category. These districts are having about 39 per cent area and 33 per cent population. Two districts having about 9 per cent area and 10 per cent population are found in low developed category. Immediate improvements in infrastructural facilities are needed in these districts.

Table 2. Number of Districts, Percentages of Area and Population lying under Different Stages of Development

Stage of Development	Number of Districts	Area (%)	Population (%)
Agricultural Development			
High	5	18.3	28.7
High Middle	7	31.9	27.3
Low Middle	6	30.4	25.5
Low	4	19.4	18.5
Infrastructural Facilities			
High	5	19.5	25.2
High Middle	8	32.5	32.5
Low Middle	7	38.6	32.5
Low	2	9.4	9.8
Socio-economic Development			
High	3	11.0	16.8
High Middle	10	39.5	42.2
Low Middle	5	26.6	21.0
Low	4	22.9	20.0

With regard to socio-economic development, three districts having about 11 per cent area and 17 per cent population of the State are found to be in high developed category. Ten districts are found to be in high middle developed category. The districts cover about 40 per cent area and 42 per cent population of the State. Five districts having about 27 per cent area and 21 per cent population of the State are found in low middle developed category. Four districts are observed to be in low developed category. These districts are having about 23 per cent area and 20 per cent population of the State. Population density in the high developed area is generally higher than that in low developed area.

The districts of East Godavari, West Godavari and Karimnagar are found to be in high developed category in agricultural sector, infrastructural facilities and socio-economic sector whereas the districts of Mahboobnagar and Ranga Reddy are in low developed category in all these sectors.

4.3 Inter-relationship among Different Sectors of Economy

For proper development and better level of living, it is essential that all the sectors of economy should flourish together. System of education envisages all round development of manpower and human resources required for socio-economic activities. The correlation coefficients between development of different sectors of economy are given in Table 3.

It is observed from Table 3 that the correlation coefficients between the development of infrastructural facilities; and agricultural and socio-economic developments are positive and highly significant which indicates that the infrastructural facilities influence both agricultural development and socio-economic development in the positive direction. In the same way, the correlation coefficient between agricultural development and overall socio-economic development is found to be positive and highly significant. Therefore, the level of development in agricultural sector influences the level of development in overall socio-economic sector in the positive direction. The development in overall socio-economic sector depends on both agricultural development and availability of infrastructural facilities.

Table 3. Correlation Coefficients

Factors	Agricultural Development (D_1)	Infrastructural Facilities (D_2)	Socio-economic Development (D_3)
Agricultural Development (D_1)	1	0.607**	0.775**
Infrastructural Facilities (D_2)		1	0.973**
Socio-economic Development (D_3)			1

** Correlation coefficient is significant at 0.01 probability level.

4.4 Potential Targets of Developmental Indicators for Low Developed Districts

It is quite useful and important to examine the extent of improvements needed in different developmental indicators for enhancing the level of development of low developed districts. This will help the planners and administrators to readjust the resources for bringing about uniform regional development. For estimation of potential targets of developmental indicators, it is essential to identify the model districts for low developed districts. In case of overall socio-economic development, four districts namely Anantpur, Kurnool, Mahboobnagar and Ranga Reddy are found to be low developed. Model districts for each of these districts are identified on the basis of composite index of development and distance between these districts with their model districts and are presented in Table 4.

Model districts are better developed in comparison to low developed districts. The districts of Chittoor, Nizamabad, Nellore and Nalgonda are found to be the model districts for all the four low developed districts of the State. The best values of the developmental indicators of model districts are taken as potential targets of low developed districts. The present values of developmental indicators along with the potential targets for the low developed districts are presented in Table 5.

Potential targets are quite high in comparison with the present achievements for most of the indicators. Suitable actions are required for achieving the potential

Table 4. Model Districts for Low Developed Districts

S.No.	Low Developed Districts	Model Districts
1	Anantpur	Chittoor, Vizianagaram, Nizamabad, Nellore, Nalgonda
2	Kurnool	Chittoor, Nellore, Nalgonda, Nizamabad
3	Mahboobnagar	Chittoor, Nellore, Nizamabad, Karimnagar, Nalgonda
4	Ranga Reddy	Chittoor, Vizianagaram, Nizamabad, Nalgonda, Nellore

targets. The broad suggestions for improving the level of development of low developed districts are given below:

Anantpur District

This district is low developed in overall socio-economic development. Irrigation facilities are required to be created in the district and the cultivators should be encouraged to enhance the application of fertilizers. Productivity levels of various crops are quite low. Action is needed to enhance yield rates of different crops by use of irrigation and fertilizers. In some parts of the district due to non-availability of sufficient irrigation facilities, improved dry land farming system should be advocated among the cultivators. Farmers should be motivated for rearing cattle and buffaloe. Road transport, communication system, educational and medical facilities etc. are required for improvements in the district. Immediate actions should be taken for improving these infrastructural facilities.

Kurnool District

This district is low developed in overall socio-economic development. Irrigation facilities should be enhanced in the district. High yielding dryland farming practices should be advocated among the cultivators. The level of crop productivities is small and it requires improvement by use of irrigation and fertilizers. Farmers should be encouraged to adopt improved animal husbandry practices. Literacy rate is quite low in the district. Suitable actions are required for enhancing the literacy rate and also for improvement in road transport

Table 5. Present Value of Developmental Indicators of Low Developed Districts along with Potential Target

S.No.	Developmental Indicators	Low Developed Districts				Potential Target
		Anantpur	Kurnool	Mahboob-nagar	Ranga Reddy	
01	Net area sown (%)	54.7	46.5	44.6	37.1	54.7
02	Area irrigated (%)	13.2	18.2	18.1	23.8	65.8
03	Fertilizer consumption (kg/ha)	30.9	74.1	38.7	161.4	171.4
04	Cropping intensity	0.96	1.11	1.10	1.08	1.48
05	Yield rate of rice	2881	2694	2389	2656	3354
06	Yield rate of groundnut	467	1046	897	1100	2298
07	Yield rate of sugarcane	86673	86979	00	83448	88322
08	Yield rate of cotton	170	178	207	263	279
09	Yield rate of foodgrains	1701	1468	1115	1271	2787
10	No. of cattle (per lakh population)	18.6	15.7	25.5	8.8	30.2
11	No. of buffaloe (per lakh population)	8.8	11.9	10.1	5.6	23.3
12	Production of milk (per lakh population)	5.6	6.0	5.3	2.4	14.4
13	Animal operated implements (000 no.) (per lakh population)	16.5	20.4	22.5	6.4	22.5
14	Work participation rate	48.9	49.5	51.8	39.9	52.2
15	Workers in non-agricultural activities (%)	32.3	31.6	26.6	59.8	59.8
16	Total literacy rate	57	54	46	68	74
17	No. of primary school (per lakh population)	88	59	71	49	115
18	No. of doctors (per lakh population)	12.3	16.9	7.5	6.3	17.5
19	No. of factories (per lakh population)	15	23	8	42	42
20	No. of PO (per lakh population)	26	29	24	12	32
21	Road length (per ‘000 sq.km of area)	34	25	45	46	58
22	No. of beneficiaries under WSHP (per lakh population)	56	46	54	35	76
23	GDP at current prices	165	162	129	205	251
24	GDP at constant prices	107	100	80	132	137

and communication systems. Educational and medical facilities should be enhanced in the district.

Mahboobnagar District

The district is found to be in low developed category in agricultural sector, infrastructural facilities and overall socio-economic field. Improvements are required to be made in the field of irrigation facilities and applications of fertilizers. Action should be taken to popularize the improved animal husbandry practices. Literacy rate is quite low in the district. Suitable actions should be taken for enhancing the literacy rate and also to improve the road transport, communication systems, medical and educational facilities in the district.

Ranga Reddy District

This district is found to be in low developed category in agricultural sector, infrastructural facilities and overall socio-economic sector. High percentages of labour force are engaged in non-agricultural activities. In agricultural sector, as far as possible irrigation facilities should be created. In non-irrigated areas, improved dry land farming system should be adopted. Improved animal husbandry practices should be adopted in the district. The present literacy rate is satisfactory but it needs continuous improvement. Actions should be taken to enhance the facilities for road transport and communication systems. Educational and medical facilities also need improvement in the district.

5. CONCLUSIONS

The broad conclusions emerging from the study are as follows:

(i) With respect to socio-economic development, the district of East Godavari, West Godavari and Karimnagar are found to be better developed in comparison to other districts of the State. The districts of Anantpur, Kurnool, Mahboobnagar and Ranga Reddy are found to be low developed. Coastal districts are generally found to be better developed.

(ii) In agricultural sector, five districts namely East Godavari, West Godavari, Krishna, Guntur and Karimnagar are better developed as compared to other districts. Visakhapatnam, Mahboobnagar, Ranga Reddy and Adilabad districts are low developed.

(iii) Infrastructural facilities in respect of road transport, communication system, availability of educational and medical facilities are found to be better in the districts of East Godavari, West Godavari, Chittoor, Nizamabad and Karimnagar. These facilities are poor in the districts of Mahboobnagar and Ranga Reddy.

(iv) Infrastructural facilities are found to be very highly associated with both agricultural development and socio-economic development. Agricultural development is found to be positively influencing the overall socio-economic development in the State.

(v) Wide disparities in the level of development have been observed in different districts.

(vi) For enhancing the level of development of low developed districts, model districts have been identified and potential targets of important developmental indicators have been estimated.

(vii) It would be better to examine and evaluate the level of development at smaller level (say tehsil, taluka or block level) for making location specific recommendations for improvement of level of development.

REFERENCES

Narain, P., Rai, S.C. and Shanti Sarup (1991). Statistical evaluation of development on socio-economic front. *J. Ind. Soc. Agril. Statist*., **43,** 329-345.

Narain, P., Rai, S.C. and Shanti Sarup (1994). Regional dimensions of socio-economic development in Andhra Pradesh. *J. Ind. Soc. Agril. Statist.*, **46**, 156-165.

Narain, P., Rai, S.C. and Bhatia, V.K. (1999). Inter–district variation of development in southern region. *J. Ind. Soc. Agril. Statist.,* **52,** 106-120.

Narain, P, Rai, S.C., Sharma, S.D. and Bhatia, V.K. (2007). Statistical evaluation of social development at district level. *J. Ind. Soc. Agril. Statist.,* **61,** 216-226.

Regional dimensions of India's economic development. *Proceedings of Seminar held on April 22-24, 1982*, sponsored by Planning Commission, Govt. of India and State Planning Institute, Govt. of U.P., Lucknow.

Districts at a Glance in Andhra Pradesh (2003). Directorate of Economics & Statistics, Andhra Pradesh, Hyderabad.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 43-45

Symposium on
Accelerated Growth of Agriculture through Information Technology

Chairman : Dr. S.D. Sharma
Convenor : Dr. P.K. Malhotra

The following four papers covering various aspects related with the theme of the symposium were presented by the following speakers:

1. Emerging Computing and Communication Techniques for Accelerated Growth is Dairying – An NDRI Perspective — Dr. D.K. Jain
2. Accelerated Growth of Agriculture through Knowledge Management in Plant Genetic Resources — Dr. R.C. Agrawal
3. ICT for Accelerated Growth Resources – Status and Prospects — Dr. Anil Rai
4. A Decentralized Process for Web based Data Management of Agricultural Education — Dr. R.C. Goyal

After detailed discussions, the following recommendations emerged out:

- More exclusive investment in needed for primary data collection to get quality data.
- High level co-ordination committee meetings need to be made more interactive and effective to achieve better co-ordination of activities between centre and states.
- The awareness of simple basic statistical concepts should be improved amongst all policy makers and users. In other words, there should be a positive movement towards basic statistical literacy.

ABSTRACTS OF THE PAPERS PRESENTED

1. Emerging Computing and Communication Techniques for Accelerated Growth in Dairying – An NDRI Perspective

D.K. Jain, A.K. Sharma and A.P. Ruhil

Globalization and growing competition have accelerated the need for knowledge intensive work performance in all the sectors of economy. In the dairy sector, constant application of latest ideas and better work technologies is essential to enhance productivity in the interest of economic well being of the stakeholders and for ensuring food security. Hence, knowledge acquisition has become a significant factor for accelerated growth of livestock sector in general and dairy sector in particular. It has acquired ever increasing importance in view of its significant contribution to the national economy, relatively higher growth in milk sector and its role as an employment provider during the stress period. To this effect, the Computer Centre at National Dairy Research Institute (NDRI), Karnal has made several efforts during the past two decades. The Institute started with a small facility for scientific data processing and gradually developed a Management Information System (MIS) for Animal Management. Ever since, it has expanded its activities to incorporate several advanced aspects of knowledge and information system for accelerating growth in Dairying. This includes development of web-enabled databases/software on various facets of dairying such as National Collection of Dairy Cultures; multimedia-based transferable technologies developed by the Institute; developing various expert systems and decision support systems as well as computational models based on emerging advanced computing and communication techniques/ technologies for precision dairy farming such as wireless sensor network for animal management for organized and unorganized herds, tracking nomadic herds for disease surveillance, soft computing models for intelligent predictions in dairy production and processing applications; dynamic and state-of-the-art Website on Animal Science, Dairy Science and Education along with transferable technologies, *etc*. The Centre is also in the process of implementing NAIP sponsored agroweb sub-project through which it is expected to strengthen the knowledge base of all the stakeholders including dairy farmers, feed and dairy industry and other potential users. Besides this, the Centre has designed and developed

Computer Curriculum for the Dairy Science students at undergraduate and postgraduate levels to equip them with latest computing and communication know-how.

National Dairy Research Institute, Deemed-to-be University, Karnal-132 001 (Haryana)

2. Accelerated Growth of Agriculture through Knowledge Management in Plant Genetic Resources

R.C. Agrawal

Plant genetic resources (PGR) for food and agriculture consists of the diversity of genetic material contained in traditional varieties and modern cultivars grown by farmers as well as wild relatives of crops and other wild plant species that can be used for food, feed, fiber, clothing, shelter, wood/timber, energy, etc. One of the major challenges for food security in the next generation is the effective management of plant genetic resources worldwide. Thus, knowledge management in plant genetic resources becomes very important at the national, regional and global levels to effective conservation of rapidly disappearing genetic stocks for possible future use and also for immediate utilization of already conserved and evaluated/characterized germplasm in the ongoing crop improvement programmes which can lead to the accelerated growth of agriculture. The recent advancement in information technology has led to an explosion in the compilation and collation of information in all fields, including PGR.

The National Bureau of Plant Genetic Resources (NBPGR) has been entrusted with the responsibility to plan, conduct, promote, coordinate and take lead in activities concerning germplasm collection, conservation, evaluation, introduction, exchange, documentation and sustainable management of diverse germplasm of crop plants and their wild relatives with a view to ensure their availability for use to the researchers. The Indian Plant Genetic Resources Management System (IPGeRMS) under the aegis of the Indian Council of Agricultural Research (ICAR), spearheaded by the NBPGR, is emerging as a dynamic system which holds prominent place among the global genebanks.

Genebank Information Management System (GBIMS), Plant Genetic Resources Passport Information Management System (PGRPIMS), Electronic catalogues for the recording of the evaluation/characterization data, database for the inventory of the import and export of the accession have been developed at NBPGR and the complete information related to Plant Genetic Resources as stated above is being documented using the Information Technology Tools for the PGR and the same is being used by the concerned PGR researchers for the knowledge management of the genetic resources in India for the conserved and evaluated/characterized germplasm in the ongoing crop improvement programmes.

National Information Sharing Mechanism on the Implementation of Global Plan of Action (GPA) for the Conservation and Sustainable Utilization of PGR for Food and Agriculture (PGRFA) (NISM-GPA) has also been developed in India for enhancing coordination of plans and activities on conservation and sustainable utilization of PGRFA amongst stakeholders and for sharing information, as well as for priority setting for the GPA implementation. A national network for the knowledge management in PGR is also being executed by the NBPGR to link all the National Active Germplasm Sites (about 50) in India.

National Bureau of Plant Genetic Resources, New Delhi-110 012

3. ICT for Accelerated Growth and Development of Indian Agriculture – Status and Prospects

Anil Rai, K.K. Chaturvedi and P.K. Malhotra

Introduction of personal computers (PC) in 70’s created vision of paperless offices, simplified computing, automated process control, management information system, expert system etc. Growth of companies with promises to provide specialist software to farmer and growers during 80’s is a clear indicator of high expectation and visualization of ICT in the field agricultural research and development. A review of agricultural software availability and its use has been made by an international team in 1997 called FARMSOFT (Gelb *et al.* 2006). This review was based on eight countries i.e. Belgium, France, Germany, Israel, Italy, Portugal, Spain and the Netherlands. It was found that in 1996 out of 1315 software related to agriculture, 39% were in category of management, 26% were in category of animal husbandry, 8.9% in category of crop production, 7.0% in category of machinery and process control, 0.1% in category of irrigation and 19% in other categories. Comparison of change in software inventory in 1996 with respect to 1994, it was found that there is around 15% growth in the agricultural software development. The highest positive growth was in the category of management software (38.40%) followed by animal husbandry (10.45%) and others (10.44%). However, highest negative growth was observed in the category of irrigation (-163.15%) followed by crop production (–53.85) and machinery and process control (–30.95). This clearly indicates that overall agricultural software inventory was growing but growth in the software related to agricultural management was sufficiently high whereas there is sharp decline in categories of irrigation and crop production. This change in pattern may be attributed to two main factors, first saturation of software and second bottleneck due to limitation of related technological innovations, socio-economic adoption and infrastructural facilities. Presently, there is increasing trend of using PC with shared resources via networks, audio and visual dissemination of data, information and knowledge in developed countries. Also, in these countries, there is significant shift in adoption of IT related technologies. There is end-user based demand driven situation for technologies such as computers embedded machines, controlled process and production, monitoring and evaluation, elementary decision making etc. Also, demand for tailor made software continues to increase along with strong network connectivity, climatic information, standardized software interface and ICT policy support. In this article attempt was made to discuss various important technologies of ICT, relevant to agricultural sector. Applications of ICT in the field of agricultural research and developments were also discussed. Technological linkages of ICT with supporting field were studied in context of India through patent analysis. Also, ICT knowledge network of Asian countries were discussed top highlight knowledge sharing among these countries. Further, international status related adoption of ICT for agricultural research and development and initiative taken in context of Indian agriculture has been discussed. Future prospects of ICT based technology has been projected based on past experience.

Indian Agricultural Statistics Research Institute, New Delhi – 110 012

4. A Decentralized Process for Web Based Data Management of Agricultural Education

R.C. Goyal

For data collection various methods like Observation, Interview, Questionnaire, Schedule, Case Study, Surveys and Panels etc. have been commonly used. Bowker (2000), Carbonaro (2002, 2000), Couper (2007, 2001, 2000) have supported and used the World Wide Web for the collection of data with strong perception that the web based data collection in academic research, extension and education might be replacing paper-and-pencil questionnaires in the near future.

This paper describes a process developed for implementing a decentralized system for web based data management of agricultural education. A practical application of the process for regular data collection, compilation and analysis of the data related to the day to day ongoing activities at the State Agricultural Universities (SAUs) and other organizations involved in imparting agricultural education in India will be presented. The advantages and limitations of the web based data collection, compilation and analysis approach may also be discussed.

The processes developed above will act as an independent data collection, compilation and analysis system at the organization level and will become a useful tool for the agricultural education data management at the universities and their affiliated/constituent colleges. Ultimately, it will lead to support and strengthen the National Information System on Agricultural Education Network in India (NISAGENET) hosted through http://www.iasri.res.in/Nisagenet/ at IASRI, New Delhi

Indian Agricultural Statistics Research Institute, New Delhi – 110 012

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 46-50

Symposium on Agricultural Statistics for Planning

Chairman: Dr. K.C. Seal
Convenor: Dr. H.V.L. Bathla

The following four papers covering aspects related with the theme of the Symposium were presented by the following speakers:

1. Quality of Agriculture Statistics: Role of NSSO — Dr. A.K. Yogi
2. Agricultural Statistics Available from National Accounts for Planning — Sh. Ramesh Kolli/Sh. Vidyadhar
3. Agriculture Statistics for Planning: Perspective, Planning of Land Use Statistics in Uttar Pradesh — Sh. Vinod Kumar Singh
4. Livestock Statistics and Data Gap — Sh. O.P. Misra

After detailed discussions, the following recommendations emerged out:

1. More exclusive investment is needed for primary data collection to get quality data.
2. High Level Co-ordination Committee meetings need to be made more interactive and effective to achieve better co-ordination of activities between centre and states.
3. The awareness of simple basic statistical concepts should be improved amongst all policy makers and users. In other words, there should be a positive movement towards basic statistical literacy.

ABSTRACTS OF THE PAPERS PRESENTED

1. Quality of Agriculture Statistics: Role of NSSO

A.K.Yogi

Agriculture continues to remain the predominant sector of Indian economy in terms of employment and livelihood, even though its share in Gross Domestic Product has declined from over 50 per cent in the initial years after independence to around 20 percent in the recent years. Growth rate of non-agriculture sector has also accelerated but strong agriculture- non-agriculture as well as rural-urban divide is seen in the society. In view of the seriousness of this issue, the 11th Plan Approach paper placed a strong emphasis on restructuring policies for achieving accelerated, broad based and inclusive growth. The Steering Committee on Agriculture and Allied Sectors for Formulation of the Eleventh Five Year Plan (2007-2012) highlighted the major concerns of the Indian agriculture and identified the causes underlying the present dismal state of agriculture and also suggested a road map for reviving agriculture with a view to placing it on high, inclusive growth path. The Committee also made recommendations for improvement in Agricultural Statistics.

At present the Directorate of Economic and Statistics, Ministry of Agriculture is the nodal agency for release of basic data on agriculture like area and production statistics for the country as a whole. While the area statistics are collected on complete enumeration basis in respect of permanently settled States and on the ad-hoc methods based on impressionistic approach in case of Assam (hill parts) and other N.E. States; and for the three states viz. Orissa, West Bengal and Kerala a scheme for Establishment of Agency for Reporting of Agricultural Statistics (EARAS) has been introduced. Crop Estimation Surveys (CES) are carried out following a specified sample design. Since the final estimates of production based on complete enumeration of area and yield through crop cutting experiments become available long after the crops are actually harvested, the Government prepares advance estimates of production for taking various policy decisions.

Role of NSSO in Quality Improvement

Timeliness and quality are the two essential features of any statistics. As far as the crop statistics are concerned, the Timely Reporting Scheme (TRS) is intended largely to take care of timeliness. As for quality, the scheme for Improvement of Crop Statistics (ICS) was introduced in rabi 1973. Now it operates in 20 States and 2 UTs. The scheme ICS was prepared jointly by the M/o Agriculture and NSSO. The programme envisaged under the ICS scheme includes:

1. Physical verification of crop enumeration done by the patwaris in a sample of 10,000 villages in each recognized season by NSSO and States on equal matching basis.
2. Checking about the accuracy of the area statistics transmitted from the village level through crop abstracts in the very same 10,000 villages; and
3. Providing technical guidance and supervision at the harvest stage in the conduct of about 30,000 crop cutting experiments distributed among principal crops in various States.

The role of NSSO confines to assisting the States in developing suitable techniques for obtaining reliable and timely estimates, providing technical guidance and ensuring adoption of uniform concepts, definitions and procedures in the Crop Estimation Surveys. It reviews the design, plans, details of implementation and the results of the surveys, participates in the training camps organized for the State field staff and supervises the primary field work undertaken by the staff of the state agencies. NSSO after processing of ICS data releases a comprehensive report "Review of Crop Statistics System through Scheme for ICS" for each season for each State. In addition another report called "Consolidated Results of Crop Estimation Survey on Principal Crops" is also released.

Findings of NSSO

Area enumeration

1. On an average timely submission of TRS statement is found in around 45% villages only. The timely submission is also found to be abnormally low in autum season, it improves in winter, rabi and falls in summer.
2. Submission of TRS Statement even without completing girdawari.
3. The errors are of three types viz. (i) crops actually sown are not reported , (ii) reporting of crops which are not sown; and (iii) incorrect reporting of crop area.

Crop cutting

1. Most of the States do not follow minimum number of experiments under CES i.e. 80-120 CC experiments for major districts and 40 experiments for minor districts.
2. The major errors observed relate to (i) selection of survey number/field, (ii) measurement of field and dimensions of plot, (iii) location of correct experimental plot, (iv) weighment of produce, (v) use of equipments.
3. The percentage of CC experiments found without errors varies from state to state and from season to season. For example less than 75% experiments observed to have been conducted without errors even in the major states like Maharashtra, Tamil Nadu, J&K and Rajasthan.

Inspite of regular reporting of weaknesses and drawbacks in season wise reports over the years, the State agencies have not taken serious measures to effect improvements in their work. Rather further deterioration continues.

Recent Reviews by Commissions and Committees

The National Statistical Commission (2001) examined the state of affairs of agricultural statistics and made several recommendations for improvements. Also the Steering Committee on Agriculture and Allied Sectors for Formulation of the Eleventh Five Year Plan (2007-12) of the Planning Commission headed by Prof. C.H. Hanumantha Rao commented on the system of agricultural statistics as "The agricultural statistics system has run down in many states. The conduct and supervision of crop cutting experiments has weakened, complete enumeration of land use and cropping and irrigation down to the plot level become difficult. The present status of implementation of various recommendations of the National Statistical Commission (NSC) clearly shows these recommendations have not been taken seriously by the concerned organizations". The also suggested that various recommendations by the NSC should be rigorously pursued and implemented at the earliest. The database for agricultural sector needs to be thoroughly reviewed for bringing lasting improvement in the basic system of Agriculture Statistics.

National Sample Survey Organization (FOD), Ministry of Statistics & Programme Implementation, Government of India, New Delhi

2. Agricultural Statistics Available from National Accounts for Planning

Ramesh Kolli

The Central Statistical Organisation (CSO) in the Ministry of Statistics and Programme Implementation, is entrusted with the responsibility of compiling National Accounts Statistics (NAS) for the country. In this role, the CSO releases annual national accounts statistics, quarterly estimates of Gross Domestic Product (GDP), and Input-Output Transactions Tables once in five years. The national accounts statistics provide a wealth of data on various sectors of the Indian economy and more particularly on agriculture sector.

The estimates of GDP for agriculture and allied economic activities are compiled by valuing the total production of each commodity in the country, and deducting inputs from this value of output. By virtue of this compilation, data on value of output at current and constant prices, state-wise and crop-wise becomes available.

The data base at all India level that is now available on value of output, inputs and value added is from 1950-51 to 2007-08, by crops, both at current and constant prices. At state level, crop-wise information on value of output at current and constant prices is available from 1960-61 to 2006-07. On livestock sector, such detailed data is available at state level from 1990-91 onwards.

With the above available data, the users can construct price indices, crop-wise and state-wise. This extensive and vast data is extremely useful in understanding the price movements of various crops in different states over a long period of time. Such detailed data is not available from any source, except from the national accounts statistics.

With the help of the same data source, users can also construct volume indices of production of agricultural crops (crop-wise) by states, as also for the livestock products. Analysis of these data provides production dynamics of various crops across the states.

From the demand side, the national accounts provide detailed data on investment made in agriculture and allied sectors, and again by institutional sectors, such as public sector, private corporate sector and private household sector.

The paper gives, along with a gist of datasets that are available from national accounts on agriculture and allied activities for the purpose of planning and decision making, a summary of data sources and estimation procedures adopted in the compilation of national accounts, and few summary tables at relatively aggregated level.

Central Statistical Organisation, New Delhi

3. Agriculture Statistics for Planning: Perspective Planning of Land Use Statistics in Uttar Pradesh

Vinod Kumar Singh

Man derives all his sustenance from the land and as land is limited, the need to make the most judicious use of land is imperative. It should be so used as to produce enough to satisfy the minimum need of mankind and remain enriched for posterity.

The Agricultural Statistics collected by the Department of Revenue, Uttar Pradesh should be used for planning the most judicious use of land.

Uttar Pradesh has about 242 lakh hectares of reporting area. The present land utilisation is somewhat imbalanced in the sense that very large proportion of area is under cultivation and relatively smaller area is under forest or orchards. Extent of barren and culturable waste is quite large of the total reporting area, about 7 percent is under forest, 68 percent is under cultivation and 4 percent is barren and culturable waste.

According to 2001 census, the population of Uttar Pradesh was 16.61 crore and is projected to be about 23.14 crore during the year 2020.

Production of food grains and other agricultural commodities is and will continue to be most important land use. The present production of foodgrains in the State is about 421 lakh tonnes. It is estimated that in the year 2020, a total quantity of 493 lakh tonnes will have to be produced annually to meet the requirement of the expanded population as well as to satisfy other demand, such as use for cattle feed, seeds, wastage in storage, etc. Requirements of oil seeds, sugarcane and other agricultural products are expected to increase at a faster rate.

As the area under forests, groves and pastures are already less and as non-agricultural use of land such as for expansion of urban communities, roads and railroads, dams and canals, industries and public utilities; it is proposed to restrict the net cultivated area to near about the existing percentage i.e. 68 percent. Additional agricultural production is proposed to be obtained by increasing the intensity of cropping from the present 153.34 percent to about 180 percent as also by increasing per hectare yields. This is proposed to be achieved by the required increase in agricultural inputs comprising irrigation, fertilizer improved seeds, pesticides etc., and by the programme of extension and education.

Lack of adequate and proper knowledge of our soil resources and of the problems caused by erosion, floods, water logging and salinity has hindered the preparation of comprehensive land use plan. It is proposed to survey the entire State within the next 2 years for which Soil Survey Organisation will have to be further strengthened. Even though increase in net cultivated area is not proposed, barren and usar lands have to be reclaimed on a priority basis in order to substitute for agricultural land necessarily being put to non-agricultural uses. It is also necessary to reclaim it for adding to grove lands, pastures and forests. Reclamation of such lands and for ravines is also necessary to prevent their further march and encroachment into agricultural lands.

Programme of land reclamation on such large scale would necessitate involvement of people. This would require research, demonstration and training.

Irrigation facilities are available from the state and private sources in an area of about 133.13 lakh hectares against the gross cropped area of 254.15 lakh hectares. The existing irrigation facilities are, generally not adequate for intensive agriculture. The irrigated area will require to be increased to 170 lakh hectare by the year 2020 in order to provide for need of agricultural production envisaged at that time.

Balanced and conjunctive use of surface water and ground water is proposed in order to prevent water logging on the one hand and excessive depletion of underground water by over pumpage on the other hand. The State tubewells will also be needed to augment canal supplies in areas where requirement cannot be met otherwise.

The area under forests is about 16.57 lakh hectares which is about 7 percent of the total reporting area. According to the National Forests Policy the area under forests should be about 20 percent. Considering other demands on the land, increase in forests area to 20 percent does not appear feasible. Forest area is, however proposed to increase to about 36 lakh hectare, i.e. 15 percent of the total area by the year 2020. The additional forest area will be found by reclaiming usar and barren lands.

To ensure the desired level of yield rates, adequate agricultural inputs will have to be provided.

Department of Agriculture, Government of Uttar Pradesh, Lucknow

4. Livestock Statistics and Data Gap

O.P. Misra

Reliable data base of Livestock Statistics plays an important role in formulation of various livestock development programmes in the country. The basic statistics on livestock include the population of livestock in terms of breed, sex, age, composition, distribution of livestock by size of land holdings, output of different livestock products, and by–products, marketing of livestock products, infrastructure facilities in the form of various farms in State, Veterinary hospitals, artificial insemination centres etc., incidence of livestock diseases, feed and fodder statistics, consumption pattern of livestock products, import and export of livestock and related products, etc.

The two main sources of livestock data includes livestock census being conducted on quinqunial basis since 1919 and estimation of production of major livestock products through conduct of integrated sample surveys. At present 18th Livestock Census with reference date 15th Oct., 2007 is in progress. Most of the States/U.Ts have completed the field work relating to census. Data entry is in progress. In order to ensure that the data can be collated at all India level, the National Informatic Centres (DADF) have been assigned the task of providing the data entry module for entry of data and related software for generation of various tables as per the tabulation plan from the census data. The 18th Livestock Census results are likely to be available by

March 2009. During the current census, lot of emphasis has been placed on collection of breed wise data. For the first time the country will have its breed wise population of livestock. This would also help in protecting the endangered species of animal i.e. such breeds which are on the verge of extinction.

In order to ensure effective implementation of the Integrated Sample Survey Scheme during the 11th Five Year Plan, following decisions has been taken:

1. Funds to meet 75% expenditure on salary of Officers/Staff would be provided under the scheme to the States by the Govt. of India.
2. An Office of Asstt. Director (Livestock Statistics) would be created for a group of every 4 districts in all States/UTs with two Supervisors-cum-Data Entry Operators and 4 Enumerators.
3. For “Web based Solution” all the Asstt. Director (Livestock Statistics) and Supervisors-cum-Data Entry Operators would be provided with a computer and necessary peripherals.
4. Studies would be conducted to fill up the data gap in livestock sector.
5. The Officers/Staff would be trained about the methodology, selection of sample, data collection, data transfer, etc. at regular intervals.
6. The time lag in finalization of all the estimates should be brought down to 3 months.
7. All States/UTs would be asked to collect, maintain and provide cost of production data on regular basis.
8. A Committee consisting of representatives from DADF, IASRI, Planning Commission and NIC would be constituted to revise the schedule/format for collection of data and methodology for conduct of sample surveys.

Department of Animal Husbandry, Dairying and Fishery, Ministry of Agriculture, New Delhi

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 51-64

Abstracts of Papers Presented

1. Nested Balanced n-array Designs and their PB Arrays

H.L. Sharma[1], R.N. Singh[2] and Roshni Tiwari

The present paper deals with the construction of nested balanced n-array designs and their PB arrays through a set of (n – 1) balanced incomplete block designs. An application in relation to intercropping designs through PB arrays has also been added.

[1] *J.N. Agricultural University, Jabalpur*
[2] *Agricultural Research Institute (RAU), Patna*

2. Construction of a Series of Ternary Group Divisible Designs

H.L. Sharma and Roshni Tiwari

In the present paper, we exhibit a new method for construction of a ternary group divisible (TGD) design using a series of regular group divisible (RGD) and L_2 designs. The association scheme of the constructed TGD is the same as that of corresponding group divisible design. The authors claim that TGD based on RGD and L_2 designs would be certainly useful in fractional factorial plans and certain related problems.

J.N. Agricultural University, Jabalpur

3. Statistical Evaluation of Rainfed Practices of Crops from Long–term Experiments Based on Multivariate Modeling and Analysis

G.R. Maruthi Sankar, P.K. Mishra and G. Ravindra Chary

Long–term field experiments (more than 15 years) have been conducted in fixed sites at different All India Coordinated Research Project for Dryland Agriculture (AICRPDA) locations during 1984 to 2007. Multivariate statistical models and analysis have been explored for the data recorded on variables of different factors over years viz., (i) weather; (ii) soil; (iii) crop; and (iv) controllable variables like organic and inorganic fertilizer through different sources. The weather factor included variables like (i) daily, weekly, monthly, crop growing stages, seasonal (kharif and rabi) and annual rainfall; (ii) number of rainy days; (iii) minimum and maximum temperature; (iv) relative humidity (morning and after noon); (v) evaporation; (vi) length and duration of dry spells occurred from sowing to harvest; (vii) water use efficiency. The soil factor comprised of variables like (i) soil moisture available at sowing, different crop growing stages and harvest; (ii) soil fertility of N, P, K and sulphur nutrients. The crop factor included variables like (i) date of sowing and harvest; (ii) crop growing period; (iii) plant uptake of N, P, K and sulphur nutrients; (iv) grain yield at harvest. The controllable factor included variables like (i) organic and inorganic fertilizer treatments; and (ii) variety. In order to homogenize the data, a classification of data has been made based on different rainfall groups viz., < 500 mm (arid); 500 – 750 mm (dry semi-arid); 750 – 1000 mm (wet semi-arid); and 1000 – 1250 mm (dry sub-humid); > 1250 mm (moist sub-humid) observed in different years. The detailed study has been explored with the objective of assessing the efficiency of input fertilizer treatments over years for (i) attaining sustainable crop yield and monetary returns; (ii) minimizing cost of cultivation of crops; and (iii) maintenance of maximum soil fertility of nutrients after harvest of crops.

Multivariate statistical assessment has been made based on the input–output correlations of variables within a factor and also between factors. The sustainability of treatments was examined by de-trending the crop yields with efficient estimates of 'prediction error' measured based on statistical models of yield calibrated through variables of each of the four factors examined. The crop yield models were also calibrated with variables of different factors under each rainfall group and also over rainfall groups for different crops tested at different locations. The sustainability was measured based on 'Sustainable yield index (SYI)' as described by Vittal *et al.* (2002) and modified by Maruthi Sankar *et al.* (2007). An assessment of convergence of the mean yield attained by a treatment over years to the potential or maximum yield attained in the study period was made. Ranks were assigned to the statistical parameters of mean, coefficient of variation, coefficient of determination (R^2), prediction error, SYI, apart from superiority of a treatment for the yield attained in individual years and maintenance of soil fertility. Based on the rank statistics, efficient fertilizer practices were identified for rainfed crops grown at different AICRPDA locations. Optimum fertilizer doses were derived at varying soil fertility and moisture levels. The statistical results of (i) sustainability yield index of fertilizer treatments; (ii) efficient fertilizer treatments based on rank analysis; and (iii) optimal fertilizer doses for attaining sustainable yield of rainfed crops viz., sorghum (Solapur and Kovilpatti); pearl millet (Kovilpatti and Rajkot); groundnut (Rajkot and Anantapur); finger millet (Bangalore); cotton and green gram (Akola); maize (Arjia, Indore and Rakh Dhiansar); black gram (Arjia, Phulbani and Rakh Dhiansar); soybean (Indore); rice and pigeonpea (Phulbani) are discussed in this paper.

Central Research Institute for Dryland Agriculture, Hyderabad

4. Computer Aided Generation of Linear Trend-free Designs for Factorial Experiments

Susheel Kumar Sarkar, Krishan Lal, Rajender Parsad and V.K. Gupta

The present article deals with development of algorithms for computer aided generation of designs for 2^k factorial experiments (without and with confounding) that permits the estimation of main effects free from linear trend present in the experimental units. The algorithms have been developed using the criterion of component-wise product. The procedures of identifying 2- and 3-factor interactions that are linear trend free have also been incorporated in the algorithm. In factorial experiments as the number of factors and/or levels increase, the number of treatment combinations becomes so large that it is not possible to accommodate them without losing homogeneity within block and one has to go for confounding in factorial experiments. Confounded factorial experiments are frequently used in agricultural experiments. Thus above algorithms have also been extended to confounded factorial experiments to generate confounded factorial experiments with any number of factors k (≥3) that are linear trend-free for main effects and identify two and three factor interactions that are linear trend free.

Indian Agricultural Statistics Research Institute, New Delhi

5. Estimation of Heritability of Mastitis Disease using ANOVA Method

Subrat Keshori Behera, A.K. Paul, S.D. Wahi and Pal Singh

Mastitis disease being a threshold character in dairy cattle breeding needs an in-depth study of its inheritance. The Analysis of Variance Estimator method and a modified method given by Fleiss is being applied to find the heritability of mastitis disease. The concept of intra class correlation coefficients has been used in calculating heritability. The results obtained showed the similarity with the results sighted in literatures.

Indian Agricultural Statistics Research Institute, New Delhi

6. A New Method of Construction of Neighbour Balanced BIBRC Designs

Anurup Majumder and Aatish Kumar Sahu

A new method of construction of a series of BIBRC (v, b, r, p, q, λ) designs, where v is a prime or prime power, is provided. The series of BIBRC designs is developed from mutually orthogonal mates (MOM) of BIB designs. The BIBRC designs obtainable from the above mentioned method are neighbour balanced as they satisfy the condition that every pair of treatments occurs in neighbouring plots n_r times in rows, n_c times in columns and n_d times in diagonals. Available literature reveals that the developed neighbour balanced BIBRC designs require a lesser number of sets for particular values of v, p and q. Moreover, the series of the BIBRC designs is more general, as it includes the series ($v = tpq + 1$, $b = tv$, $r = tpq$, p, q, and $\lambda = (p - 1)(q - 1)$ of Agrawal and Prasad (1982), as a special case. Also it should be noted that the designs of Agrawal and Prasad (1982) are not neighbour balanced.

Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia

7. Genotype × Environment Interaction and Yield Stability in Mustard (*Brassica juncea L.*) - A Statistical Approach

M.K. Chaudhari and B.H. Prajapati

Eight genotypes of mustard were evaluated for stability at six different environments (Locations) in Gujarat during the rabi seasons of the years 2003-04, 2004-05 and 2005-06 and significant G × E interaction for seed yield and its attributes were noticed. The major portion of G × E interaction was linear in nature for all the characters indicated that prediction of genotypes tested would be possible. Differential response of genotypes to environments was observed in year-wise as well as pooled analysis for seed yield and its attributes, hence stability analysis was carried out. Based on mean seed yield and stability parameters, G_5 (SKM 9927), G_1 (SKM 0139) and G_8 (GM-2) showed higher seed yield, unit regression coefficient (b_i = 1) and least deviation from regression $(\bar{S}^2_{di} = 0)$. Hence genotypes G_5 (SKM 9927), G_1 (SKM 0139) and G_8 (GM-2) are found stable and high yielding genotype for different mustard growing areas of the state.

S.D. Agricultural University, Sardar Krushinagar

8. Computer aided Generation of Linear Trend-free Response Surface Designs

Susheel Kumar Sarkar and Krishan Lal

The treatment combinations of the ordinary full factorial need not be the best for fitting the relationship. Therefore, it is necessary to search for a suitable set of treatment combinations by using which a stipulated relation can be fitted. The special class of designed experiments for fitting the response surfaces is called response surface design. Response surface designs have wide applications in agricultural, biological and industrial experiments. Similar to factorial experiments, experimental units in response surface design may exhibit trend over space or time. Among response surface designs, Box-Behnken design has been studied and linear trend-free design has been obtained. The developed algorithm helps the experimenters who are conducting quantitative factors using response surface designs. There may be a trend in the experimental material and hence we need trend-free design. It provides a complete solution in the sense that it is capable of generating the trend free Box-Behnken design. Trend-free designs are quite useful for such experimental situations. But the construction for such design is not easily available. It is, therefore, required to give easy method of construction, possibly computer aided for the construction of these designs. Thus, algorithms have been developed to generate complete factorial experiments each at two levels with any number of factors k (^{3}3) that are linear trend-free for main effects using the criterion of component-wise product.

Indian Agricultural Statistics Research Institute, New Delhi

9. Study of Correlation between Various Body Measurements among Different Types of Deoni Cattle

M.M. Appannavar and M.D. Suranagi

The data collected on various body measurements of different types of Deoni breed of cattle were analyzed to know the degree of correlation between different types of Deoni breed of cattle as Deoni breed of cattle over the years have arisen as morphological diversions viz., Balankya: The animals with complete white body coat and without any spot on the body, Wannera: The animals with white body and black shades on sides of the face and Waghya: the animals with white and black spot/ patches scattered all over the body. The correlation between Body Length and Height, Height and Chest Girth and Body Length and Chest Girth among Balankya were 0.9053, 0.8746 and 0.9388 respectively. Similar correlations among Wannera were 0.8793, 0.8122, 0.9330 and in Waghya were 0.8538, 0.8138, 0.9268 respectively. Though the correlations between various body measurements of all the three were highly significant, but degree of correlation were higher among Balankya and Wannera which were truly resembling Deoni breed of cattle than Waghya which had white and black spot/patches scattered all over the body.

Karnataka Veterinary, Animal and Fisheries Sciences University, Bidar

10. Estimation of Reliability in Non-accumulating Damage Shock Model (Random Threshold)

M.D. Suranagi[1] and S.B. Munoli[2]

A component is subjected to a sequence of shocks occurring randomly in time as events of a Poisson process. The shocks damage the component and the amount of damage due to a shock is an exponential random variable. The component fails whenever the damage due to a shock exceeds the random threshold of the component, otherwise the component works as good as new one (non-accumulating damage shock model). The reliability function of the model is derived. The maximum likelihood estimator (MLE) and Bayes estimator of the reliability functions are obtained.

[1] *Karnataka Veterinary, Animal and Fisheries Sciences University, Bidar*
[2] *Karnataka University, Dharwad*

11. A Study on Expenditure of Patna Rickshaw Pullers

R.N. Singh[1] and R.C. Bharati[2]

A survey of expenditure pattern on rickshaw pullers of Patna was conducted. Since rickshaw pullers are of low-income group, lognormal distribution was fitted. The fit was found to be satisfactory. This study is based on primary data of one thousand rickshaw pullers belonging to Patna capital city.

[1] *Agricultural Research Institute, Patna*
[2] *ICAR-RCER, Patna*

12. Population Dynamics of Cattle in India

Shiv Prasad and Rajendra Singh

The total cattle population showed increasing trend from 1951 (155.29 million) to 1992 (204.58 million) and then decreased continuously, reaching to 185.18 million in 2003. There was remarkable growth in case of exotic/crossbred cattle during reported period (1982-2003). Thus, the decrease in total cattle population was due to decrease in population of indigenous cattle. There was increasing trend in sex ratio over years in all categories considered here, showing that male population was not sustained in comparison to female population. It was higher among young cattle (less than 3 years) as compared to adult cattle in all the census years, showing that the rate of disposal of adult female cattle is higher than male cattle. The sex ratio in exotic/crossbred cattle was higher than indigenous cattle during the reported years from 1982 to 2003. It shows that exotic/crossbred male cattle are generally disposed by the farmers due to their worthless in agricultural operations. The sex ratio in exotic/crossbred cattle varied from 1508 to 2418 in young and 1080 to 6574 in adult during 1982 to 2003. It is revealed from this observation that people are adopting crossbreeding program but do not like to keep crossbred male cattle. The negative growth rate was recorded during 1966-72 in males (-0.14% for under 1 year and –0.80% for 1 to 3 years age) and females (- 0.75% for 1 to 3 year age), 1977-82 in males under 1 year (- 0. 46%) and over 3 years (-0. 62%). After division of cattle population as indigenous and exotic/crossbred cattle in the year 1982, the negative growth was observed in adult crossbred males during 1982-87 (- 2.38%) and 1997-2003 (-6.09%) and in indigenous males of 1 to 3 years (- 0.01%) and over 3 years (- 0.09%) during 1987-92. After 1992, it is observed that there was negative growth in indigenous male and female populations of each age group except males under one year during1992-97 (0.05%) and females during 1997-2003 (0.07%). The negative growth rate was recorded in case of male cattle used for breeding purpose during 1987-2003 and in working males during 1977-2003. No growth model was found suitable with R^2 greater than 90% to describe the growth pattern of cattle population. However, the multiple regression equation of cattle population on time, buffaloes population (lakh) and number of tractors (thousands) was found fit to explain the variability in cattle population with $R^2 = 94.69\%$.

Indian Veterinary Research Institute, Izatnagar

13. Factors Affecting the Adoption of Crop Insurance in Bihar

N.K. Azad[1], R.C. Bharati[2], S. Chakraborti[1] and S.P. Singh[3]

To identify the factors affecting the adoption of crop insurance, micro-level study was undertaken in Bihar with 600 farmers. Multistage Stratified Random Sampling was used for selecting two villages in each of the three Agro-Climatic Zones. From each selected village, 100 farmers were interviewed. Based on the data collected on age, education and category, multiple regression equation was fitted. Age, education and category contributed significantly on adoption of crop insurance. Age being the most important factor of adoption contributed about 50% towards adoption. It was observed that younger farmers with larger land holding adopt more crop insurance. It was also observed that with the increase in bank branches, adoption rate increased (r = 0.982). Further, it was found that the factors age, education and category were not independent.

[1] *Palli Siksha Bhavana, Sriniketan*
[2] *ICAR-RCER, Patna*
[3] *RAU, Pusa*

14. Forecasting Models for Sunflower Production of Andhra Pradesh in Presence of Shifts - A Comparative Study

K. Alivelu[1], B.S. Kulkarni[2], G. Ramakrishna Rao[3], S. Ravichandran[4] and C. Sarada[1]

Study of shifts in cropping pattern over the years help in understanding the rationale in allocation of area under the crops. Similarly information on advance forecast of crop production also help in strategic planning. Statistical methods that are in vogue for dealing with these issues have its own limitations. A multivariate procedure based on cluster analysis has been proposed to identify the years of discontinuity in the data. These shifts also hamper the applicability of the conventional models for forecasting. Shifts disturb the ‘continuity’ in the year-to-year variations in the data. Two alternative models that favour the ‘discontinuity’ in the year-to-year variations were explored for forecasting the sunflower production. These models were the Spline Regression and Artificial Neural Networks. The methodology was applied to the 35 years of crops data of Andhra Pradesh State covering the years 1970-71 to 2005-06. The analysis revealed that the proposed cluster analysis procedure was effective in identifying the shifts as well as the years of discontinuity in the crops data. A neural network model with 9 neurons in the hidden layer was found to have minimum residual variance. These results indicated that ANN had relatively lower values of goodness of fit statistics and bias. Hence, ANN model provided a better fit for this production series and was used to generate the forecasts. The forecast for 2006-07 as given by ANN was 3.25 lakh tonnes.

[1] *Directorate of Oilseeds Research, Hyderabad*
[2] *ANGRAU, Hyderabad*
[3] *JNTU, Hyderabad*
[4] *DRR, Hyderabad*

15. A Composite Index to Select Castor Genotypes using Multi-stage Principal Component Analysis

C. Sarada, P. Lakshmamma, K. Alivelu and Lakshmi Prayaga

The study explored multi-stage principal component analysis for selecting the efficient castor genotypes with high performance in terms of shoot and root traits. The principal component analysis is utilized for the study as it would give scope for selecting the traits with higher loadings in the extraction of variance in the traits. 10 shoot traits viz., plant height (cm), number of nodes, leaf number, number of secondary branches, number of tertiary branches, stem girth (cm), total dry matter (TDM), leaf area index (LAI), spad chlorophyll meter reading (SCMR) and specific leaf area (SLA) and 6 root traits root length, root volume, root dry weight, root density, root to shoot weight ratio, root to shoot length ratio are considered for the study. In the first stage indices were developed separately for shoot and root traits using principal component analysis based on the correlation matrices. Thus, developed indices were utilized as variables for second stage principal component analysis to develop a composite index for selecting the castor genotypes. The genotypes were ranked based index in ascending order to select the efficient genotypes with high ranks. The results indicated that the developed index performed well in selecting the genotypes with high performance with respect to shoot and root traits considered.

Directorate of Oilseeds Research, Hyderabad

16. Inter-regional Variation in Crop Production in Uttar Pradesh

Sandeep K. Sharma[1], Virendra P. Singh[2] and R.K. Tomar[1]

The present study envisages the inter-regional variation in crop production in Uttar Pradesh and tries to identify the most agriculturally developed regions in the State. For the analysis, the State was divided into four economic regions, Eastern, Central, Western, and Bundelkhand. Twelve indicators like per capita net area sown, percent area and productivity of food-grains, per capita production, percentage of gross irrigated area to gross sown area, irrigation intensity, fertilizer consumption, cropping intensity, etc. were selected, based on which a composite development index of crop production for each region of Uttar Pradesh was worked out following the Taxonomic method developed by a group of Polish mathematicians in early 1950s. Many disparities were observed between the regions in terms of these indicator values, but the Central and Western regions showed relatively better situations as compared to other regions. Based on the composite index, the study reveals that the Central region is the most developed region in Uttar Pradesh so far as agricultural development is concerned.

[1] *Indian Agricultural Research Institute, New Delhi*
[2] *NCERT, New Delhi*

17. Time Series Modeling for Forecasting Seasonal Paddy Yield of Tamil Nadu

Ramasubramanian V.[1], Chandrahas[1] and A. Dhandapani[2]

Timely availability of reliable forecasts of crop statistics are crucial in deciding about shift in production patterns, farmers' choice, impact of government policies, etc. One among the various statistical approaches for forecasting crop size includes time series models especially for short term. The paper deals with development of forecast models based on time series approaches for obtaining seasonal paddy yield forecasts of Tamil Nadu state. The various models considered are the general linear Gaussian state space, exponential smoothing via state space, exponential smoothing and ARIMA. Three seasons per year yield data for the period 1987-2006 have been taken up for model fitting and validation. The forecasting performance of various models was judged on the basis of Mean Absolute Percentage Error and Percent Root Mean Squared Error. For forecasting paddy yield of Tamil Nadu, general linear Gaussian state space model has come out to be the best followed by exponential smoothing via state space and then by seasonal ARIMA and by exponential smoothing. The study revealed that state space models can be employed as a viable alternative for forecasting paddy yield of Tamil Nadu.

[1] *Indian Agricultural Statistics Research Institute, New Delhi*
[2] *National Centre for Integrated Pest Management, New Delhi*

18. GARCH and EGARCH Nonlinear Time-series Modelling and Forecasting of Volatile Data

Prajneshu and Himadri Ghosh

Two parametric nonlinear time-series models, viz. Generalized Autoregressive Conditional Heteroscedastic (GARCH) and Exponential GARCH (EGARCH) models are thoroughly studied. A heartening feature of these models is that these are capable of describing volatile data sets. Procedures for estimation of parameters of these models are also discussed. As an illustration, these models are applied for modeling and forecasting of all India monthly export data of fruits and vegetables seeds using EViews, Ver. 4 software package and by writing computer programs in C. Comparative study of the fitted models is carried out from the viewpoint of dynamic one-step ahead forecast error variance along with Mean Square Prediction Error (MSPE), Mean Absolute Prediction Error (MAPE), and Relative Mean Absolute Prediction Error (RMAPE). It is concluded that for data set under consideration, EGARCH model has performed better than GARCH model for both modelling and forecasting purposes.

Indian Agricultural Statistics Research Institute, New Delhi

19. GRAMBUG: An Expert System for Chickpea (Gram) Insect-pests Management

Devraj[1] and Renu Jain[2]

Chickpea (Gram) is the most important pulse crop of India accounting for 29% of area and 38% of production of total pulses in the country. Madhya Pradesh is the highest producer of chickpea followed by Maharashtra, Rajasthan, Andhra Pradesh, Karnataka, and

Gujarat. In spite of all the advances made in the crop protection technology, around 18-20 percent crop is annually lost due to ravages of insect-pests. This crop is prone to attack by a large number of insect-pests damaged right from seedling to maturing and in storage. However, only few of them (viz., Gram Pod Borer, Cutworm and Termites) are of economic importance inflicting serious yield losses in specific locations and seasons in India. When a crop shows symptoms, it is important to make a correct diagnosis to support appropriate control measures. Diagnosing insect-pests in Chickpea requires considerable expertise. Only a few experts have the ability to do this job, and each expert has his own specific domain. To retain expertise and to make it more accessible, an expert system, called GRAMBUG, has been developed.

In this paper, we present the design, the structure of the knowledge base and the inference mechanism used in GRAMBUG. GRAMBUG is an automatic identification tool that can help farmers and extension workers to identify major insect-pests and suggest the appropriate treatments. The knowledge was obtained from literatures, farmers, extension workers and from experts using automatic knowledge acquisition interface. Information stored in the database is presented to users through a series of interactive question-answer sessions. The selection of next question to be asked is done dynamically on the basis of symptoms given by the user through previous questions. The user input is matched with the expert's data and on that basis system predicts all possible insect-pests along with their reliability factor. Digitized photographs of insect-pest symptoms are shown by the system to support an interacting session for final diagnosis of the predicted insect-pests. The system performance has been tested at Indian Institute of Pulses Research for the diagnosis of Chickpea insect-pests. GRAMBUG system for insect-pest diagnosis was found quite effective in terms of time and cost in addition to correct insect-pest identification. The methodology can also be applied to other pulse crops without making the design changes in the system.

[1] *Indian Institute of Pulses Research, Kanpur*
[2] *Krishna Girls Engineering College, Kanpur*

20. Establishing Campus Area Networking (CAN) using a Combination of Wireless and Wired Connectivity – An Optimum Solution

R. Ganesan and C.R. Girija

Computer networking is undergoing drastic transformation in response to end users' need for mobility and connectivity while exchanging information among computers. There are two basic ways by which a computer network can be established - "wired" using Ethernet cables or "wireless" using radio waves, also known as "Wi-Fi." The technology of wireless connectivity has come of age, and is now a viable, low cost alternative to the traditional wired technologies. Wireless technology gives users the mobility to move around within a reasonably broad coverage area and still be connected to the network. Most wireless LANs today use the 2.4-gigahertz (GHz) or 5 GHz frequency band, the only portion of the RF spectrum reserved around the world for unlicensed devices. The freedom and flexibility of wireless networking can be applied both within buildings and between buildings. In many places, combination of wired and wireless technology is used to meet all the networking needs. As the market for wireless LAN is continuously increasing, this paper discusses an overview of wireless technology and also how the Campus Area Network (CAN) has been optimally established using a combination of wired and wireless technology at Rajiv Gandhi College of Veterinary and Animal Sciences (RAGACOVAS) campus, Puducherry connecting computers located in nine buildings within the campus of 1 km radius.

Rajiv Gandhi College of Veterinary and Animal Sciences, Puducherry

21. Information Technology and Networking for Enhanced Oilseeds Productivity

S.V. Ramana Rao, P. Madhuri, R. Venkattakumar and M. Padmaiah

The revolution in the Information Technology and the intensified pace of versatility of the same is an important growth engine for Indian agriculture in general and oilseeds in particular. The domestic oilseed economy predominantly rainfed is in a jinx due to market and market forces. The application of tools viz., GIS, DSS etc. make them vital for furthering productivity through the "Geo-referenced" approach. Further, AER based specific ramifications for enhanced sustainability of oilseeds based production system in the long run can be brought out through' application of IT. Proper DSS results in addressing to specific biotic and abiotic issues at micro/macro level realizing in enhanced productivity of oilseeds thereby making oilseeds more globally competitive due to economies of scale. Enormous potential prevails for linkage mechanisms in input and output marketing through PPP. The networking through IT can help in improving the supply chain/procurement management of oilseeds which can enhance the processing efficiency and assure greater availability of domestic edible oils by safeguarding the producer and the consumer. This calls for institutional refurbishments by policy planners for reaping the benefits of IT for increasing productivity of oilseeds thereby reducing the burden on the huge exchequer on imports.

Directorate of Oilseeds Research, Hyderabad

22. SSDA: A C# Library for Analysis of Sample Survey Data

Anu Sharma, S.B. Lal and V.K. Mahajan

SSDA is an object oriented C# library for analysis of survey data. It implements the logic of standard procedures for the estimation of parameters for various sampling designs within a framework designed to be easy to use, extend, and integrate with other .net compatible software tools. This library could be easily customized and extended by adding new modules. This library is available in the form of dynamic link libraries (.dll) and can be called by adding their reference in the software project. This reusable library is highly useful for programmers and statisticians involved in statistical software development. A software has been developed using these methods for the survey data analysis.

Indian Agricultural Statistics Research Institute, New Delhi

23. Class of Unbiased Modified Ratio and Product Estimators for Finite Population Mean using Coefficient of Variation of Auxiliary Variable in Sample Surveys

Virendra Pratap Singh

This paper presents classes of unbiased ratio and product type estimators for finite population mean of study variable using coefficient of variation (CV) of auxiliary variable employing jackknife technique envisaged by Quenouille (1956). Asymptotic expressions for variance formulae are derived. Numerical illustrations are provided to examine the performance of the proposed estimators.

National Council of Educational Research and Training, New Delhi

24. Estimation of Soil Nutrients using Remote Sensing and Geographic Information System

K.N. Singh[1], Abhishek Rathore[1], Karan Singh[2], A.K. Tripathi[1], A. Subba Rao[1] and Salman Khan[1]

Optimum return on the investment and minimum environmental pollution are major issues to be addressed while giving soil test based nutrient recommendations. A comprehensive knowledge of the basic soil resources is of fundamental importance for efficient land use planning. Green revolution by using high yielding varieties and improved management technology has increased crop production at the cost of soil productivity and possible risk of soil degradation. Decrease in the soil fertility and imbalanced use of nutrients are important factors responsible for stagnation or decrease in the crop yield over the years. Thus, it should be firmly understood that further increase in food production must be attained by judicious use of soil resource base. Fertilizers being the costliest inputs, the scientific approaches towards profitable agriculture would imply the use of plant nutrients according to the actual needs of the soil-crop situations. An attempt has been made to prepare soil fertility maps using remote sensing data (Imagery). For proper recommendation of fertilizer applications for different crops, the knowledge of fertility status is essential. However, it is not always possible to collect soil samples from every location and analysed it for different nutrients. Cloud free data for the months December 2005, January, February and March 2006 of Hoshangabad district was considered for the study. The result showed that most of the relationships between different nutrients and mean of NDVI values were significant for all the nutrients. This is preliminary indication that the nutrients can be predicted with the help of remote sensing data. The February month may be more appropriate for this purpose.

[1] Indian Institute of Soil Science, Bhopal

[2] Central Institute of Agricultural Engineering, Bhopal

25. Methods of Estimation of Crop Production for Small Areas

B.V.S. Sisodia, Anupam Singh and L.K. Dube

Crop production statistics are generally estimated at district level through crop-cutting experiments. Yearly time series data on crop-production at district level are available in various official bulletins. However, need has been felt to have crop production statistics at block/ panchayat level for formulation of various agricultural development programmes and also for crop-insurance schemes. Some statistical prediction techniques have been proposed to develop estimators for the crop production at block level by postulating regression model using the time series data at district level in the present paper. The variances of the estimators are derived and their relative efficiencies have been theoretically examined. An empirical study has also been carried out to illustrate the utility of the proposed methods.

Narendra Deva University of Agriculture & Technology, Faizabad

26. On Robust Estimation in Stratified Sampling under Super-population Model

R.P. Kaushal, B.V.S. Sisodia and Sunil Kumar

Following Royall and Herson (1973b), a BLU predictor of population total under the model $\xi(0, 1: x_{hk})$ in stratified sampling, the model is common across to strata, is constructed. Its robustness and optimality is studied when some general polynomial model of degree *J*, i.e. $\xi(\delta_0, \delta_1, ..., \delta_{\mathrm{j}}; x_{hk})$ is true in real practice. It has been found that the proposed predictor is robust and optimal for stratified balanced sample and is also more efficient than that of due to Royall and Herson (1973b) when slopes are varying from stratum to stratum under certain conditions.

Narendra Deva University of Agriculture & Technology, Faizabad

27. A General Class for Estimation of Population Mean when $\overline{\mathrm{X}}$ is not Known

Manish Sharma[1] and Sharad Bhatnagar[2]

For estimating population mean, the ratio and product methods are not operational when $\overline{X}$ is not known. A simple alternative is then to employ double sampling in which $\overline{X}$ is first estimated from a large preliminary sample. It is pointed out that this may not be often feasible. For such situations, a class of estimators is proposed and its large sample properties are studied with the ratio method, product method and other estimators.

[1] *Sher-e-Kashmir University of Agricultural Sciences & Technology-J, Jammu.*

[2] *CCS Haryana Agricultural University, Hisar*

28. Estimators for Farming Practices, Resources and Activities

Jagbir Singh, K.K. Tyagi, K.K. Kher, A.K. Gupta and V.K. Jain

In this paper an attempt has been made to develop Minimum Variance Linear Unbiased Estimators (MVLUEs) for the parameters under the study “Estimation of extent of farming practices, resources and activities with energy use” for kharif and rabi seasons of the agricultural year 2002-03 by making use of Projective Geometry approach. The MVLUEs have been obtained for the parameters viz. (i) extent and seasonal variation of use of farming practices (such as land use for different kinds of farming etc.), (ii) extent and seasonal variation of available farming resources (such as fertilizers etc.) and their use and (iii) extent of farming activities (such as ploughing etc.).

Indian Agricultural Statistics Research Institute, New Delhi

29. Estimation of Small Area Quantities with Zero-Inflated Data

Hukum Chandra, H.V.L. Bathla and U.C. Sud

The paper examines small area estimation (SAE) for data with larger proportion of zeros (i.e. hereafter zero-inflated data) than would be expected under standard distributional assumptions. Presence of substantial proportion of zeros in the data makes model assumptions invalid. Consequently, commonly used methods of SAE based on a linear mixed model, for example, the empirical best linear unbiased predictor (EBLUP, Rao, 2003, Chapter 5), Pseudo-EBLUP (Prasad and Rao 1999) and model-assisted empirical best predictor of Jiang and Lahiri (2006), may not be efficient due to model misspecification. We propose SAE technique under the mixture model (Fletcher *et al.* 2005) that account for excess zeros in the data. Empirical results generated from limited simulation studies show that the proposed method works well and produces an efficient set of small area estimates.

Indian Agricultural Statistics Research Institute, New Delhi

30. Use of Two-phase Sampling for Crop Yield Forecasting

Ranjana Agrawal and Gaurav Goel

Models based on plant characters is one of the approaches for crop yield forecasting. This approach requires periodical data collection on plant characters from farmers' fields. In this paper, use of two-phase sampling has been suggested for reducing cost on data collection wherein data on less costlier characters are collected from all sampling units and data on characters involving high cost and labour are collected from a sub-sample. The approach has been demonstrated on sugarcane in Meerut district. The results indicated that forecasts based on suggested methodology were comparable to those based on data on all characters collected from all the units.

Indian Agricultural Statistics Research Institute, New Delhi

31. Methodological Issues in Marketing Research in Agriculture

S.P. Bhardwaj

Marketing research is the function that links the consumer, customer, and public to the market through available information. Market research is now on the agenda of all the trading economies whether they are large or small. Marketing research is a form of business research and is generally divided into two categories: consumer market research and business-to-business (B2B) market research. To conduct market research, it is crucial to define the research objectives clearly. Market researchers can utilize many types of research techniques and methodologies to capture the data that they require. All of the available methodologies based on quantitative or qualitative information. Agricultural marketing is witnessing major changes owing to liberalization and globalization of markets. In this context agriculture has to be market driven, more cost effective, competitive, innovative and responsive to high tech and I.T. applications. Market intelligence (M.I.) may be defined as gathering, analysis, interpretation and dissemination of trade information relevant to current and potential markets. Major building blocks of M.I. are Competitor Intelligence, Product Intelligence, Customer Understanding (Market Research) and Market Understanding (Analysis). Exploratory research is unstructured and qualitative in nature and information is collected by focus group interviews, reviewing literature or books, discussing with experts, etc. Conclusive research is numerically oriented, requires significant attention to the measurement of market phenomena and often involves statistical analysis to draw some conclusion about the problem. It is essentially, structured and quantitative research, and the output of this research is the input to management information systems (MIS). Exploratory research is also conducted to simplify the findings of the conclusive or descriptive research. Choice of marketing model was probably developed by economists and psychologists together.

Indian Agricultural Statistics Research Institute, New Delhi

32. A Comparative Study for Growth Pattern of Ginger Yield in India

Pal Singh, Amrit Kumar Paul and Savita Wadhwa

Nonlinear growth models play a vital role in many branches of agricultural and biological sciences. Three nonlinear growth models are presented in this study. The parameters are estimated using the Levenberg-Marquardt iterative method of nonlinear regression relating ginger yield growth data. Based on the performance of parameters and goodness of fit statistics, Monomolecular model is best fit than Logistic model and Gompertz model.

Indian Agricultural Statistics Research Institute, New Delhi

33. Average Linkage Method for Clustering Rice Producing States of India

Amrit Kumar Paul, Md. Wasi Alam and Pal Singh

In this article, fifteen years rice production level of thirty four states (including union territory) of India have been studied more precisely, based on homogeneous grouping pattern of rice producing states which were obtained by using average linkage method of hierarchical cluster analysis and squared euclidean distance. This method constructs four valid mutually exclusive clusters of rice producing states based on true distance of fifteen years rice production data. Years of rice production have also been grouped on the basis of similarity in terms of production level of state as well as national average (NAAV). Grouping of years based on rice production level can assist in discovering the meteorological or pathological or biological reasons for a group of particular production level of rice. Since, rice production ranging from 0.2 (Chandigarh) to around 12000 (WB) in 000 tonnes, hence in order to reduce the impact of extreme values or outliers in calculation of NAAV, 5% trimmed mean has been used for calculating the NAAV. Group wise trend of rice production of different states have been shown along with national average production of rice. Trend of rice production of some typical states have been shown along with NAAV. This study can play an important role for the research workers or planners whose aim is to augment the rice production level of the states like Kerala, AP etc. whose production has declined in recent past.

Indian Agricultural Statistics Research Institute, New Delhi

34. Frontier Production Function for Estimation of Technical Efficiency of Aquaculture: A Study of Punjab State of India

D.R. Singh and A.K. Vasisht

The level of technical efficiency of a particular firm is characterized by the relationship between observed production and some ideal production. The measurement of firm-specific technical efficiency is based upon deviations of observed output from the best production or efficient production frontier. A firm is defined as being technically efficient for a given technology, if it fully realizes its own technical efficiency potential by following the best practice techniques of the chosen technology and produces on its production frontier consistent with its socio-economic physical environment. Technical efficiency is defined and measured as the ratio of the firm's/farm's actual output to its maximum possible frontier output for a given level of inputs and the chosen technology. The study was undertaken on the primary data collected from the selected fish farmers of Punjab during 2007-08. The stochastic frontier production function of the Cobb-Douglas type was used to estimate the technical efficiency for the individual farms. An attempt has also been made to identify the determinants of technical efficiency. In addition to technical efficiency, the study also finds out the allocative efficiency, which is a measure of how efficiently and rationally farmers are allocating the resources in the production process. The cost of cultivation and benefit-cost analysis of aquaculture in Punjab State was also performed. The analysis revealed that the benefit-cost ratio was less than unity in all the category of farmers.

Further, the farms are technically sound but the economic efficiency is quite low. Various factors like the low price of output or the high cost of inputs etc., may influence this outcome. There is a tremendous scope to enhance the profitability of the fish farmers of Punjab State by improving their technical, allocative and economic efficiencies in aquaculture.

Indian Agricultural Statistics Research Institute, New Delhi

35. Multivariate Structural Time-Series Modelling and its Applications in Rice Forecasting

S. Ravichandran and P. Muthuraman

Autoregressive Integrated Moving Average (ARIMA) is widely used for modelling stationary time-series data sets. Structural time series (STM) modelling (Harvey 1996) is also utilized for modelling stationary or non-stationary time series data sets. In STM modelling system, characteristics of the data decide the particular type of model from STM family. Multivariate STM modelling can be utilized effectively when there are several independent and dependent variables. This can be used to describe simultaneously several features and properties of various time series. Here, cyclical components are constructed, which, although different for different series, having common parameters. Thus, the cycles in different series have similar properties; in particular, their movements are centered around the same period. This will be reasonable if cyclical movements arise as a result of a common cycle. As regards the trends themselves, constraints can be imposed so that there are fewer trends than series. The idea of imposing common trends is that when the series all have the same underlying source of growth. On the other hand, constraints arise from long-run co-integrating relationships. Both interpretations are possible, the co-integration interpretation is based on a triangular representation (Phillips 1991). Having established facts associated with a group of series, the next step is to construct models which aim to capture the dynamic interactions between them. In some circumstances, such models may have advantages over vector autoregressive (VAR) representations. Multivariate structural time series models are designed to handle this kind of situation. Since, trends are modelled explicitly, short-term dynamics can be captured by a low-order VAR. Co-integration appears when there are common trends (Harvey and Stock 1988). In agriculture, the dependent variable is a function of several independent variables and hence using Multivariate STM procedure, rice yield forecast for the country is made for the year 2008-09 by keeping all-India rice yield as independent variable and all the other yield influencing parameters such as area, production, and rainfall as independent parameters. All-India rice yield forecast for 2008-09 would be 2185 kg ha^{-1}.

Directorate of Rice Research, Hyderabad

36. Growth Model of Weed Variables Grown with Wheat

T. Rai, Ranjana Agrawal and Madan Mohan

The present investigation is made to study the growth behaviour of weed variables which affect the wheat yield. The experiments were conducted in the field of Agronomy Division of IARI, New Delhi for wheat during the rabi season for consecutively three years from 1999-2000, 2000-01 and 2001-02. During the period of crop growth, two types of weeds, i.e. broad leaves and grazy leaves were identified for recording of observations on weed counts, dry matter accumulations and leaf area at weekly intervals starting from thirty four days after sowing during rabi season till a fortnight before the harvest. The growth of weed appears to increase slowly during early stage. It increases at an increasing rate for a period of time and then approaches to an asymptotic maximum value. Dry matter of both types of weeds indicated logistic pattern. Leaf area of both types of weeds followed quadratic pattern. Hence, to examine the growth behaviour of weeds, two types of models i.e. logistic and quadratic were fitted. The weed counts of both type of weeds remained almost static.

Indian Agricultural Statistics Research Institute, New Delhi

37. Sequential Estimation of Genetic Parameters for Multiple Alleles at the Locus

K. Dutta[1] and A.S. Acharya[2]

The concept of sequential sampling for the two alleles at the locus is available in statistical literature. We generalize definitions and terminology of two dimensional sampling plans generated for the sequentially sample data from gene population for multidimensional sampling plans. Methods of estimations of genetic proportions available in statistics literature, is applicable only when Hardy-Weinberg law holds good and when there is no rare gene in the population. Here, we have introduced the procedure of estimation of genetic proportions for the multiple alleles at the locus which is applicable under both the situations. Moreover, estimation of parameters like proportions, ratios, percentages are widely used in agriculture, bio-medical science, population studies and tribal health etc. We have described here the general procedure named as sequential sampling method for estimation of different proportions, as fixed sample size method fails when there is rarest character in a population. It has been shown that the sequential sampling method, which has advantage of the use of adequate information with small samples provide estimators with small variance.

[1] *Sambalpur University, Sambalpur*
[2] *Regional Medical Research Centre (ICMR), Bhubaneswar*

38. The Influence of the Rainfall on the Yield of Wheat at Faizabad

K.K. Pandey, V.N. Rai and R.P. Kaushal

Fisher (1924) studied the influence of rainfall on the yield of wheat and showed that it is total amount which influence the crop yield. Murlitharan and Lathika (2005) discussed that the aspects of modeling of rainfall data using a modified version of Weibull distribution for ten years period 1961 to 1970. Normal, Lognormal, and Pearson's type distribution were studied for each year separately for rainfall. We found that Pearson's type IV in year 1994 and 1998 and Pearson's type VI in year 1990 performed better. For forecasting of wheat yield, three models Linear, Nonlinear (Cobb-Douglus), Square root with and without coefficient '*c*' are used and the square root without coefficient '*c*' model selected as the best.

N.D. University of Agriculture and Technology, Faizabad

39. Model Selection Criteria

S. Ismail[1], G. Mohan Naidu[1], D. Giri[2], C. Subbarami Reddy[3] and P. Balasiddamuni[3]

In recent years, Agricultural Statisticians have shown an increasing interest in the problems of model selection and model building. Model selection is an important part of any statistical analysis and many statisticians have examined and suggested different methods in the literature for selecting the 'best model'. The selection of a model uses some criteria in which the choice of the model is refined on the basis of some preliminary data search. Under the criteria of model selection, among a large number of potentially important regressors, only a subset of explanatory variables may be finally chosen.

For a long time, computer-based stepwise procedures based on testing hypothesis have been the dominant approach in the literature. Subset regression procedures are widely available, which can be used in model fitting when a subset of given regressors is assumed to be adequate for describing a dependent variable.

In the present study, a new model selection technique besides a modified stepwise regression method for selection of regressors, has been proposed by using different types of residuals.

[1] *S.V. Agricultural College, Tirupati*
[2] *S.V.A. Govt. Degree College, Srikalahasti*
[3] *S.V. University, Tirupati*

40. Specification Tests for Model Building

G. Mohan Naidu[1], S. Ismail[1], Y. Vijaya Sekhara Reddy[2], D. Giri[3], C. Subbarami Reddy[4] and P. Balasiddamuni[4]

Agricultural production function models are in general, either linear or non linear. The first and foremost step in the model building is the specification of the model. Over specification yields unbiased estimates of the parameters, but large variances; under specification yields biased estimators of the parameters and understates the variances of these estimates. In agricultural model building, a model can be misspecified in a number of ways. Two major sources are incorrect functional form and invalid assumption on the distribution of the error term in the model. One of the most important components of model building is tests for specification errors.

In this paper, an attempt has been made by proposing tests for model specification by using studentized and predicted residuals.

[1] *S.V. Agricultural College, Tirupati*
[2] *S.V.G.S. Degree College, Nellore*
[3] *S.V.A. Govt. Degree College, Srikalahasti*
[4] *S.V. University, Tirupati*

41. A Study of Growth Pattern and Technological Impact on Pulse Production in Uttar Pradesh

M.K. Sharma and B.V.S. Sisodia

The time series data on area, production and productivity of pulse crop for the period 19960-61 to 2005-06 have been considered for this study. The entire period has been divided in two period namely period I (1960-61 to 1989-90) and period II (1990-91 to 2005-06). The analysis of data revealed that the area and production of arhar, gram and moong have declined while those of lentil, pea and urd have increased substantially during the second period. The productivity of all the pulse crops except arhar and moong has increased during the second period. The annual growth rate of area of arhar, gram and pea has been found to be negative to the tune of 1.04, 2.90 and 2.40 per cent, respectively, since 1960-61 onwards. However, it was more pronounced during the second period for arhar and gram to the tune of about 2.70 and 3.24 per cent, respectively. The area under lentil, moong and urd has increased with the annual rate of 3.50, 6.0 and 3.00 per cent, respectively, during the last forty-five years. It was more prominent during the first period for lentil (5.74%) and moong (11.60%) while during the second period for urd (4.90%).

The production of arhar and gram has declined considerably at the annual rate of about 2.80% and 2.40% during the second period, respectively, though negative growth rate was also recorded during the first period but relatively less. The production of lentil and urd has increased with annual rate of about 5.20 and 3.53 per cent during 1960-61 to 2005-06, respectively. It was more distinguished during the second period for urd (3.78%), and during the first period for lentil (5.74). The production of moong has increased with the annual rate of about 13.69% during the first period but declined at the rate of about 5.00% during the second period. Unlike moong, pea production plunged by about 5% per annum during the first period but showed positive growth of 0.65% annually during the second period.

The productivity of arhar, gram, pea, lentil and moong grew relatively at high rate, i.e. 1.56, 1.40, 1.35, 1.73 and 1.88 per cent annually during the first period. However, moong productivity declined during the second period at the rate of about 1.60 per cent per annum. The productivity of urd has been found to decline at the rate of 0.60 to 0.90 per cent during both the periods, but on an overall it witnessed a positive growth of about 0.52 per cent per annum. The second period has witnessed low instability in the area, production and productivity of all the pulse crops except in case of production of arhar. The negative differential production during both periods has been observed for arhar, gram, pea and total pulses where decrease in area has affected negatively while increase in productivity has affected positively except in case of arhar and total pulses during the second period. However, positive differential production has been recorded for lentil and urd during both the periods where increase in area and productivity has affected positively except in case of urd which productivity has affected negatively during second period. That means the effect of change in area has mattered more on differential production of almost all the pulse crops. A significant structural difference in the production process of arhar and urd has been found between the two periods. The contribution of land to the production of pea, lentil, moong and urd has been found significant during both the periods. The technological impact on production of lentil and pea has, however, has been found significant during the period under study.

Narendra Deva University of Agriculture & Technology, Faizabad

42. Relative Importance of Plant Attributes in Discriminating Kabuli and Desi Varieties of Chickpea

Hemant Kumar, Shiv Sewak and Shiv Kumar

Twenty four varieties of chickpea were evaluated in a randomized complete block design with three replications at the main research farm of the Indian Institute of Pulses Research, Kanpur during rabi season 2005-06. Observations were recorded on days to 50% flowering, number of leaflets per plant, number of pods per plant, number of primary branches per plant, days to maturity, plant height, grain yield per plant and 100-seed weight to assess their relative importance in discriminating desi and kabuli varieties of chickpea. The results showed the Wilks' lamda value of 0.401 which indicating better discriminating power of the model. The probability value ($p = 0.036$) of the F-test of Wilks' lamda indicated that desi and kabuli varieties were significantly distinct. The classification matrix indicates that the discriminant function was able to classify about 92 % of the 24 varieties correctly. More specifically out of 19 desi varieties predicted to be in group 1 (desi), 17 fell rightly in group 1 and 2 found in group 2 (kabuli). Out of 5 kabuli varieties predicted to be in group 2, all 5 were classified in group 2. Thus, only 2 out of 24 varieties were not grouped correctly through the above model. Out of 8 independent variables studied, number of leaflets per plant was the best predictor followed by days to 50% flowering, plant height and 100-seed weight. In contrast grain yield per plant, number of primary branches per plant, days to maturity, and number of pods per plant were the poor discriminators.

Indian Institute of Pulses Research, Kanpur

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 65-67

Indian Society of Agricultural Statistics

Secretary's Report for the Year 2008

The Indian Society of Agricultural Statistics is a scientific body which was founded on January 03, 1947 with the main objective of promoting and undertaking research in Statistics and its application to Agriculture, Animal Husbandry, Fishery, Agricultural Economics, Computer Applications and allied fields. The Society was fortunate to have Late Dr. Rajendra Prasad, the then Union Minister of Agriculture, Government of India as its Founder President. He guided the Society for the first sixteen years of its inception even after becoming the President of the Republic of India. The Society had the privilege of receiving patronage and guidance from several eminent personalities from time to time as its Presidents in the past who took keen interest and were a source of great inspiration. In fact, the Society could attain its present status due to the untiring efforts of its Presidents in the past and the continued patronage and guidance from the current President Dr. Mangala Rai, Secretary, Department of Agricultural Research and Education, Ministry of Agriculture, Government of India and Director General, Indian Council of Agricultural Research, New Delhi. The farsightedness, overall guidance and unstinting support from the eminent statisticians and founder members, Late Prof. PV Sukhatme and Late Dr. VG Panse have been fundamental to the growth of the Society.

The Society organizes annually a conference in different parts of the country which provides a wide platform for exchange of ideas on various issues of national as well as regional importance through symposia besides holding invited lectures by eminent scientists, paper presentation and awards and incentives in various forms. Last year, the Society held its 61st Annual Conference at Birsa Agricultural University, Kanke, Ranchi from 30 November to 02 December 2007. This year, the Society is grateful to the Acharya N.G. Ranga Agricultural University, Hyderabad for inviting the Society to hold its 62nd Annual Conference at S.V. Agricultural College (ANGRAU), Tirupati from 24 to 26 November, 2008. Symposia on (i) **Accelerated Growth of Agriculture through Information Technology**, and (ii) **Agricultural Statistics for Planning** are being organized during the Conference.

The Society brings out a Journal every year consisting of three issues (April, August and December) in a volume. The Journal serves as a medium for disseminating research findings on agricultural statistics and allied fields. The Hindi Supplement continues to be a special feature of the Journal. The high standard of the Journal has been maintained due to the sincere efforts of the Editorial Board and the Referees. The Society is thankful to them for their keen interest in its activities related to the publication of the Journal. With a view to promoting research in Statistics and improving the standard of its Journal, the Society has been awarding prizes for the best papers published in the Journal for every biennium from 1987 in the fields of Design of Experiments, Sampling Theory, Statistical Genetics, Statistical Methodology, Applied Statistics and Computer Applications. A special issue of the Journal Volume 62, No. 2, August 2008 was brought out in the memory of Dr. K. Kishen in 2008. This issue contains 10 invited articles from India and abroad.

The membership of the Society which is drawn from all parts of India as well as from abroad during the year was

Permanent Institutional Members	29
Life Members	606
Annual Members	04
	639

In addition to its regular members, the Society has a number of institutional subscribers to its Journal in India and abroad. The number of Indian subscribers during the year was 185. During the year under report, 3 new permanent institutional members (Govt. Vidarbha Institute of Social & Humanities, Amravati; Dr. Gaur Hari Singhania Institute of Management & Research, Kanpur; Brahmanand Mahavidyalaya, Rath, Hamirpur) and 24 new life members were enrolled. Thus, the total number of members and subscribers during the year was 824.

In order to perpetuate the memory of its Founder President, Late Dr. Rajendra Prasad, a memorial lecture is being organized during the Conference since 1965. The Society has organized 44 lectures so far and the memorial lecture is being organized during this Conference would be 45th in the series. Also since 1973, the Society has been organizing a lecture in the memory of Late Dr. VG Panse, who had been the guiding spirit behind the Society and its activities. The Society has so far organized 29 lectures in his memory and the current memorial lecture would be 30th in the series.

The Society has a Research Unit to undertake research on specific problems of current interest under the guidance of a Research Direction Committee.

This year, a study relating to the estimation of level of development of different districts of Andhra Pradesh was conducted. The level of development was obtained with the help of composite index based on optimum combination of fifty socio-economic indicators. The district-wise data for the year 2001-02 in respect of these indicators were utilized for 22 districts of the State. The level of development was estimated separately for agricultural sector, infrastructural facilities and socio-economic sector. The district of West Godavari was ranked first in overall socio-economic development and the district of Guntur was found on the first position in respect of agricultural development. Wide disparities were observed in the level of development among different districts. Infrastructural facilities were found to be positively associated with the levels of development in agricultural sector and overall socio-economic field. Agricultural development was also influencing the overall socio-economic development in the positive direction.

The problem of finance for scientific activities, printing of the Journal and other ad-hoc publications could be overcome to a certain extent through grant-in-aid received from the Ministry of Agriculture, Government of India and Indian Council of Agricultural Research. The Society wishes to acknowledge very gratefully the financial assistance received from them during the year under report.

The Society continues to be a member of the International Statistical Institute, Netherlands, Indian Association of Social Science Institutions, New Delhi and Federation of Indian Societies of Agricultural Sciences and Technology, New Delhi.

The 61st Annual Conference of the Indian Society of Agricultural Statistics (ISAS) was held at Department of Agricultural Statistics, Birsa Agricultural University, Kanke, Ranchi – 834 006 (Jharkhand) from 30 November to 02 December, 2007. The Conference was inaugurated by Dr. Pronab Sen, Chief Statistician of India and Secretary, Ministry of Statistics and Programme Implementation, Government of India on 30 November 2007. Dr. Mangala Rai, Secretary, DARE, Government of India and Director General, ICAR, New Delhi and President of the Society presided over the Inaugural Function. Dr. A.K. Sarkar, Dean, Faculty of Agriculture, Birsa Agricultural University, Ranchi delivered the Welcome Address. Shri A.K. Sarkar, IAS and Principal Secretary and Commissioner (Agriculture), Government of Jharkhand and Dr. N.N. Singh, Vice Chancellor, Birsa Agricultural University, Ranchi addressed the participants. Review of the activities of the Society was presented by Dr. S.D. Sharma, Secretary, Indian Society of Agricultural Statistics. Dr. Mangala Rai, delivered the Presidential Address. Dr. K. Sinha, Professor and Chairman, Department of Agricultural Statistics, Birsa Agricultural University, Ranchi and Organizing Secretary of the Conference proposed a Vote of Thanks. Dr. V.K. Gupta, ICAR National Professor, IASRI, New Delhi was the Sessional President of the Conference and he delivered the Technical Address on **Orthogonal Arrays and their Applications**. Dr. Rajendra Prasad Memorial Lecture was delivered by Dr. Pronab Sen, Chief Statistician of India and Secretary, Ministry of Statistics and Programme Implementation, Government of India, New Delhi. The topic of his lecture was **Reflections on Planning in India**. Dr. V.G. Panse Memorial Lecture was delivered by Dr. A.K. Srivastava, Former Joint Director, Indian Agricultural Statistics Research Institute, New Delhi on **Small Area Estimation – A Perspective and Some Applications**. Two symposia on (i) Role of Agriculture in Containing Acute Rural Distress – Some Statistical Issues, and (ii) Statistical Aspects of Research in Agricultural and Environmental Sciences – Status and Scope were organized during the Conference. The methodologies and the findings of the research study carried out by the Research Unit of the Society were presented by Shri S.C. Rai. The topic of the study was Statistical Evaluation of Socio-economic Development of Different States in India. A total of 43 research papers were presented by different delegates during the Conference in the fields of Design of Experiments, Sample Surveys, Statistical Genetics, Applied Statistics, Statistical Methodology and Computer Applications.

Plenary Session was chaired by Dr. S.D. Sharma, Director, IASRI and Secretary, ISAS, New Delhi where the recommendations of both the symposia were presented by Dr. H.V.L. Bathla and Dr. Prajneshu and these were approved.

The Society conveyed its grateful thanks to the authorities of the Birsa Agricultural University, Ranchi for organizing the 61st Annual Conference.

The accounts of the Society for the year ending 31 March, 2008 were audited by M/s V. Garg & Co., Chartered Accountants, Professional Auditors and will be presented in the General Body Meeting.

The work of the Society during the year was made possible through the advice and help of the members of the Executive Council, Editorial Board and the Research Direction Committee. The burden of the entire Secretariat of the Society has been willingly borne by my colleagues, Dr. V.K. Bhatia, Shri R.S. Khatri, Dr. Rajender Parsad, Dr. A.K. Vishandass, Shri S.C. Rai and Dr. A.K. Srivastava. In the end, I wish to thank the staff of the Society for their devoted work.

S.D. SHARMA
Secretary

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 69-75

Efficient Estimation in Poststratification under Optimal and Non-optimal Conditions

M.C. Agrawal and S.C. Senapati[1]
University of Delhi, Delhi

(Received: February 1999, Revised: November 2008, Accepted: March 2009)

SUMMARY

Employing the customary predictive format, as alluded to by Basu (1971), Smith (1976) and several others, for estimation of the population total or the population mean under a fixed population set-up, we have generated a sequence of efficient unbiased poststratification-based estimators. The proposed sequence of estimators is found, under optimal and non-optimal conditions, to be more efficient than the customary poststratified estimator and the usual simple mean. The performance of the proposed sequence of estimators has been examined from the point of view of conditional randomization inference.

Key words: Poststratification, Conditional randomization inference, Non-optimal better estimators.

1. INTRODUCTION

To ascertain whether or not the traditional estimators conform to a certain intuitive in-built feature, Basu (1971), Smith (1976) and several others have pleaded for a plausible predictive format under a fixed-population set-up. Agrawal and Sthapit (1997) have tapped this format to arrive at a sequence of efficient ratio-based and product-based estimators. Agrawal and Panda (1993) have suggested a suitably weighted combination of the customary poststratified estimator ($\bar{y}_{ps}$, say) and simple mean ($\bar{y}$, say) which performs better than either of the estimators $\bar{y}_{ps}$ and $\bar{y}$. In this paper, we have invoked a plausible predictive format under poststratified sampling and have used $\bar{y}_{ps}$ as an input predictor for the non-surveyed part of the population in a repetitive manner, thus obtaining a sequence of poststratification-based estimators.

Agrawal and Panda (1995) proposed a poststratified estimator through use of optimum weights. Here, in this paper, we consider, apart from the optimum situation, a decomposition of optimum weights into non-optimal sub-weights with a view to retaining the superiority of the suggested poststratified estimator over the customary poststratifeid estimator and simple mean.

2. POSTSTRATIFICATION-BASED ESTIMATION AND THE RELATED PERFORMANCE

Consider a population of size N stratified into k strata, the size of the i^{th} stratum being N_{i} such that $\sum_{i=1}^{k} N_i = N$. A simple random sample of size n is drawn from the population and the sample units are then assigned to the k strata. Suppose that n_i (i = 1, 2, · · · , k) is the number of units that fall into the i^{th} stratum such that $\sum_{i=1}^{k} n_i = n$, n_i varying from sample to sample. Assuming the probability of n_i being zero to be small, the usual unbiased estimator of the population mean $\bar{Y}$ in poststratified sampling is given by

$$\bar{y}_{ps} = \sum_{i=1}^{k} W_i \bar{y}_i$$

[1] *Ravenshaw University, Cuttack*

where *ps* stands for poststratification, $W_i = N_i/N$ and $\bar{y}_i$ is the mean of the n_i sample units that fall into stratum $i\,(i = 1, 2, \cdots, k)$. With a view to arriving at a predictive format under the fixed population set-up, we express the population total *Y* as

$$Y = \sum_{i=1}^{k} \sum_{j\in s_i} y_{ij} + \sum_{i=1}^{k} \sum_{j\in \bar{s}_i} y_{ij} = \sum_{i=1}^{k} n_i \bar{y}_i + \sum_{l\in \bar{s}} y_l \qquad (2.1)$$

where s_i denotes the sample of size n_i selected from the i^{th} stratum and $\bar{s}_i$ is its complement, and $\bar{s}$ is the complement of overall sample $s = \overset{k}{\underset{i=1}{\square}} s_i$. It is clear from (2.1) that, to estimate the total *Y*, we have to predict $y_l\,(l \in \bar{s})$ because the first component on the right hand side of (2.1) is known. This is tantamount to stating

$$\hat{Y} = \sum_{i=1}^{k} n_i \bar{y}_i + \sum_{l\in \bar{s}} \hat{y}_l \qquad (2.2)$$

where $\hat{y}_l$ is the implied predictor of $y_l\,(l \in \bar{s})$. Invoking the customary post-stratified estimator $\bar{y}_{ps}$ as an intuitive predictor of y_l in (2.2), we obtain

$$\hat{Y} = \sum_{i=1}^{k} n_i \bar{y}_i + (N - n)\bar{y}_{ps}$$

or say

$$\hat{Y} = \sum_{i=1}^{k} \frac{n_i \bar{y}_i}{N} + \frac{N-n}{N}\bar{y}_{ps} = y_{ps}^{(1)}$$

In the next step, we utilize $\bar{y}_{ps}^{(1)}$ as an intuitive predictor of y_l in (2.2) and this leads to $\bar{y}_{ps}^{(2)}$ given by

$$\bar{y}_{ps}^{(2)} = (1-\lambda^2)y + \lambda^2 \bar{y}_{ps}$$

where $\lambda = 1 - \frac{n}{N}$. Repetition of this process *r* times will culminate in

$$\bar{y}_{ps}^{(r)} = (1-\lambda^r)\bar{y} + \lambda^r \bar{y}_{ps} \qquad (2.3)$$

Having obtained $\bar{y}_{ps}^{(r)}$ given by (2.3), we may, in fact, extend the scope of r to cover negative integer values or even all real values without causing any problem. However, we hereafter consider $r \geq 0$. Such an estimator $\bar{y}_{ps}^{(r)}$, which is unbiased for population mean $\bar{Y}$, will be called poststratification-based estimator of order *r*. It may be noted that, for $r = 0$, $\bar{y}_{ps}^{(r)}\,(i.e,\ \bar{y}_{ps}^{(0)}) = \bar{y}_{ps}$ and, as $r \to \infty, y_{ps}^{(r)} \to \bar{y}$. Now, noting that

$$\text{V}(\bar{y}) \;=\; \frac{\lambda}{n}S^2$$

$$\text{V}(\bar{y}_{ps}) \;=\; \frac{\lambda}{n}\sum W_i S_i^2 + \frac{N(N-1)\lambda}{n^2}\sum_{i=1}^{k}(1-W_i)S_i^2$$

and

$$\text{Cov}(\bar{y}, \bar{y}_{ps}) \;=\; \text{Cov}\left(\sum_{i=1}^{k}\frac{n_i \bar{y}_i}{n}, \sum_{i=1}^{k} W_i \bar{y}_i\right)$$

$$= \frac{\lambda}{n}\sum_{i=1}^{k} W_i S_i^2$$

the variance of $\bar{y}_{ps}^{(r)}$ can be expressed as

$$\text{V}(\bar{y}_{ps}^{(r)}) \;\square\; \frac{\lambda}{n}S^2 + \frac{\lambda^{r+1}}{n}Q(\lambda^r - 2)$$

$$+\frac{\lambda^{2r+1}}{n^2}\left[\frac{N}{N-1}\sum_{i=1}^{k}(1-W_i)S_i^2\right] \qquad (2.4)$$

where $Q = S^2 - \sum_{i=1}^{k} W_i S_i^2$ and S^2 and S_i^2 are, respectively, the population mean square and the mean square for the i^{th} stratum. The optimum value of *r* which minimizes the variance expression given in (2.4) is obtainable from

$$\lambda^{r*} \;=\; \frac{Q}{R} \qquad (2.5)$$

where *r** denotes the optimum value of *r*, $R = Q + Q_1$ and

$$Q_1 = \frac{1}{f(N-1)}\sum_{i=1}^{k}(1-W_i)\,S_i^2$$

Barring some exceptional sampling situations, the value of Q will be positive. Henceforth, we will assume $Q > 0$. It can be easily verified that the use of (2.5) will reduce (2.4) to

$$\mathrm{V}_{\min}(\bar{y}_{ps}^{(r)}) = \mathrm{V}(\bar{y}_{ps}^{(r*)}) = \frac{\lambda}{n}S^2 - \frac{\lambda Q^2}{nR}$$

which will be smaller that $\mathrm{V}(\bar{y}_{ps})$ or $\mathrm{V}(\bar{y})$. Rewriting the variance of $\bar{y}_{ps}$ as

$$\mathrm{V}(\bar{y}_{ps}) = \frac{\lambda}{n}\sum_{i=1}^{k}W_iS_i^2 + \frac{\lambda}{n}Q_1 = \frac{\lambda}{n}S^2 + \frac{\lambda}{n}(Q_1 - Q)$$

we note that poststratification will be resorted to only when $\frac{Q}{Q_1} > 1$, for otherwise, the simple mean will score over the poststratified estimator $\bar{y}_{ps}$.

Since determination of r^* via (2.5) will not be easy in view of involvement of the population quantities, we discuss non-optimal efficient solution in order to ensure superior performance of $\bar{y}_{ps}^{(r)}$ for values of r other than r*. The following inequality, obtained from the comparison of relevant variances of $\bar{y}_{ps}^{(r)}$, $\bar{y}_{ps}$ and $\bar{y}$ given above, will render $\bar{y}_{ps}^{(r)}$ more efficient compared to either of $\bar{y}_{ps}$ and $\bar{y}$

$$\frac{\tau-1}{\tau+1} < \lambda^r < \frac{2\tau}{\tau+1} \quad \text{where } \tau = \frac{Q}{Q_1} \qquad (2.6)$$

An idea about τ can be had from a pilot or past survey, thus enabling us to decide on τ.

For different values of f and τ, we have used (2.6) to prepare Table 1 which displays bounds on r for which $\bar{y}_{ps}^{(r)}$ performs better than $\bar{y}$ and $\bar{y}_{ps}$.

Table 1. Range of r for different τ and f values

τ\\f	.01	.05	.10	.20
0.1	$r > 169.62$	$r > 33.23$	$r > 16.18$	$r > 7.64$
0.5	$r > 40.34$	$r > 7.9$	$r > 3.84$	$r > 1.82$
0.9	$r > 5.38$	$r > 1.05$	$r > .51$	$r > .24$
1	$r > 0$	$r > 0$	$r > 0$	$r > 0$
2	$0 < r < 109.31$	$0 < r < 21.42$	$0 < r < 10.43$	$0 < r < 4.92$
5	$0 < r < 40.34$	$0 < r < 7.90$	$0 < r < 3.85$	$0 < r < 1.82$
10	$0 < r < 19.96$	$0 < r < 3.91$	$0 < r < 1.90$	$0 < r < .90$

To appreciate mathematically the significance of τ which is a pivotal quantity, we suppose $S_i^2 = S_W^2$ which implies that proportional allocation is optimal in the Neyman sense. Then, for large N_i (i = 1, 2, · · · , k)

$$Q = S^2 - \sum_{i=1}^{k}W_iS_i^2 = \sum_{i=1}^{k}W_i(\bar{Y}_i - \bar{Y})^2$$
$$= S_b^2 \cdot \frac{k-1}{N}$$

where $S_b^2 = \frac{1}{k-1}\sum_{i}^{k}N_i(\bar{Y}_i - \bar{Y})^2$

and $Q_1 = \frac{1}{f(N-1)}\sum_{i=1}^{k}(1-W_i)\,S_i^2 = \frac{(k-1)\,S_W^2}{f(N-1)}$

and thus τ = f. F where $F = \frac{S_b^2}{S_W^2}$ which is a ratio of 'between' mean squares to 'within' mean squares and is obtained from ANOVA table.

3. PERFORMANCE-SENSITIVITY OF THE PROPOSED ESTIMATOR DUE TO NON-OPTIMALLY OF *r*

We would like to determine the loss in efficiency of $\bar{y}_{ps}^{(r)}$ arising from the use of values of r other than optimum r (i.e. r*) value. To evaluate this loss, we define a quantity P_{I} which is the proportional inflation in variance of $\bar{y}_{ps}^{(r)}$ resulting from lack of knowledge of r^* as

$$P_{\mathrm{I}} = \frac{\mathrm{V}(\bar{y}_{ps}^{(r)}) - \mathrm{V}(\bar{y}_{ps}^{(r^*)})}{\mathrm{V}(\bar{y}_{ps}^{(r^*)})} \qquad (3.1)$$

After some algebra, P_{I} can be expressed as

$$P_{\text{I}} = \left(\frac{\lambda^{r} - \lambda^{r*}}{1-\lambda^{r*}}\right)^{2} G$$

where $G = \dfrac{\text{V}(\bar{y}_{ps}^{(r)}) - \text{V}(\bar{y}_{ps}^{(r*)})}{\text{V}(\bar{y}_{ps}^{(r)})}$, indicating the gain in efficiency of $\bar{y}_{ps}^{(r)}$ (using r^{*}) relative to $\bar{y}_{ps}$. The estimator $\bar{y}_{ps}^{(r)}$ will continue to fare better than $\bar{y}_{ps}$ provided

$$P_{\text{I}} < G \Rightarrow \left|\frac{\lambda^{r} - \lambda^{r^{*}}}{1-\lambda^{r*}}\right| < 1 \Rightarrow 2\lambda^{r^{*}} - 1 < \lambda^{r} < 1 \quad (3.2)$$

implying thereby that $\bar{y}_{ps}^{(r)}$ will always (irrespective of choice of r) be more efficient than $\bar{y}_{ps}$ if $\lambda^{r*} < \frac{1}{2} \; (\Rightarrow \tau < 1)$. But, for $\lambda^{r^{*}} > \frac{1}{2} \; (\Rightarrow \tau > 1)$ we can manipulate (3.2) to obtain

$$r < \frac{\log(2\lambda^{r*} - 1)}{\log \lambda} \square \frac{2Q_1}{f\,Q} = \frac{2}{f\tau} \quad (3.3)$$

Now, turning to the case involving $\bar{y}$, we can express $P_{\text{I}} = \delta^{2} \cdot G'$ where $G' = \dfrac{\text{V}(\bar{y}) - \text{V}(y_{ps}^{(r^{*})})}{\text{V}(\bar{y}_{ps}^{(r*)})}$ and $\lambda^{r} = (1+\delta)\lambda^{r^{*}}$, G' indicating the gain in efficiency of $\bar{y}_{ps}^{(r)}$ (using r^{*}) relative to $\bar{y}$. The estimator $\bar{y}_{ps}^{r}$ (for a non-optimal r) will be more efficient than $\bar{y}$ provided

$$P_{\text{I}} < G' \Rightarrow |\delta| < 1 \Rightarrow \lambda^{\text{r}} < 2\lambda^{r^{*}} \Rightarrow \tau > \frac{\lambda^{r}}{2-\lambda^{r}} \quad (3.4)$$

Alternatively, $\bar{y}_{ps}^{(r)}$ will fare better than $\bar{y}$ if

$$r > \frac{1}{f\tau} - \frac{\ln 2}{f} \quad (3.5)$$

where τ > 1. Combining (3.3) and (3.5), we conclude that, for $\bar{y}_{ps}^{(r)}$ to perform better than $\bar{y}$ and $\bar{y}_{ps}$, we have (for τ > 1)

$$\frac{1}{f\tau} - \frac{\ln 2}{f} < r < \frac{2}{f\tau} \quad (3.6)$$

Note also that, if $\dfrac{\lambda^{r}}{2-\lambda^{r}} < \tau < 1$ (for and r), $\bar{y}_{ps}^{(r)}$ will be superior to $\bar{y}$ and $\bar{y}_{ps}$. It can be verified from (3.4) that, if τ < 1, the values of r that render $\bar{y}_{ps}^{(r)}$ more efficient than $\bar{y}$ are given by

$$r > \frac{\tau - \ln(2\tau)}{f} \quad (3.7)$$

Alternatively, taking $r = (1+\delta')r*$ where δ' is the proportional deviation in r^{*}, we can express

$$P_{\text{I}} = \left\{\left(\frac{Q}{R}\right)^{\delta'} - 1\right\}^{2} \quad G' = \frac{\delta'^{2}}{r^{2}} G' \text{ if } \tau > 1$$

Numerical Illustration

Example: The following data have been taken from Sarndal *et al.* (1992, p.119)

Stratum i	N_i	$\sum_{j=1}^{N_i} y_{ij}$	$\sum_{j=1}^{N_i} y_{ij}^{2}$
1	105	1098.9	21855.05
2	19	3445.9	1822736.83

(i) For n = 30, f = 0.242, λ = 0.758, we have

$$\lambda^{r*} = 0.64 \Rightarrow r^{*} = 1.6 \text{ and } \tau = \frac{\lambda^{r*}}{1-\lambda^{r*}} = 1.775.$$

Since τ > 1, it is clear from (3.6) that, for $\bar{y}_{ps}^{(r)}$ to be more efficient than $\bar{y}$ and $\bar{y}_{ps}$, we should have 0 < r < 4.66.

(ii) For n = 15, f = 0.121, λ = 0.897, we have $\lambda^{r^{*}}$ = .47, r^{*} = 0.5.851 and τ = 0.887.

As τ < 1, we conclude from (3.2) that $\bar{y}_{ps}^{(r)}$ (whatever be r) will always be better than $\bar{y}_{ps}$.

However, for $\bar{y}_{ps}^{(r)}$ to perform better than $\bar{y}$, we get $r > 2.61$ from (3.7). Thus a choice of $r > 2.61$ will ensure superiority of $\bar{y}_{ps}^{(r)}$ vis-a-vis $\bar{y}_{ps}$ and $\bar{y}$.

For the above example with $n = 30$, Table 2 presents appraisal of the impact of departure from r^* in terms of loss in the efficiency of $\bar{y}_{ps}^{(r)}$ as a result of employing some non-optimum r instead of r^*.

Table 2. Loss in efficiency of $\bar{y}_{ps}^{(r)}$ as a result of departure from r^*

δ'	P_{I}
0.05	0.000091
0.10	0.000357
0.15	0.000786
0.20	0.001367
0.25	0.002090
0.30	0.002944
0.35	0.003921
0.40	0.005012
0.45	0.006208
0.50	0.007500

Table 2 clearly reflects that proportional deviations to the extent of 50% from r^* cause only .75% proportional inflation on $V(\bar{y}_{ps}^{(r)})$ relative to $V(\bar{y}_{ps}^{(r^*)})$. In other words, there is insignificant or no loss in efficiency of $\bar{y}_{ps}^{(r)}$ when we conceive departures from r^*, at least to the extent envisaged in the above table.

4. CONDITIONAL RANDOMIZATION INFERENCE

Following the work of authors such as Holt and Smith (1979), Smith (1991), Valliant (1993), Agrawal and Panda (1995) with regard to the use of conditional inference in poststratification, we would now like to examine the performance of $\bar{y}_{ps}^{(r_c)}$ in the conditional case by conditioning the mean square error on actual sample size from different strata and the same is then expressible as

$$\mathrm{MSE}(\bar{y}_{ps}^{(rc)} \mid n) = \mathrm{V}(\bar{y}_{ps}^{(r_c)} \mid n) + \{Bias(\bar{y}_{ps}^{(r_c)} \mid n)\}^2$$

$$= \sum_{i=1}^{k} \psi_i \left(\frac{n_i}{n} - \lambda^{rc} k_i \right)^2 \mid (1 - \lambda^{rc})^2 \left(\sum_{i=1}^{k} K_i \bar{Y}_i \right)^2 \qquad (4.1)$$

where $W_i = \dfrac{N_i}{N}, K_i = \dfrac{n_i}{n} - \dfrac{N_i}{N}$ and $\psi_i = \left(\dfrac{1}{n_i} - \dfrac{1}{N_i} \right) S_i^2$

The subscript c' in the above discussion indicates 'conditional' case. Using the optimal value of r_c, say, r_c^* we can find

$$\mathrm{MSE}\left(y_{ps}^{(r_c^*)} \mid n \right) = \sum_{i=1}^{k} \psi \left(\frac{n_i}{n} \right)^2 + \left(\sum_{i=1^k} K_i \bar{Y}_i \right)^2$$

$$- \frac{\left[\sum_{i=1}^{k} K_i \psi_i \frac{n_i}{n} + \left(\sum_{i=1}^{k} K_i \bar{Y}_i \right)^2 \right]^2}{\sum_{i=1}^{k} K_i^2 \psi_i + \left(\sum_{i=1}^{k} K_i \bar{Y}_I \right)^2} \qquad (4.3)$$

which will be smaller that $\mathrm{V}(\bar{y}_{ps} \mid n)$ or $\mathrm{MSE}(\bar{y} | n)$ given by

$$\mathrm{V}(\bar{y}_{ps} \mid n) = \sum_{i=1}^{k} W_i^2 \psi_i \qquad (4.4)$$

$$\mathrm{MSE}(\bar{y} \mid n) = \sum_{i=1}^{k} \left(\frac{n_i}{n} \right)^2 \psi_i + \left(\sum_{i=1}^{k} K_i \bar{Y} i \right)^2 \qquad (4.5)$$

5. PERFORMANCE-SENSITIVITY UNDER CONDITIONAL RANDOMIZATION INFERENCE

It is of interest to know that potential loss in efficiency of the proposed post stratified estimator $\bar{y}_{ps}^{(r_c)}$ if we use some λ^{r_c} other than $\lambda_c^{r^*}$. For this purpose, we define, under conditional randomization inference, a measure P_{IC} similar to P_{I}, i.e.,

$$P_{\mathrm{IC}} = \frac{\mathrm{MSE}(y_{ps}^{(r_c)} \mid n) - \mathrm{MSE}(\bar{y}_{ps}^{(r^*c)} | n)}{\mathrm{MSE}(y_{ps}^{(r^*c)} | n)}$$

We can, after simplification, express

$$P_{IC} = \left(\frac{\lambda^{r_c} - \lambda^{r_c^*}}{1 - \lambda_c^{r*}}\right) G_c'$$

where $G_c' = \dfrac{\mathrm{V}(\bar{y}_{ps} \mid n) - \mathrm{MSE}(y_{ps}^{(r_c^*)} \mid n)}{MSE(y_{ps}^{(r_c^*)} \mid n)}$

In the case of conditional randomization, $\bar{y}_{ps}^{(r_c)}$ will continue to fare better than y_{ps} provided

$$P_{IC} < G_c' \Rightarrow \left|\frac{\lambda^{r_c} - \lambda^{r_c^*}}{1 - \lambda^{r_c^*}}\right| < 1 \Rightarrow r_c < \frac{\log(2\lambda^{r_c^*} - 1)}{\log \lambda} \qquad (5.1)$$

Proceeding exactly in the same way as in Section 3, we can show that, in the conditional case, $\bar{y}_{ps}^{(r_c)}$ will perform better than $\bar{y}$ if

$$r_c > \frac{\log(2\lambda^{r_c^*})}{\log \lambda} \qquad (5.2)$$

Combining (5.1) and (5.2), we conclude that, in the conditional case, $\bar{y}_{ps}^{(r_c)}$ fares better than $\bar{y}$ and $\bar{y}_{ps}$ if

$$\frac{\log(2\lambda^{r_c^*})}{\log \lambda} < r_c < \frac{\log(2\lambda^{r_c^*} - 1)}{\log \lambda} \qquad (5.3)$$

The bounds on r_c given by (5.3) may be termed as 'efficiency bounds'. Taking $r_c = r_c^*(1 + \delta_c{}')$ where $\delta_c{}'$ denotes proportional deviation in r_c^*, we can express

$$P_{IC} = \left\{(\lambda^{r_c^*})^{\delta_c'} - 1\right\}^2 \cdot \frac{MSE(\bar{y} \mid n) - MSE(\bar{y}_{ps}^{(r_c^*)} \mid n)}{MSE(\bar{y}_{ps}^{(r_c^*)} \mid n)}$$

To illustrate the above results relating to $\bar{y}_{ps}^{(r_c)}$ under the framework of conditional randomization inference, we consider the following theoretical numerical example due to Holt and Smith (1979) which shows that there exists a sequence of non-optimal efficient estimators (based on use of some r_c other than r_c^*) ensuring superior performance of $\bar{y}_{ps}^{(r_c)}$ compared to simple mean and traditional poststratified estimator.

Example: A population which is postsratified into two strata has the following characteristics.

$$\overline{Y} = 0,\ S^2 = 2,\ N_1/N = N_2/N = \frac{1}{2}$$

$$S_1^2 = S_2^2 = 1,\ \ \overline{Y}_1 = -1, \overline{Y}_2 = 1 \text{ and } n = 20$$

However, instead of ignoring finite population correction factors as assumed by Holt and Smith (1979), we retain them by considering $N_1 = N_2 = 100$ and $S_1^2 = S_2^2 = 2$ in respect of the two strata.

From the standpoint of conditional randomization inference, we need, because of reasons of symmetry, to discuss the configurations from $n_1 = 1$, $n_2 = 19$ to $n_1 = 9$, $n_2 = 11$. We have excluded the case $n_1 = n_2 = 10$ as λ_c^{r*} is not defined in view of K_i becoming zero when $\dfrac{n_i}{n} = \dfrac{N_i}{N}$ $(i = 1, 2)$.

To appraise the performance of $\bar{y}_{ps}^{(r)}$ under conditional randomization inference, we have prepared Table 3 which reflects the performance of $\bar{y}_{ps}^{(r_c)}$ vis-a-vis $\bar{y}_{ps}$ and $\bar{y}$ when r_c^* is employed. More importantly, we display possible values of r_c (efficiency bounds of r_c) for which $\bar{y}_{ps}^{(r_c)}$ performs better that $\bar{y}_{ps}$ and $\bar{y}$.

Table 3. A Comparison of $\bar{y}_{ps}^{(r_c)}, \bar{y}_{ps}$ and $\bar{y}$ under the given configurations of sample sizes

n_1	r_c^*	MSE $(y_{ps}^{r_c^*} \mid n)$	Efficiency bounds of r_c	V $(\bar{y}_{ps} \mid n)$	MSE $(\bar{y} \mid n)$
1	4.05	0.3683	$0 < r < 11.24$	0.5163	0.8919
2	2.35	0.2288	$0 < r < 5.47$	0.2678	0.7236
3	1.71	0.1702	$0 < r < 3.81$	0.1861	0.5751
4	1.39	0.1386	$0 < r < 3.02$	0.1462	0.4464
5	1.20	0.1194	$0 < r < 2.57$	0.1233	0.3375
6	1.08	0.1070	$0 < r < 2.29$	0.1090	0.2484
7	1.01	0.0985	$0 < r < 2.14$	0.0999	0.1785
8	0.95	0.0938	$0< r < 2.00$	0.0942	0.1296
9	0.92	0.0909	$0 < r < 1.94$	0.0910	0.0999

Table 4. Loss of efficiency of $\bar{y}_{ps}^{(r_c)}$, as a result of departures from r_c^*

n_1	r_c	P_{IC}
1	5.0580	0.014500
2	2.9340	0.007800
3	2.1440	0.004600
4	1.7386	0.002900
5	1.4988	0.001770
6	1.3470	0.001000
7	1.2630	0.000560
8	1.1869	0.000230
9	1.1525	0.000057

Table 4 underscores the fact that, for above example, the values of r_c that embody deviations of 25% from r^*_c cause very little inflation in minimum conditional mean square error of $y_{ps}^{(r_c)}$ as indicated by column P_{IC}. In other words, there is insignificant loss in efficiency of $\bar{y}_{ps}^{(r_c)}$ when we conceive departures from r_c^*.

ACKNOWLEDGEMENT

The author thanks the referee for his suggestions leading to improvement in presentation of the paper.

REFERENCES

Agrawal, M.C. and Panda, K.B. (1993). An efficient estimator in poststratification. *Metron*, **LI (3-4),** 179-188.

Agrawal, M.C. and Panda, K.B. (1995). On efficient estimation in poststratification. *Metron*, **LIII(3-4)**, 107-115.

Agrawal, M.C. and Sthapit, A.B. (1997). Hierarchic predictive ratio-based and product-based estimators and their efficiency. *J. Appl. Statist.*, **24(1)**, 97-104.

Basu, D. (1971). An essay on the logical foundations of statistical inference, Part 1. In : *Foundations of Statistical Inference* (ed.) V.P. Godambe and D.A. Sprott, 203-233.

Holt, D. and Smith, T.M.F. (1979). Poststratification, *J. Roy. Statist. Soc.*, **A142**, 33- 46.

Smith, T.M.F. (1976). The foundations of survey sampling- A review. *J. Roy. Statist. Soc.,* **A139**, 183-204.

Smith, T.M.F. (1991). Poststratification. *The Statistician*, **40**, 315-323.

Valliant, R. (1993). Poststratification and conditional variance estimation. *J. Amer. Statist. Assoc.*, **88**, 89-96.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 77-83

Estimation of Parameters of Morgenstern Type Bivariate Logistic Distribution by Ranked Set Sampling

Manoj Chacko and P. Yageen Thomas
University of Kerala, Trivandrum 695 581

(Received: February 2006, Revised: December 2008, Accepted: January 2009)

SUMMARY

Ranked set sampling is applicable whenever ranking of a set of sampling units can be done easily by a judgement method or based on the measurement of an auxiliary variable on the units selected. In this work, we derive different estimators of the parameters associated with the distribution of the study variate *Y*, based on ranked set sample obtained by using an auxiliary variable *X* correlated with *Y* for ranking the sample units, when (*X*, *Y*) follows a Morgenstern type bivariate logistic distribution. The theory developed in this paper is illustarted using a real data. Efficiency comparison among these estimators are also made.

Key words: Ranked set sample, Morgenstern type bivariate logistic distribution, Best linear unbiased estimator, Concomitants of order statistics.

1. INTRODUCTION

The concept of ranked set sampling was first introduced by McIntyre (1952) as a process of improving the precision of the sample mean as an estimator of the population mean. This is applicable whenever ranking of a set of sampling units can be done easily by a judgement method see, Chen *et al.* (2004). Ranking by judgement method is not recommendable if the judgement method is not mathematically much relevant with the problem of study. In certain situations, one may prefer exact measurement of some easily measurable variable associated with the study variable rather than making ranking by a crude judgement method. Suppose the variable of interest say *Y*, is difficult or much expensive to measure, but an auxiliary variable *X* correlated with *Y* is readily measurable and can be ordered exactly. In biological studies, such as in the root zone analysis of bamboo plants (*Bambusa arundinacea*), the shoot height of the plant is a correlated character with root weight. Clearly shoot height can be measured very easily whereas root weight measurement requires uprooting of the sampled plants. Hence, in such situations, we can choose the most desired plants with respect to their shoot length value and on which we measure the root weight for further analysis as presented in any ranked set sampling (RSS). Thus, as an alternative to McIntyre (1952) method of ranked set sampling, Stokes (1977) used an auxiliary variable for ranking of the sampling units, which is as follows: Choose *n* independent bivariate samples, each of size *n*. In the first sample, the *Y* variate associated with smallest ordered *X* is measured, in the second sample, the *Y* variate associated with the second smallest, *X* is measured. This process is continued until the*Y* variate associated with the largest is measured.

Stokes (1977) suggested the ranked set sample mean as an estimator for the mean of the study variate *Y*, when an auxiliary variable *X* is used for ranking the sample units, under the assumption that (*X*, *Y*) follows a bivariate normal distribution. Barnett and Moore (1977) improved it by deriving the Best Linear Unbiased Estimator (BLUE) of the mean of the study variate *Y*, based on ranked set sample obtained on the study variate *Y*. Chacko and Thomas (2007) obtained the BLUE of the parameter involved in the study variate *Y*, under the assumption that (*X*, *Y*) follows bivariate Pareto distribution. Unbalanced RSS arising from Morgentern type bivariate exponential distribution have been considered by Chacko and Thomas (2008).

Chacko and Thomas (2006) used the concomitants of record values arising from a Morgenstern type bivariate logistic distribution to estimate some of its parameters. Sampling to get a given number of record values will require several selection (uncertain number) of units and moreover the obtained concomitants of record values are correlated, which makes one to determine the variance and covariance of concomitants of record values to use them for inference problem. However, in case of Stokes method of ranked set sampling, the number of units to be selected is definite and there exists no correlation between one observation to another as they are drawn from independent samples so that handling the observations in a ranked set sample for inference problem will be very easy.

In this work, we consider the case when (*X*, *Y*) follows Morgenstern type bivariate logistic distribution (MTBLD) with cumulative distribution function (cdf) defined by (Kotz *et al.* 2000):

$$F_{X,Y}(x, y) = \left(1+\exp\left\{-\frac{x-\theta_1}{\sigma_1}\right\}\right)^{-1}\left(1+\exp\left\{-\frac{y-\theta_2}{\sigma_2}\right\}\right)^{-1} \left(1+\alpha\left[\frac{\exp\left\{-\frac{x-\theta_1}{\sigma_1}\right\}}{1+\exp\left\{-\frac{x-\theta_1}{\sigma_1}\right\}}\right]\left[\frac{\exp\left\{-\frac{y-\theta_2}{\sigma_2}\right\}}{1+\exp\left\{-\frac{y-\theta_2}{\sigma_2}\right\}}\right]\right) \tag{1.1}$$

In Section 2, we derive unbiased estimators of the parameters θ_2 and σ_2 involved in MTBLD defined by (1.1) based on a ranked set sample. In Section 3, we derive the BLUE of θ_2 and σ_2 involved in MTBLD based on the ranked set sample and have made an efficiency comparison with corresponding unbiased estimators developed in Section 2. In Section 4, we illustrate the methods developed in Section 2 and Section 3 using real data.

2. UNBIASED ESTIMATORS OF θ_2 AND σ_2

Let (*X*, *Y*) be a bivariate random variable which follows a MTBLD with cdf defined by (1.1). Suppose *n* sampling units each of size *n* are taken. Let $X_{(r)r}$ be the *r*th order statistic of the auxiliary variate *X* in the *r*th sample and let $Y_{[r]r}$ be the measurement made on the variate associated with $X_{(r)r}$, *r* = 1, 2, ..., *n*. By using the approach of Scaria and Nair (1999) for obtaining means and variances of concomitants of order statistics arising from Morgenstern family of distributions, we get the mean and variance of $Y_{[r]r}$ for $1 \le r \le n$ as

$$\mathrm{E}[Y_{[r]r}] = \theta_2 - \alpha\left(\frac{n-2r+1}{n+1}\right)\sigma_2 \tag{2.1}$$

$$\mathrm{Var}[Y_{[r]r}] = \left(\frac{\pi^2}{3} - \alpha^2\left(\frac{n-2r+1}{n+1}\right)^2\right)\sigma_2^2 \tag{2.2}$$

Since $Y_{[r]r}$ and $Y_{[s]s}$ for $r \neq$ s are drawn from two independent samples, we have

$$\mathrm{Cov}[Y_{[r]r}, Y_{[s]s}] = 0, r \neq s \tag{2.3}$$

If we write

$$\xi_r = -\alpha(n-2r+1)/(n+1) \tag{2.4}$$

and
$$\delta_r = \frac{\pi^2}{3} - \alpha^2\left(\frac{n-2r+1}{n+1}\right)^2 \tag{2.5}$$

then from (2.1) and (2.2), we can write

$$\mathrm{E}[Y_{[r]r}] = \theta_2 + \xi_r\sigma_2 \tag{2.6}$$

$$\mathrm{Var}[Y_{[r]r}] = \delta_r\sigma_2^2 \tag{2.7}$$

In the following theorem, we propose estimators θ_2^* and σ_2^* of θ_2 and σ_2 involved in (1.1) and prove that they are unbiased estimators for θ_2 and σ_2.

Theorem 2.1. Let $Y_{[r]r}$, *r* = 1, 2, ..., *n* be the ranked set sample observations on a study variate *Y* obtained out of ranking made on an auxiliary variate *Y*, when (*X*, *Y*) follows MTBLD as defined in (1.1). Then the ranked set sample mean given by

$$\theta_2^* = \frac{1}{n}\sum_{r=1}^{n} Y_{[r]r} \tag{2.8}$$

is an unbiased estimator of θ_2 and

$$\sigma_2^* = \frac{1}{\sum_{r=1}^{[n/2]}\xi_r}\sum_{r=1}^{[n/2]} T_r \tag{2.9}$$

is an unbiased estimator of σ_2, where $T_r = (Y_{[r]r} - Y_{[n-r+1]\overline{n-r+1}})/2$ and [.] is the usual greatest integer function. The variances of the above estimators are given by

$$\mathrm{Var}[\theta_2^*] = \frac{\sigma_2{}^2}{n}\left[\frac{\pi^2}{3} - \frac{\alpha^2}{n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right] \quad (2.10)$$

and

$$\mathrm{Var}[\sigma_2^*] = \frac{\sigma_2^2}{2(\sum_{r=1}^{[n/2]}\xi_r)^2}\sum_{r=1}^{[n/2]}\left[\frac{\pi^2}{3} - \alpha^2\left(\frac{n-2r+1}{n+1}\right)^2\right]$$

(2.11)

Proof.

$$\mathrm{E}[\theta_2^*] = \frac{1}{n}\sum_{r=1}^{n}\mathrm{E}[Y_{[r]r}] = \frac{1}{n}\sum_{r=1}^{n}\left[\theta_2 - \alpha\frac{n-2r+1}{n+1}\sigma_2\right]$$

(2.12)

Note that

$$\sum_{r=1}^{n}(n-2r+1) = 0 \quad (2.13)$$

Applying (2.13) in (2.12), we get

$$\mathrm{E}[\theta_2^*] = \theta_2$$

Thus θ_2^* is an unbiased estimator of θ_2. The variance of θ_2^* is given by

$$\mathrm{Var}[\theta_2^*] = \frac{1}{n^2}\sum_{r=1}^{n}\mathrm{Var}(Y_{[r]r})$$

Thus using (2.2) in the above sum, we get

$$\mathrm{Var}[\theta_2^*] = \frac{\sigma_2^2}{n}\left[\frac{\pi^2}{3} - \frac{\alpha^2}{n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right]$$

From (2.9) we have

$$\mathrm{E}[\sigma_2^*] = \frac{1}{\sum_{r=1}^{[n/2]}\xi_r}\sum_{r=1}^{[n/2]}\mathrm{E}[T_r]$$

$$= \frac{1}{2\sum_{r=1}^{[n/2]}\xi_r}\sum_{r=1}^{[n/2]}\mathrm{E}[Y_{[r]r} - Y_{[n-r+1]\overline{n-r+1}}]$$

On using (2.1) in the above equation and simplifying, we get

$$\mathrm{E}[\sigma_2^*] = \sigma_2$$

The variance of σ_2^* is given by

$$\mathrm{Var}[\sigma_2^*] = \frac{1}{(\sum_{r=1}^{[n/2]}\xi_r)^2}\sum_{r=1}^{[n/2]}Var[\mathrm{T}_r]$$

$$= \frac{1}{4(\sum_{r=1}^{[n/2]}\xi_r)^2}\sum_{r=1}^{[n/2]}Var[\mathrm{Y}_{[r]r}] + Var[\mathrm{Y}_{[n-r+1]\overline{n-r+1}}]$$

On using (2.2) in the above equation and simplifying, we get

$$\mathrm{Var}[\sigma_2^*] = \frac{\sigma_2^2}{2(\sum_{r=1}^{[n/2]}\xi_r)^2}\sum_{r=1}^{[n/2]}\left[\frac{\pi^2}{3} - \alpha^2\left(\frac{n-2r+1}{n+1}\right)^2\right]$$

Thus the theorem is proved.

We compare the variance of θ_2^* with the Cramer Rao Lower Bound $\pi^2\sigma_2^2/(3n)$ of any unbiased estimator of θ_2 involved in the marginal distribution of *Y* in (1.1). Clearly the ratio of $\pi^2\sigma_2^2/(3n)$ with variance of θ_2^* denoted by $e_1(\theta_2^*)$ is given by

$$e_1(\theta_2^*) = \frac{\frac{\pi^2}{3}}{\left[\frac{\pi^2}{3} - \frac{\alpha^2}{n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right]} \quad (2.14)$$

Since

$$\left[\frac{\pi^2}{3}-\frac{\alpha^2}{n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2\right] = n\sigma_2^2\text{Var}(\theta_2^*) \geq 0$$

We have $\frac{\alpha^2}{n}\sum_{r=1}^{n}\left(\frac{n-2r+1}{n+1}\right)^2 \leq \frac{\pi^2}{3}$ and hence $e_1(\theta_2^*) \geq 1$. Thus we conclude that there is some gain in efficiency of the estimator θ_2^* due to ranked set sampling. It is to be noted that $\text{Var}(\theta_2^*)$ is a decreasing function of α^2 and hence its variance is least when $\alpha = \pm 1$. Thus the gain in efficiency of the estimator θ_2^* is increasing as $|\alpha|$ is increasing.

Again on simplifying (2.14), we get

$$e_1(\theta_2^*) = \frac{\frac{\pi^2}{3}}{\frac{\pi^2}{3} - \alpha^2[\frac{2}{3}(\frac{2+1/n}{1+1/n}) - 1]}$$

Clearly

$$\lim_{n\to\infty} e_1(\theta_2^*) = \lim_{n\to\infty}\frac{\pi^2/3}{\pi^2/3-\alpha^2\left[\frac{2}{3}(\frac{2+1/n}{1+1/n})-1\right]}$$

$$= \frac{\pi^2/3}{\pi^2/3-\alpha^2/4}$$

Clearly the maximum asymptotic value for $e_1(\theta_2^*)$ is attained when $|\alpha| = 1$ and in this case $e_1(\theta_2^*)$ tends to $4\pi^2/(4\pi^2 - 3)$.

3. BEST LINEAR UNBIASED ESTIMATORS OF θ_2 AND σ_2

In this section we provide better estimators θ_2^* than and σ_2^* of θ_2 and σ_2 respectively by deriving the BLUE $\hat{\theta}_2$ and $\hat{\sigma}_2$ of θ_2 and σ_2 respectively provided the parameter α is known. Suppose n sampling units each of size n are taken from the population with cdf defined by (1.1). Let $Y_{[n]} = (Y_{[1]1}, Y_{[2]2}, \ldots, Y_{[n]n})'$. Then from (2.6), (2.7) and (2.3), the mean vector and the dispersion matrix of $Y_{[n]}$ are given by

$$\text{E}[Y_{[n]}] = \theta_2\mathbf{1} + \sigma_2\xi \qquad (3.1)$$

$$\text{D}[Y_{[n]}] = \sigma_2^2\mathbf{G} \qquad (3.2)$$

where $\xi = (\xi_1, \xi_2, \ldots, \xi_n)'$ and $\mathbf{G} = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$ in which ξ_r and δ_r are as defined in (2.4) and (2.5) respectively and 1 is a column vector of ones. If the parameter involved in ξ and **G** is known, then (3.1) and (3.2) together defines a generalized Gauss-Markov set up and hence the BLUEs $\hat{\theta}_2$ and $\hat{\sigma}_2$ of θ_2 and σ_2 are obtained as

$$\hat{\theta}_2 = \Delta^{-1}\left[\xi'\mathbf{G}^{-1}(\xi\mathbf{1}' - \mathbf{1}\xi)\mathbf{G}^{-1}\right]Y_{[n]} \qquad (3.3)$$

and

$$\hat{\sigma}_2 = \Delta^{-1}\left[1'\mathbf{G}^{-1}(\mathbf{1}\xi' - \xi\mathbf{1}')\mathbf{G}^{-1}\right]Y_{[n]} \qquad (3.4)$$

with variances given by

$$\text{Var}(\hat{\theta}_2) = \sigma_2^2(\xi'\mathbf{G}^{-1}\xi)/\Delta \qquad (3.5)$$

and

$$\text{Var}(\hat{\sigma}_2) = \sigma_2^2(\mathbf{1}'\mathbf{G}^{-1}\mathbf{1})/\Delta \qquad (3.6)$$

where $\Delta = (\xi'\,\mathbf{G}^{-1}\xi)(1'\mathbf{G}^{-1}1) - (\xi'\mathbf{G}^{-1}\mathbf{1})^2$. On substituting the values of ξ and **G** in (3.3) and (3.4) and simplifying, we get

$$\hat{\theta}_2 = \sum_{r=1}^{n}\left\{\frac{\delta_r^{-1}(\sum_{i=1}^{n}\xi_i^2\delta_i^{-1}) - \xi_r\delta_r^{-1}(\sum_{i=1}^{n}\delta_i^{-1})}{(\sum_{i=1}^{n}\delta_i^{-1})(\sum_{i=1}^{n}\xi_i^2\delta_i^{-1}) - (\sum_{i=1}^{n}\xi_i\delta_i^{-1})^2}\right\}Y_{[r]r} \qquad (3.7)$$

and

$$\hat{\sigma}_2 = \sum_{r=1}^{n}\left\{\frac{\xi_r\delta_r^{-1}(\sum_{i=1}^{n}\delta_i^{-1}) - \delta_r^{-1}(\sum_{i=1}^{n}\xi_i\delta_i^{-1})}{(\sum_{i=1}^{n}\delta_i^{-1})(\sum_{i=1}^{n}\xi_i^2\delta_i^{-1}) - (\sum_{i=1}^{n}\xi_i\delta_i^{-1})^2}\right\}Y_{[r]r} \qquad (3.8)$$

Table 3.1. Variances and efficiences of the estimators

n	α	Var(θ_2^*)	Var$(\hat{\theta}_2)$	Var(σ_2^*)	Var$(\hat{\sigma}_2)$	e_1	e_2
2	0.25	1.641	1.641	236.370	236.371	1.000	1.000
	0.50	1.631	1.631	58.718	58.718	1.000	1.000
	0.75	1.614	1.614	25.819	25.819	1.000	1.000
	1.00	1.589	1.589	14.304	14.304	1.000	1.000
4	0.25	0.819	0.819	81.934	65.387	1.00001	1.253
	0.50	0.810	0.810	20.249	16.038	1.0002	1.263
	0.75	0.794	0.794	8.826	6.899	1.001	1.280
	1.00	0.772	0.770	4.828	3.699	1.003	1.305
6	0.25	0.546	0.546	47.548	36.558	1.00002	1.301
	0.50	0.538	0.538	11.725	8.922	1.0002	1.314
	0.75	0.526	0.525	5.091	3.802	1.001	1.339
	1.00	0.509	0.506	2.769	2.009	1.005	1.379
8	0.25	0.409	0.409	33.146	25.158	1.00002	1.317
	0.50	0.403	0.403	8.163	6.123	1.0003	1.333
	0.75	0.393	0.392	3.537	2.597	1.002	1.362
	1.00	0.379	0.377	1.918	1.360	1.006	1.410
10	0.25	0.327	0.327	25.345	19.123	1.00002	1.325
	0.50	0.322	0.322	6.237	4.646	1.0004	1.342
	0.75	0.314	0.313	2.699	1.964	1.002	1.374
	1.00	0.302	0.300	1.460	1.024	1.007	1.427
12	0.25	0.273	0.273	20.482	15.403	1.00002	1.330
	0.50	0.268	0.268	5.038	3.738	1.0004	1.348
	0.75	0.261	0.260	2.178	1.577	1.002	1.381
	1.00	0.251	0.249	1.177	0.819	1.007	1.437
14	0.25	0.234	0.234	17.170	12.887	1.00002	1.332
	0.50	0.230	0.230	4.221	3.125	1.0004	1.351
	0.75	0.223	0.223	1.824	1.316	1.002	1.385
	1.00	0.214	0.213	0.984	0.682	1.008	1.444
16	0.25	0.204	0.204	14.773	11.073	1.00003	1.334
	0.50	0.201	0.201	3.631	2.684	1.0004	1.353
	0.75	0.195	0.195	1.568	1.129	1.002	1.389
	1.00	0.187	0.186	0.845	0.583	1.008	1.449
18	0.25	0.182	0.182	12.959	9.705	1.00003	1.335
	0.50	0.179	0.179	3.184	2.351	1.0004	1.355
	0.75	0.173	0.173	1.374	0.988	1.002	1.391
	1.00	0.166	0.165	0.741	0.510	1.008	1.453
20	0.25	0.164	0.164	11.540	8.637	1.00003	1.336
	0.50	0.161	0.161	2.835	2.091	1.0004	1.356
	0.75	0.156	0.156	1.223	0.878	1.002	1.393
	1.00	0.149	0.148	0.659	0.453	1.009	1.456

The variances given by (3.5) and (3.6) can also be simplified as

$$\mathrm{Var}(\hat{\theta}_2) = \frac{\sum_{i=1}^{n} \xi_i^{2} \delta_i^{-1}}{(\sum_{i=1}^{n} \delta_i^{-1})(\sum_{i=1}^{n} \xi_i^{2} \delta_i^{-1}) - (\sum_{i=1}^{n} \xi_i \delta_i^{-1})^2} \sigma_2^2$$

and $\mathrm{Var}(\hat{\sigma}_2) = \dfrac{\sum_{i=1}^{n} \delta_i^{-1}}{(\sum_{i=1}^{n} \delta_i^{-1})(\sum_{i=1}^{n} \xi_i^{2} \delta_i^{-1}) - (\sum_{i=1}^{n} \xi_i \delta_i^{-1})^2} \sigma_2^2$

We have computed $\mathrm{Var}(\theta_2^*)$, $\mathrm{Var}(\hat{\theta}_2)$, efficiency $e(\hat{\theta}_2 \mid \theta_2^*) = \mathrm{Var}(\theta_2^*) \big/ \mathrm{Var}(\hat{\theta}_2)$ of $\hat{\theta}_2$ relative to θ_2^*, $Var(\sigma_2^*)$, $Var(\hat{\sigma}_2)$, efficiency $e(\hat{\sigma}_2 \mid \sigma_2^*) = \mathrm{Var}(\sigma_2^*) \big/ \mathrm{Var}(\hat{\sigma}_2)$ of $\hat{\sigma}_2$ relative to σ_2^* for $\alpha = 0.25(0.25)1$ and $n = 2(2)20$ and the same are given in Table 3.1. From this table one can easily see that $\hat{\theta}_2$ is relatively more efficient than θ_2^*. Further, we observe from the table that for fixed α, both $e(\hat{\theta}_2 \mid \theta_2^*)$ and $e(\hat{\sigma}_2 \mid \sigma_2^*)$ increasing with *n*. Also from the table we notice that the precision obtained for the BLUE $\hat{\sigma}_2$ is more than that obtained for $\hat{\theta}_2$.

Remark 3.1. If we have a situation with α unknown, we introduce an estimator (moment type) for α as follows. For MTBLD the correlation coefficient between the two variables is given by $\rho = 3\alpha/\pi^2$. If *r* is the sample correlation coefficient between $X_{(i)i}$ and $Y_{[i]i}$, $i = 1, 2, ..., n$ then the moment type estimator for α is obtained by equating with the population correlation coefficient ρ and is obtained as

$$\hat{\alpha} = \begin{cases} -1 \text{ if} & r \le -3/\pi^2 \\ 1 \text{ if} & r \ge 3/\pi^2 \\ r\pi^2/3 & \text{otherwise} \end{cases} \tag{3.9}$$

4. AN ILLUSTRATION

In this section, as an application of theory developed on RSS in the previous sections, we consider a bivariate data set from Platt *et al.* (1969) relating to 396 Confir (*Pinus Palustrine*) trees. In Chen *et al.* (2004) also, the above bivariate data set is reproduced in which the first component *X* of a bivariate observation represents the diameter in centimetres of the Confir tree at breast height and the second component *Y* represents height in feet of the tree. Clearly *X* can be measured easily but it is somewhat difficult to measure *Y*. Also observations, such ar girth (function of diameter) or height follows normal distribution. It is well known that logistic distribution is having more or less similar properties of a normal distribution (Malik 1985, p. 123) and hence it is known as an alternative model to normal distribution. Assume that (*X*, *Y*) follows Morgenstern type bivariate logistic distribution. We select 10 random samples each of size 10 from the 396 tree data and rank the sampling units of each sample according to the *X* variate values (diameter of the tree). From the *i*th sample, we measure the *Y* variate (height of the tree) corresponding to the ith order statistic of the *X* variate. The obtained RSS observations are reported in Table 4.1.

Table 4.1. RSS observations

i	$X_{(i)i}$	$Y_{[i]i}$
1	6.3	11
2	10.1	28
3	3.8	6
4	4.5	10
5	6.0	16
6	15.9	28
7	38.6	42
8	17.8	38
9	41.4	177
10	51.7	219

The sample correlation between *X* and *Y* is 0.883. Thus, from (3.9), an estimate of α is taken as 1. We have obtained the RSS estimators θ_2^* and $\hat{\sigma}_2$ derived in Section 2, the BLUEs $\hat{\theta}_2$ and $\hat{\sigma}_2$ based on RSS;

${\sigma_2}^{-2} Var(\theta_2^*)$, ${\sigma_2}^{-2} Var(\hat{\theta}_2)$, ${\sigma_2}^{-2} Var(\sigma_2^*)$ and ${\sigma_2}^{-2} Var(\hat{\sigma}_2)$ and are given in the Table 4.2 .

Table 4.2. Estimators and their variances

Estimator	Estimate	variance / ${\sigma_2}^2$
θ_2^*	57.500	0.3017
$\hat{\theta}_2$	60.745	0.2997
σ_2^*	95.062	1.9443
$\hat{\sigma}_2$	108.062	1.0236

The usual traditional estimators of involved in MTBLD is the sample mean $\overline{Y}$ and its variance is $\pi^2 {\sigma_2}^2 / (3n)$. For n = 10, this variance is $0.3287{\sigma_2}^2$, which is clearly larger than $\mathrm{Var}(\theta_2^*)$ and $\mathrm{Var}(\hat{\theta}_2)$. This establishes the advantage of estimating the mean height of trees more closely to the true value of the mean using RSS.

ACKNOWLEDGEMENTS

The authors are grateful to the referee for his valuable comments.

REFERENCES

Barnett, V. and Moore, K. (1997). Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering. *J. Appl. Statist.*, **24,** 697-710.

Chacko, M. and Thomas, P. Y. (2006). Concomitants of record values arising from Morgenstern type bivariate logistic distribution and some of their applications in parameter estimation. *Metrika*, **60,** 301-318.

Chacko, M. and Thomas, P.Y. (2007). Estimation of a parameter of bivaraite Pareto distribution by ranked set sampling. *J. Appl. Statist.*, **34,** 703-714.

Chacko, M. and Thomas, P.Y. (2008). Estimation of a parameter of Morgenstern type bivariate exponential distribution by ranked set sampling. *Ann. Instt. Statist. Math.*, **60,** 301-318.

Chen, Z., Bai, Z. and Sinha, B.K. (2004). *Lecture Notes in Statistics, Ranked Set Sampling, Theory and Applications.* Springer, New York.

Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000). *Distributions in Statistics: Continuous Multivariate Distributions.* Second ed., John Wiley and Sons, New York.

Malik, H.J. (1985). *Logistic Distribution*. Encyclopedia of Statistical Sciences, **5**, (eds S. Kotz and N.L. Johnson), John Wiley and Sons, New York.

McIntyre, G.A. (1952). A method of unbiased selective sampling, using ranked sets. *Austr. J. Agril. Res.*, **3,** 385-390.

Platt, W.J., Evans, G.M. and Rathbun, S.L. (1988). The population dynamics of a long-lived Conifer (*Pinus Palustris*). *Amer. Naturalist*, **131,** 491-525.

Scaria, J. and Nair, N.U. (1999). On concomitants of order statistics from Morgenstern family. *Biometrical J.*, **41**, 483-489.

Stokes, S.L. (1977). Ranked set sampling with concomitant variables. *Comm. Statist. – Theory and Methods,* **6**, 1207-1211.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 85-89

Length-weight Relationship and Growth Pattern of *Tor putitora* (Hamilton) under Monoculture and Polyculture Systems: A Case Study

N. Okendro Singh, Md. Wasi Alam[1], Amrit Kumar Paul[1] and Surinder Kumar[2]
Directorate of Coldwater Fisheries, Bhimtal, Uttaranchal

(Received: July 2006, Revised: July 2008, Accepted: January 2009)

SUMMARY

The population of the endangered coldwater fish species, Tor putitora has been sharply declined in the recent past and is threatened with multifaceted dangers. In the present investigation, an attempt has been made to develop the length-weight relationship of *Tor putitora* under monoculture and polyculture systems for direct use in fishery assessment and also to describe growth pattern in terms of weight of this fish species in the above two different culture systems. The von-Bertalanffy model was found to be the best suitable model to describe the growth pattern of *Tor putitora*.

Key words: Endangered, Fish stock, Growth pattern, Monoculture, Polyculture.

1. INTRODUCTION

In the recent past, tremendous downfall in the catches of *Tor putitora* (golden mahseer) has been experienced not only in the Himalayan region but also in other parts of India. The reasons behind this unabated declining trend are many, mostly related to anthropogenic activities and the environmental degradation, which in the pace of development has severely damaged the eco-systems holding mahseer and allied fisheries. Besides environmental stresses, indiscriminate killing of adults and juveniles has been recognized as a major factor responsible for declining of mahseer in the Himalayan region. *Tor putitora* is an endangered coldwater fish species that is a popular fish as food and as a source of recreation for anglers. As mahseer stock is threatened with multifaceted dangers, which are partly due to overexploitation and consequently reduced yield from many fish stocks. The size of fish plays an important role in fish stock assessment. In the present scenario, it may be worthy to know the growth pattern of this fish species under monoculture and polyculture systems, so that we could be able to provide proper management of mahseer stocks more precisely. Since, the size of fish is a primary driver for many key processes in fisheries systems and the majority of the data underlying stock assessments are size structured, the fitting of non-linear models will be of immense help to provide formal advice on stock management. Again, length-weight relationship of fish is important in fisheries biology because they allow the estimation of the average weight of the fish of a given length group by establishing a mathematical relationship between the two (Beyer 1987). They are also useful for assessing the relative well being of the fish population (Bolger and Conolly 1989). As length and weight of fish are among the important morphometric characters, they can be used for the purpose of taxonomy and ultimately in fish stock assessment. Once if we establish the mathematical relationship between length and weight of fish, we will be able to see the changes during various developmental events of life such as metamorphosis, the onset of maturity, etc. Thus, the present study aims primarily to establish the length-weight relationship more precisely of *Tor putitora* under monoculture and polyculture systems, which can be very useful in fish

[1] *Indian Agricultural Statistics Research Institute, New Delhi – 110 012*

[2] *Department of Statistics, DSB Campus, Kumaun University, Nainital, Uttarakhand*

stock assessment, secondly, to identify most suitable non-linear model for describing the growth pattern of *Tor putitora* over different period of time under monoculture and polyculture systems.

2. MATERIALS AND METHODS

A non-linear statistical model is one in which at least one derivative with respect to at least one parameter being a function of the parameter(s). Details of non-linear models have been given by Ratkowsky (1990). The length-weight relationship was calculated using the formula: $W = aL^b$ (Pauly 1984), where 'W' is the weight of the fish in gm and 'L' is the length of the fish measured in mm; 'a' and 'b' are parameters, the later being called the rate of allometric. Also, by taking logarithmic transformation on both sides of the above equation, we get the linearize model. log W = log a + b log$L$.

The following non-linear models (Seber and Wild 1989) have been tried to explain the growth pattern in the dataset considered.

(i) Von-Bertalanffy model

$$W_t = W_\infty - \left(W_\infty - K\right)\exp\left(-bt\right) + e$$

(ii) Logistic model

$$W_t = \frac{W_\infty}{\left[1 + b\exp\left(-Kt\right)\right]} + e \; ; \; b = \frac{W_\infty}{W\left(0\right)} - 1$$

(iii) Gompertz model

$$W_t = W_\infty \exp\left[-b\exp\left(-Kt\right)\right] + e$$

$$\mathrm{b} = \ln\left[W_\infty / W\left(0\right)\right]$$

(iv) Richards model

$$W_t = W_\infty \left[1 + b\exp\left(-Kt\right)\right]^{\left(-1/d\right)} + e$$

$$b = \left[\mathrm{W}_\infty^d / \mathrm{W}^d\left(0\right)\right] - 1$$

where, W_t is the observed fish weight during time t; K, b, W_∞ , d are the parameters, and e is the error term. The parameter K is the intrinsic growth rate and the parameter W_∞ represents asymptotic size (in weights) of the fish for each model. Symbol b represents different functions of the initial value $W(0)$ and d is the added parameter in Richards model.

Model Fitting

Being a part of the project programme conducted at NRCCWF, Bhimtal under the NATP, Joshi *et al*. (2004) monitored growth performance of Tor putitora in different culture systems. The mahseer stock were rearing at Pantnagar, Uttarakhand and raising under monoculture and polyculture systems with the stocking densities of 4000 and 1600 per hectare respectively in earthen ponds of 0.1 hectare each. During the rearing period of about two years, the average length and weight of mahseer was obtained at different unequal time intervals for monoculture and polyculture systems separately. In monoculture system, the specimens of this fish species ranged 234-431 mm in length and 320-950 gm in weight while in polyculture system, specimens varied between 244-445 mm in length and 325-980 gm in weight. Further, the initial and average weight gain in *Tor putitora* at an unequal interval during the rearing period of about two years viz. 0, 2, 5, 8, 11, 14, 17 and 22 months are respectively 320, 400, 475, 550, 650, 725, 825 and 950 gm in monoculture system whereas 325, 430, 500, 605, 710, 795, 850 and 980 gm in polyculture system. These data were further utilized for the present investigation. For model fitting, age of *Tor putitora* given in months was converted to corresponding approximate age in years. However, ordinary least squares method is used for fitting of linearized model.

There are four main methods available in literature (Seber and Wild 1989) to obtain estimate of the unknown parameters of a non-linear regression model, namely (a) Gauss-Newton Method, (b) Steepest-Descent Method, (c) Levenberg-Marquardt Technique and (d) Do Not Use Derivative (DUD) Method. Levenberg-Marquardt method is the most widely used and reliable procedure for computing non-linear least square estimates and has been applied under the present study.

To examine model performance, a measure of how the predicted and observed variables covary in time is needed. Thus, the coefficient of determination, R^2 is generally used. However, Kvalseth (1985) has emphasized that, although R^2 given by

$$\mathrm{R}^2 = 1 - \frac{\sum\left(W_t - \hat{W}_t\right)^2}{\sum\left(W_t - \bar{W}\right)^2}$$

is quite appropriate even for non-linear models, uncritical use of and sole reliance on R^2 statistics may fail to reveal important data characteristics and model inadequacies. Hence, summary statistics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Square Error (MSE) are also used.

$$\text{RMSE} = \left[\sum_{t=1}^{n}\left(W_t - \hat{W}_t\right)^2 \Big/ n\right]^{1/2}$$

$$\text{MAE} = \sum_{t=1}^{n}\left|\left(W_t - \hat{W}_t\right)\right| / n$$

$$\text{MSE} = \left[\sum_{t=1}^{n}\left(W_t - \hat{W}_t\right)^2 \Big/ (n-p)\right]$$

where

$\hat{W}_t$ Predicted fish weight of tth observation

$\bar{W}$ Average fish weight

n Number of observations, $t = 1, 2, ..., n$

p Number of parameters involved in the model

The better model will have the least values of these statistics. It is, further, recommended for residual analysis to check the assumptions made for the model to be developed. Thus, independence or the randomness assumption of the residuals needs to be tested before taking any final decision about the adequacy of the model developed. To test the independence assumption of residuals run test procedure is available in the literature (Ratkowsky 1990). However, the normality assumption is not so stringent for selecting non-linear models because their residuals may not follow normal distribution while it must be strictly follow for linear or linearize models.

The Analysis of Covariance (ANCOVA) method was applied to check if the regression lines of length-weight relationship for monoculture and polyculture are parallel using SPSS Syntax commands (Singh and Nayak 2007). The above non-linear models were fitted using the Non-linear Regression option on SPSS 12.0 version. Different sets of initial parameter values were tried to meet the global convergence criterion for best fitting of the non-linear models.

3. RESULTS AND DISCUSSION

The slopes of the two regression lines due to various culture systems are not significant different since (F = 0.506) and (p = 0.491) for the corresponding ANCOVA test as shown in Table 1. Thus, we have to fit a model of combined data for monoculture and polyculture systems regarding length-weight relationship of Tor putitora. The two different forms of allometric model (non-linear and linearized forms) are fitted to the combined data for monoculture and polyculture systems. The estimates of parameter, goodness of fit statistics and residual analysis results of the fitted models is presented in Table 2. In this case, R^2 values are approximately

Table 1. User-specified contrasts test results on comparing length-weight relationship in monoculture and polyculture systems of *Tor putitora* by ANCOVA method

Comparison of culture systems	Monoculture v/s polyculture
Mean Square	1.429×10^{-4}
F-value	0.506
p-value	0.491
Comment	Not significant

Table 2. Summary statistics of the models fitted to length-weight dataset of *Tor putitora*

	Non-linear model $W = aL^b$	Linearize model $\log W = \log a + b \log L$
Parameter Estimates		
a or Log *a*	0.016	–1.680 (0.021)*
b	1.810	1.761
Goodness of Fit Statistics		
R^2	0.994	0.993 (0.993)
RMSE	16.001	1.118×10^{-2} (18.314)
MAE	12.113	1.000×10^{-2} (14.022)
MSE	292.616	1.429×10^{-4} (293.472)
Residual Analysis		
Run Test (|Z|)	1.908	1.109
Shapiro-Wilk Test p-value	0.832	0.123

* The bracketed values are related to conversion of the linearize models to original (non-linear) models by taking antilogarithm.

same in all irrespective of the nature of models. The best-fitted model is decided based on the values of RMSE, MAE and MSE. Further, we have examined whether the assumptions about residuals are satisfied for the models or not. The run test |Z| values to check independence assumption of the residuals are below the critical value (1.96) of normal distribution at 5% level of significance, shows the suitability of the fitted models. Moreover, Shapiro-Wilk test p-values for the residuals clearly indicate that residuals are normally distributed in Table 2. Thus, the independence as well as normality assumption about residuals is satisfied for fitting of the models. On comparing the performance, the non-linear model gives better result to its corresponding linearize model. Hence, we conclude that non-linear model appears to describe more precisely the length-weight relationship of *Tor putitora* than its corresponding linearize model.

The estimates of parameters, R^2, RMSE, MAE, MSE, run test statistic (|Z|) value, Shapiro-Wilk test p-values for the above growth models are again given in Table 3. The same criteria are used to identify the best model, since R^2 values being almost equal in all the models and von-Bertalanffy model performs well for dataset of monoculture system. Further, we have examined whether the assumptions about residuals are satisfied for this model or not. The run test |Z| value to check independence assumption of the residuals is 1.146, which is below the critical value 1.96. Shapiro-Wilk test p-values given in Table 3 also clearly show that the normality assumption is not violated. Hence, von-Bertalanffy model seems to describe more precisely the growth pattern of *Tor putitora* under the monoculture system. The corresponding asymptotic size (in weight) of *Tor putitora* is 8913 gm approximately in the monoculture system. Furthermore, by working out the ratio of the weight during the last interval considered in the dataset to the asymptotic size, it is seen that only 11% of the maximum size is already achieved so far

Table 3. Summary statistics of the models fitted to the growth dataset of *Tor putitora*

	Monoculture System				Polyculture System			
	Von-Bertalanffy	Logistic	Gompertz	Richards	Von-Bertalanffy	Logistic	Gompertz	Richards
Parameter Estimates								
K	327.840	0.096	0.050	0.050	333.938	0.120	0.073	0.073
b	0.003	2.765	1.584	0.001	0.026	2.251	1.316	0.004
W_∞	8912.834	1266.355	1618.974	1618.577	1804.003	1126.581	1269.385	1269.320
d	-	-	-	0.001	-	-	-	0.000
Goodness of Fit Statistics								
R^2- value	0.998	0.998	0.998	0.998	0.997	0.995	0.996	0.996
RMSE	8.010	9.345	8.430	8.431	11.675	14.209	12.520	12.520
MAE	7.851	7.326	6.941	7.326	11.225	10.801	10.463	10.053
MSE	102.663	139.718	113.706	142.148	218.089	323.053	250.780	313.495
Residual Analysis								
Run test |Z| value	1.146	1.146	1.146	1.146	0.382	0.382	0.382	0.382
Shapiro Wilk Test p-value	0.720	0.913	0.738	0.738	0.531	0.637	0.550	0.531
Asymptotic size (in weight) achieved at present	11%	75%	58%	59%	54%	87%	77%	77%

and hence, there is immense scope for further increase in weight of this fish species under the monoculture system. Similarly, von-Bertalanffy model is again found appropriate for the dataset under polyculture system. The required assumptions about residuals of this model are also satisfied. Further, it is observed that 54% of the maximum size has already been achieved and scope for further increase in weight of this fish species is more limited under polyculture system as compare to monoculture system. Also, von-Bertalanffy model fitted to the data of monoculture system shows better fit than the polyculture system.

REFERENCES

Beyer, J.E. (1987). On length-weight relationship. *Fishbyte*, **5,** 11-13.

Bolger, T. and Conolly, P.L. (1989). The selection of suitable indices for measurement and analysis of fish conditions. *J. Fish. Biol.*, **34(2)**, 171-182.

Joshi, C.B., Sunder, S., Murgnandam, M., Sharma, A.P., Chahan, R.S., Das, M., Dhanze, J.R., Dhanze, R., Balkhi, M.H. and Samoon, M.S. (2004). Aquaculture management in coldwaters: Evaluation of mahseer fishery potentials and its farming feasibility for conservation in Himalayan region. Unpublished NATP Project Report, NRC on Coldwater Fisheries, ICAR, Bhimtal, Uttarakhand.

Kvalseth, T.O. (1985). Cautionary note about R^2. *Amer. Statist*., **39(4)**, 279-285.

Pauly, D. (1984). Fish population dynamics in tropical waters: A manual for use with programmable calculators. *ICLARM Stud. Rev*., **8**, 325.

Ratkowsky, D.A. (1990). *Handbook of Non-linear Regression Models*. Marcel Dekker, New York.

Seber, G.A.F. and Wild, C.J. (1989). *Non-linear Regression.* John Wiley and Sons, New York.

Singh, N. Okendro and Nayak, A.K. (2007). Analysis of covariance (ANCOVA) on evaluation of sex and seasonal differences in length-weight relationship of a fish species. *The SPSS Analyst*, October-December Issue, 1-12.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 91-96

Decision Support System for Nutrient Management in Crops

S. Pal, I.C. Sethi and Alka Arora
Indian Agricultural Statistics Research Institute, New Delhi

(Received: January 2006, Revised: June 2008, Accepted: January 2009)

SUMMARY

Nutrient Management plays a vital role in increasing crop production, soil upgradation and increase in profitability. Taking these things into consideration, a Decision Support System on Nutrient Management in Crops (DSSNMC) has been designed and developed at Indian Agricultural Statistics Research Institute (IASRI). DSSNMC is a Web-based Decision Support System (DSS) and provides decision to farmers on nutrient management in crops. The system will have great importance in agriculture as experts are not always available to answer farmers' queries. DSSNMC has three modules to provide decision support to farmers in three different situations. First module is the subsystem based on soil test values. Herein, the user gets an advice for fertilizer application based on the information provided for soil test values, crop to be grown, variety of that crop, sowing season, soil type and targeted yield (within a particular range). In case, soil was not tested, then the farmer can use the second module which provides decision support on the basis of location such as zone or district. The system requires information of the location of the farm in terms of zone or district, targeted yield and rest of the values like available nitrogen, phosphorus, potassium and the soil pH are taken from the data base, where standard values for different districts or zones are stored.

Third module of the system helps in controlling nutrient deficiency of standing crops based on abnormal growth as seen through deficiency symptoms shown by the crop. The basis here is the observation of the farmers, which they compare with the images, stored in the system and can use the corrective measures provided by the system. The testing and validation of the system was done using the data of different cooperating centers under All India Coordinated Research Project (AICRP) of Soil Crop Response Correlation (STCR).

Key words: Decision support system, Nutrient management, Web application, .Net technologies.

1. INTRODUCTION

Agriculture continues to remain the major contributing sector of the Indian economy even after 60 years of independence. It contributes 30% of GNP, provides 65% of employment and continues to be primary source of living. The adoption of new agricultural technology brought out the green revolution that boosted agriculture production in India. However, this process has caused nutrient imbalance in the soils due to rapid depletion of soil fertility because of heavy withdrawal of essential plant nutrients by bumper harvests requiring proper nutrient management in the soils. Many studies on soils, chemical fertilizers, plant and their relationship have been carried out, but farmers' queries could not be answered satisfactorily on nutrient management as the information is scattered at different places and the experts are not always available (Watermann 1988). The scattered information can be utilized effectively by the farmers through a "Decision Support Systems" (DSS). The present DSS (Pal 2005) is an attempt in this direction.

The nutrient application in a field is dependent on the nature of the soil fertility in the field i.e. by testing the soil and applying the fertilizers on the basis of relevant soil test values. But many a times, farmers do not have appropriate soil test values and as such under this situation fertilizer application can be done on the basis of soil type of area under cultivation, as has been identified by soil scientists. However, despite of all care many a times, a crop shows deficiency symptoms. In such cases, additional fertilizers can be applied to the crop. Kumar (1992) developed a DSS for micronutrient management in the soils. DSS developed by Patil (2002) provided recommendations on how much of chemical fertilizer is to be applied, instead of dose for a fertilizer source (as farmers require source wise application). The present system based on .net technology provides decision for fertilizer application based on the information provided for soil test values, crop to be grown, variety of that crop, sowing season, soil type and targeted yield (within a particular range).

2. SYSTEM ARCHITECTURE AND REQUIREMENTS

DSSNMC has been implemented in a three layered structure i.e. User Interface Layer (UIL), Application Layer (APL) and Database Layer (DBL). UIL is implemented using HTML (Hyper Text Markup Language) and JavaScript. The User Interface Layer consists of forms for accepting information from the user and validating those forms using JavaScript. APL has been implemented using ASP.NET. ASP.NET is a powerful and flexible technology for creating dynamic Web pages. It is a convergence of two major Microsoft Technologies, Active Server Pages (ASP) and the .NET Framework. ASP.NET is a way of creating dynamic Web pages, using the innovations present in the .NET Framework (Ullman 2005). DBL has been implemented using Microsoft Access 2000. The relational approach has been used to design the database. The fundamentals of normalization theory have been used to normalize different tables of the database (Loney 2004). All tables have proper interaction among themselves via primary key - foreign key relationship. The entity relationship (ER) diagram of DSSNMC is given in Fig. 1.

3. FUNCTIONALITIES OF DSSNMC

DSSNMC is developed as a web-based application, using .NET technology. Therefore, it is platform independent and can be accessed from any computer connected to the internet. The only requirement at the client side is a web-browser. The most commonly used web browsers are internet explorer 6.0 or above from

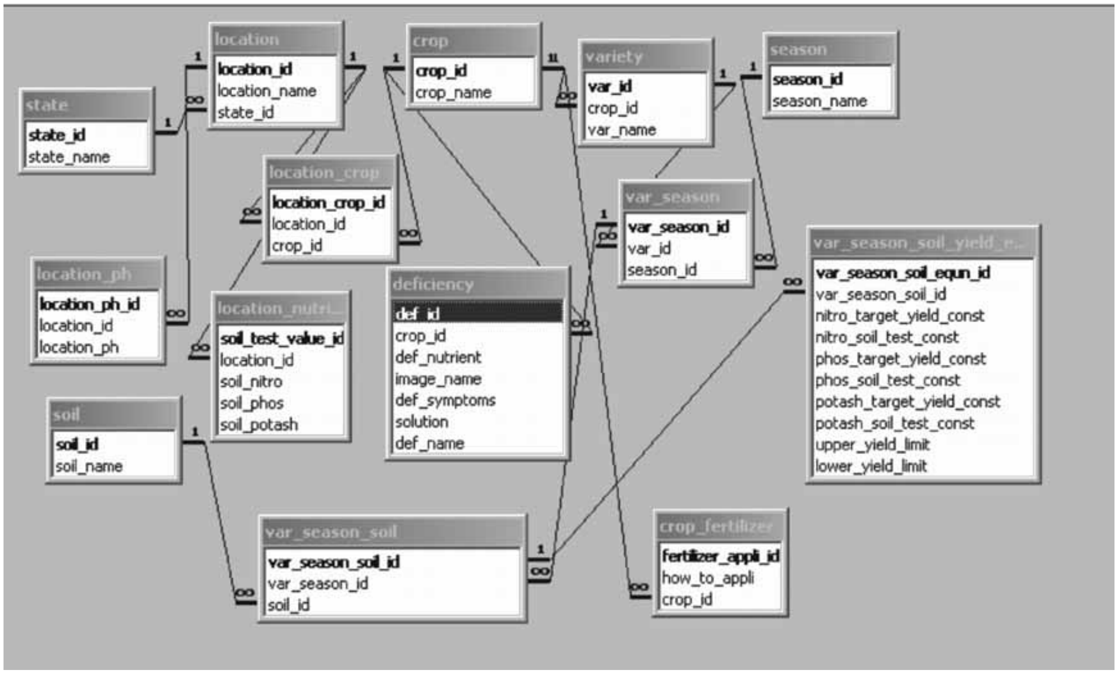


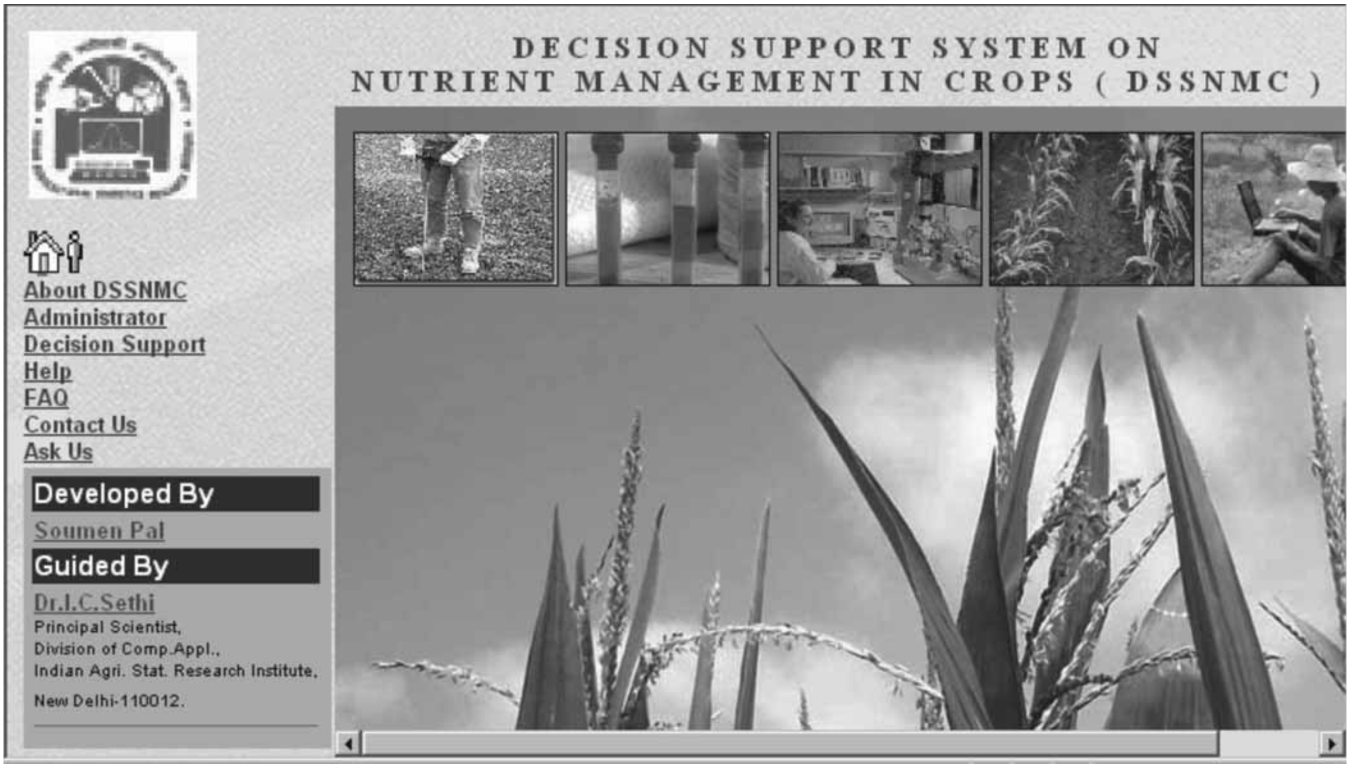
Fig. 1. ER diagram of DSSNMC

Microsoft Corporation and netscape communicator from netscape Communications. DSSNMC successfully runs on both the browsers.

DSSNMC system can be hosted from the server having Internet Information Server (IIS) installed on it. System home page is shown in Fig. 2.

3.1 Type of Users

The system has been designed keeping in view the requirements of two types of users i.e. System Administrators and End Users. Administrators are the users who manage the system and they, therefore, have the right to add, modify, delete or update any part of the information captured in the database. Therefore,



DECISION SUPPORT SYSTEM ON NUTRIENT MANAGEMENT IN CROPS (DSSNMC)

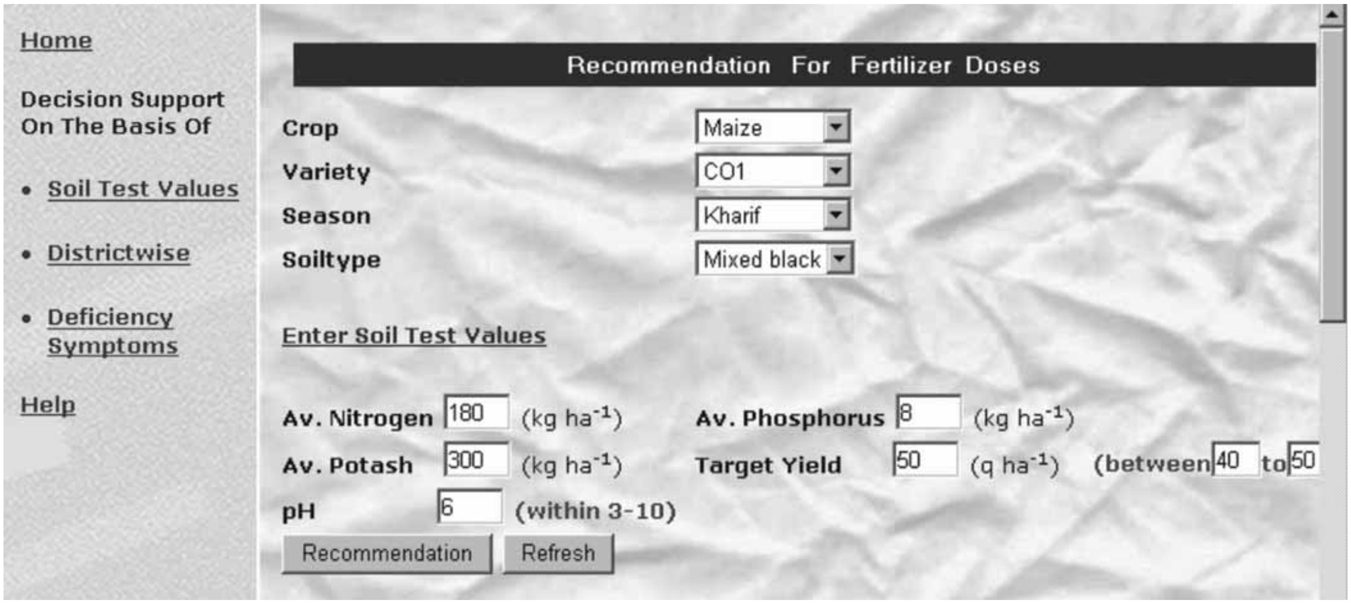
Developed By

Soumen Pal

Guided By

Dr.I.C.Sethi
Principal Scientist,
Division of Comp.Appl.,
Indian Agri. Stat. Research Institute,
New Delhi-110012.

Fig. 2. Home Page of DSSNMC



Recommendation For Fertilizer Doses

Crop: Maize
Variety: CO1
Season: Kharif
Soiltype: Mixed black

Enter Soil Test Values

Av. Nitrogen 180 (kg ha^{-1}) Av. Phosphorus 8 (kg ha^{-1})
Av. Potash 300 (kg ha^{-1}) Target Yield 50 (q ha^{-1}) (between 40 to 50
pH 6 (within 3-10)

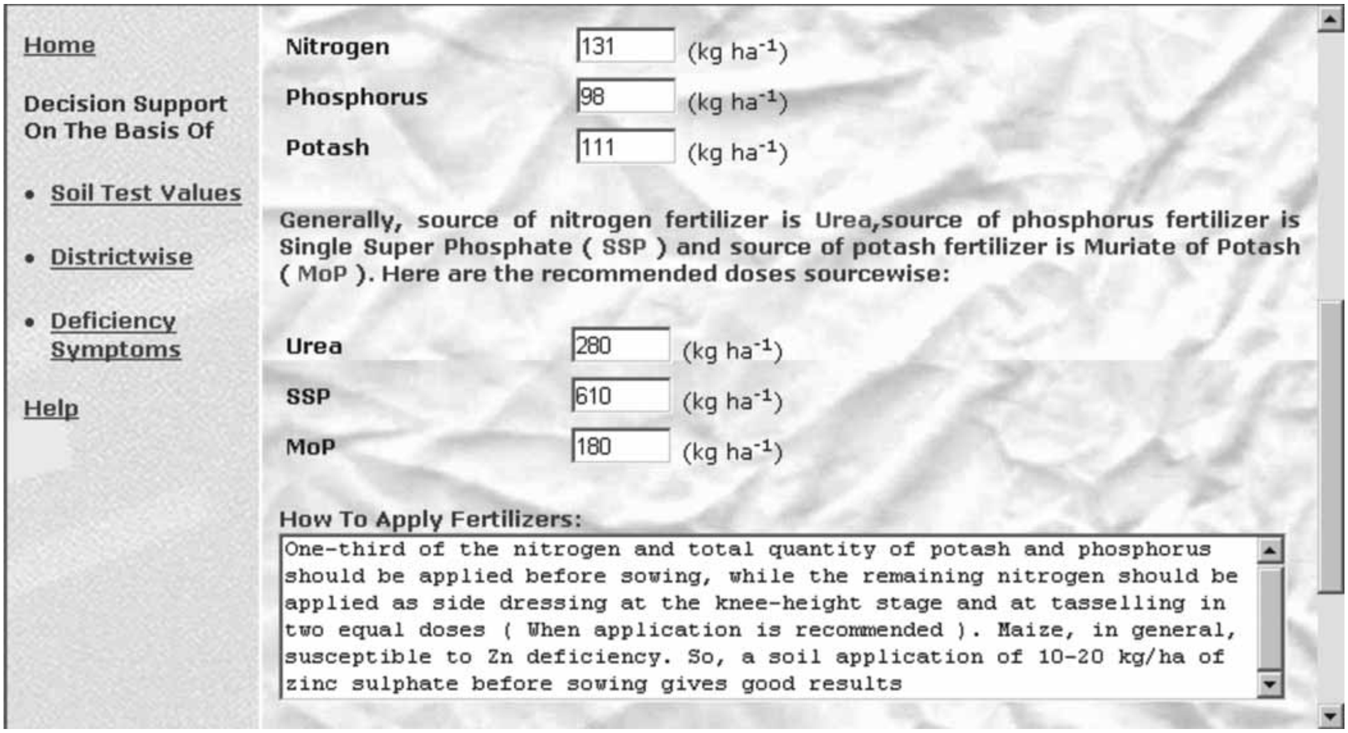
Fig. 3. Information about crop and soil test values

Administrator has user name and password protection for accessing the system. End users are the farmers, who can start the system, and get the desired decision. For this they can start the system and enter the desired data and get the decision from the system. If they do not get satisfied with the advice flashed on monitor, they can view the frequently asked questions, to quench their queries or send e-mail to concerned developers and get the satisfactory answers.

3.2 Decision Support

The present system consists of three subsystems, based on three different situations of decision making for nutrient management in crops. System design strategies for these subsystems are discussed below:

(a) **A subsystem based on soil test values:** In this subsystem, users provide soil test values along with desired crop to be grown, variety of that crop, season for that particular variety, soil type and target yield within a particular range. Recommendation for application of chemical fertilizers (for supplying the requirement of nitrogen, phosphorus and potash) based on above data is provided by the DSS. Further, general recommendation of how to apply fertilizers for a particular crop and also an additional recommendation based on soil pH value is provided (Fig. 3, 4 and 5).



Nitrogen 131 (kg ha^{-1})

Phosphorus 98 (kg ha^{-1})

Potash 111 (kg ha^{-1})

Generally, source of nitrogen fertilizer is Urea,source of phosphorus fertilizer is Single Super Phosphate (SSP) and source of potash fertilizer is Muriate of Potash (MoP). Here are the recommended doses sourcewise:

Urea 280 (kg ha^{-1})

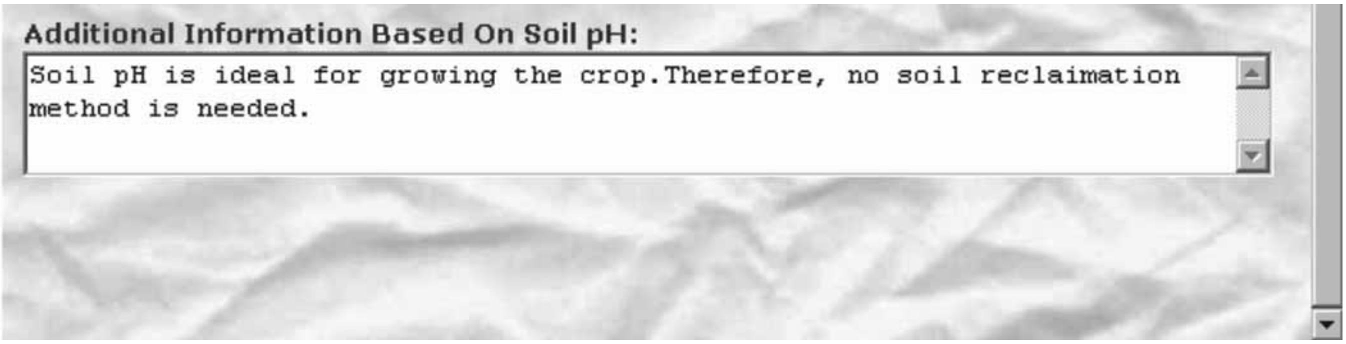
SSP 610 (kg ha^{-1})

MoP 180 (kg ha^{-1})

How To Apply Fertilizers:

One-third of the nitrogen and total quantity of potash and phosphorus should be applied before sowing, while the remaining nitrogen should be applied as side dressing at the knee-height stage and at tasselling in two equal doses (When application is recommended). Maize, in general, susceptible to Zn deficiency. So, a soil application of 10-20 kg/ha of zinc sulphate before sowing gives good results

Fig. 4. Source of fertilizer applied

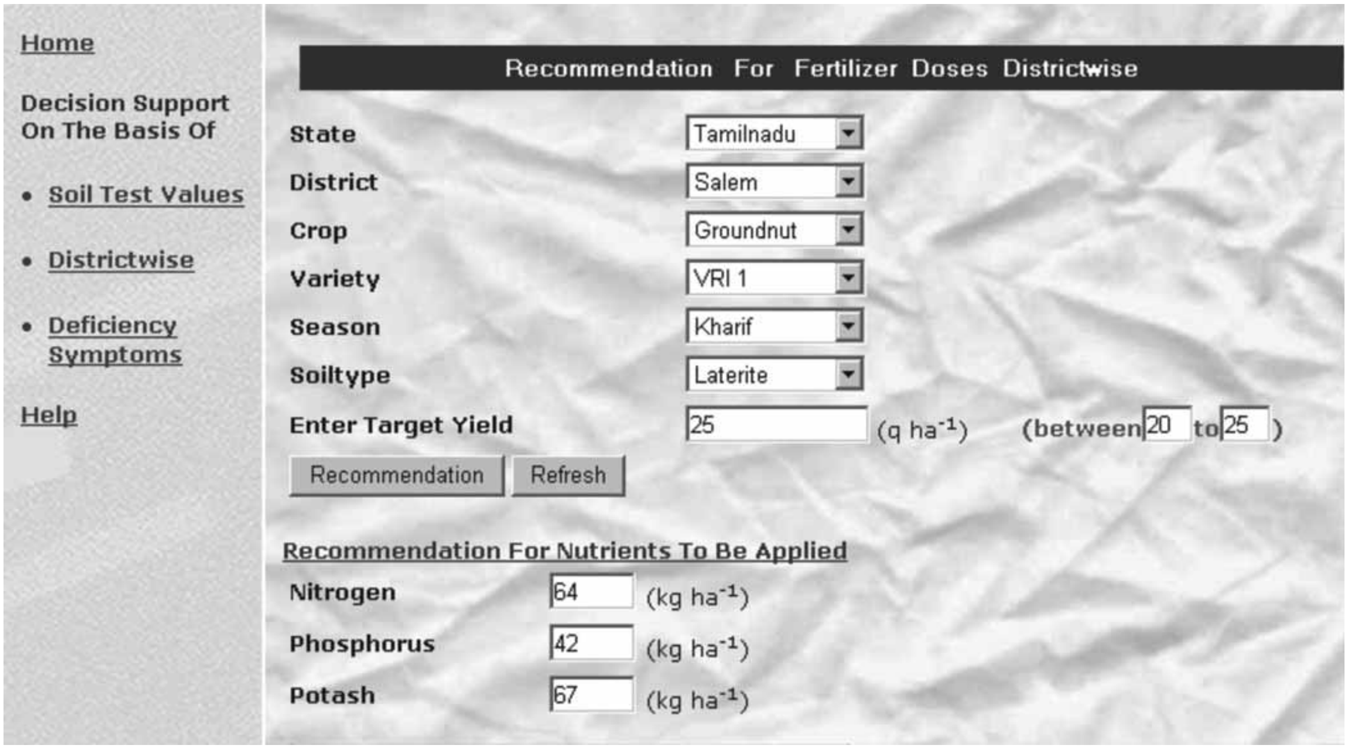


Additional Information Based On Soil pH:

Soil pH is ideal for growing the crop.Therefore, no soil reclaimation method is needed.

Fig. 5. Effect of pH value

(b) **Subsystem based on location such as zone or district:** This module will take input as the location of the farm in terms of zone or district and targeted yield and rest other values like, available nitrogen, phosphorus, potassium and the soil pH are taken from data base, where standard values for



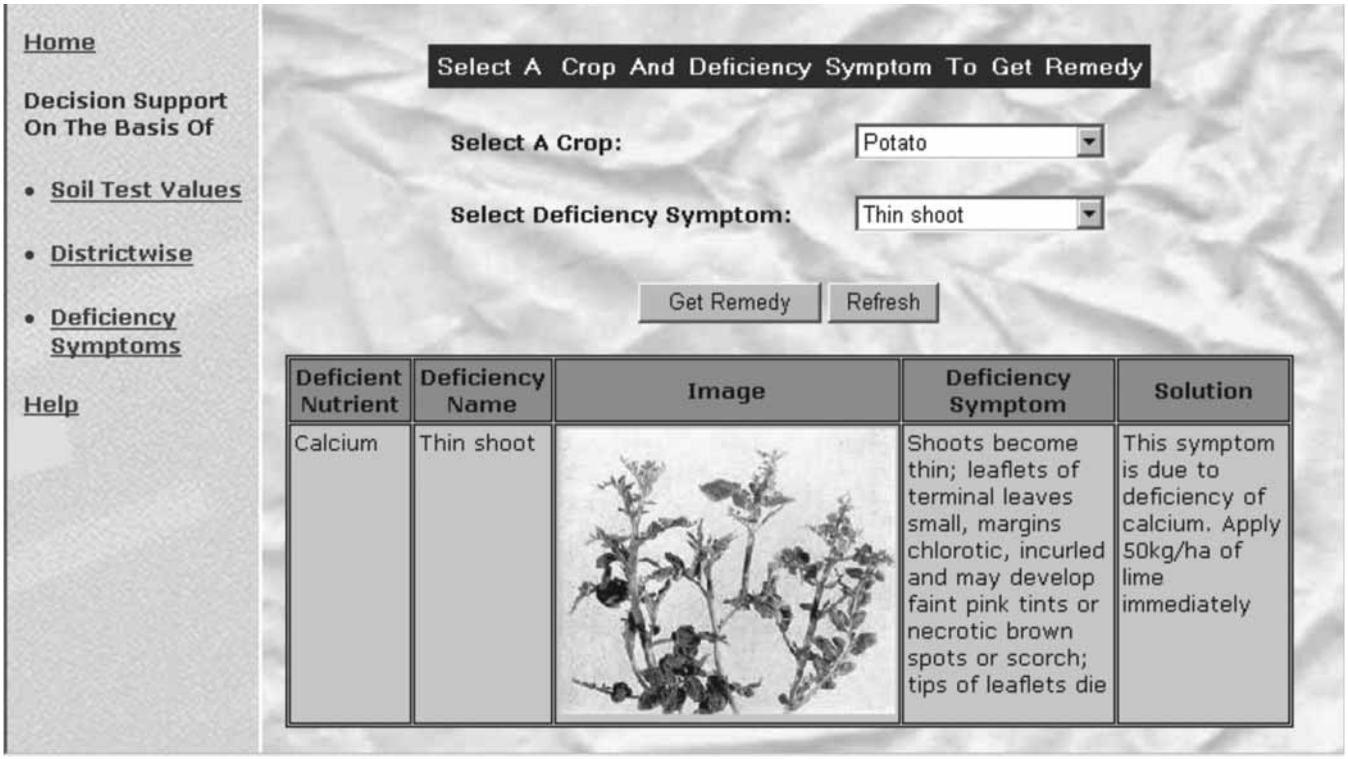
Recommendation For Fertilizer Doses Districtwise

State	Tamilnadu
District	Salem
Crop	Groundnut
Variety	VRI 1
Season	Kharif
Soiltype	Laterite
Enter Target Yield	25 (q ha^{-1}) (between 20 to 25)

Recommendation For Nutrients To Be Applied

Nitrogen	64 (kg ha^{-1})
Phosphorus	42 (kg ha^{-1})
Potash	67 (kg ha^{-1})

Fig. 6. Information about the location of the farm



Select A Crop And Deficiency Symptom To Get Remedy

Select A Crop: Potato

Select Deficiency Symptom: Thin shoot

Deficient Nutrient	Deficiency Name	Image	Deficiency Symptom	Solution
Calcium	Thin shoot		Shoots become thin; leaflets of terminal leaves small, margins chlorotic, incurled and may develop faint pink tints or necrotic brown spots or scorch; tips of leaflets die	This symptom is due to deficiency of calcium. Apply 50kg/ha of lime immediately

Fig. 7. Deficiency symptoms for a crop

particular zone or district are stored and decision is provided in the same way as in the previous module (Fig. 6).

(c) **Subsystem based on deficiency symptoms:** Proper fertility management requires the ability to recognize deficiency symptoms, either to correct them early in the growing season or to prevent them in subsequent parts of the season. In this module, user can get information for a particular crop on the basis of deficiency symptom observed by him, which he tallies with the images provided in the system (Fig. 7).

4. CONCLUSION

DSSNMC has been developed for providing decision support to the farmers, students, research workers, extension workers and others for nutrient management in crops. It can be implemented as network-based with a server. Currently, there are provisions for rice, rapeseed, wheat, groundnut, maize and potato crops in the system. But it will work for any crop. The system is menu driven and user-friendly. Hopefully, the use of this system will provide intellectual support to farmers in proper application of fertilizers to their crops.

REFERENCES

Kumar, A. (1992). Decision support system for micronutrient management in crops. Unpublished M.Sc. thesis, I.A.R.I., New Delhi.

Loney, K. and Koch, G. (2004). *Oracle 9i: The Complete Reference.* First Edition, Tata McGraw Hill, New Delhi.

Pal, S. (2005). Decision support system for nutrient management in crops. Unpublished M.Sc. Thesis, I.A.R.I., New Delhi.

Patil, A.N. (2002). Decision support system for nutrient management in wheat, mustard and bajra. Unpublished M.Sc. Thesis, I.A.R.I., New Delhi.

Ullman, C., Kauffman, J., Hart, C., Sussman, D. and Maharry, D. (2005). *Beginning ASP.NET 1.1 with Visual C#.NET* 2003. First Edition, Wiley Dreamtech India (P) Ltd., New Delhi.

Waterman, D.A. (1988). *A Guide to Expert Systems*. Addison -Wesley Publishing Company, USA.

J. Ind. Soc. Agril. Statist.
63(1), 2009 : 97-107

Machine Learning for Forewarning Crop Diseases

Rajni Jain, Sonajharia Minz[1] and Ramasubramanian V.[2]
National Centre for Agricultural Economics and Policy Research, New Delhi

(Received: June 2006, Revised: November 2008, Accepted: December 2008)

SUMMARY

With the advent of computers, the development of accurate forewarning systems for incidence of crop diseases has been increasingly emphasized. Timely forewarning of crop diseases will not only reduce yield losses but also alert the stakeholders to take effective preventive measures. Traditionally, Logistic Regression (LR) and discriminant analysis methods have been used in forewarning systems. Recently, several machine learning techniques such as decision tree (DT) induction, Rough Sets (RS), soft computing techniques, neural networks, genetic algorithms etc. are gaining popularity for predictive modelling. This paper presents the potential of three machine learning techniques viz. DT induction using C4.5, RS and hybridized rough set based decision tree induction (RDT) in comparison to standard LR method. RS offers mathematical tools to discover hidden patterns in data and therefore its application in forewarning models needs to be investigated. A DT is a classification scheme which generates a tree and a set of rules representing the model of different classes from a given dataset. A java implementation of C4.5 (CJP) is used for DT induction. A variant of RDT called RJP, combines merits of both RS and DT induction algorithms. *Powdery mildew of Mango* (PWM) is a devastating disease and has assumed a serious threat to mango production in India resulting in yield losses of 22.3% to 90.4%. As a case study, prediction models for forewarning PWM disease using variables viz. temperature and humidity have been developed. The results obtained from machine learning techniques viz. RS, CJP and RJP are compared with the prediction model developed using LR technique. The techniques RJP and CJP have shown better performance over LR approach.

Key words: Forewarning crop diseases, Machine learning, Rough sets, RDT, Decision tree, Logistic regression, Powdery mildew of mango.

1. INTRODUCTION

With the advent of computers, the development of accurate forewarning systems for incidence of crop diseases has been increasingly emphasized. Crop diseases are one of the major causes of reduction in crop yield and hence timely application of remedial measures may combat the yield loss to a great extent. Forewarning systems can help in providing prior knowledge of the time and severity of the outbreak of such diseases. Crop diseases are influenced by interaction of various factors with the most significant of them being weather. Normally data on crop disease status and information on related variables (including weather) over years are utilized for developing models/rules for forewarning of diseases. Developing forewarning systems for crop diseases is now made relatively easier by increasingly research efforts in the application of advanced and complicated statistical computing concepts which include inter alia soft computing techniques such as neural networks, fuzzy theory, rough sets etc. Timely forewarning of crop diseases will not only reduce yield losses but also alert the stakeholders to take effective preventive measures. Forewarning consists of examining the features of a newly presented case and assigning it to a predefined class. In general it can be treated as task

[1] *School of Computer Science, Jawaharlal Nehru University, New Delhi 110067*

[2] *Indian Agricultural Statistics Research Institute, New Delhi 110012.*

of classification which is characterized by the well-defined classes, and a training set consisting of pre-classified examples. The task is to build a model called classifier that can be applied to unclassified data in order to classify it. Machine Learning offers many techniques like decision tree induction algorithms, neural networks, genetic algorithms, rough sets, fuzzy sets as well as many hybridized strategies for the classification (Han and Kamber, 2001; Pujari, 2000; Komorowski *et al.,* 1999; Witten and Frank, 1999). On the other hand, traditional statistical techniques such as Logistic Regression (LR) and discriminant analysis may be employed for the task of classification. The potential of three machine learning techniques viz. DT induction using C4.5, RS and hybridized rough set based decision tree induction (RDT) has been compared with the standard LR method. As a case study, prediction models for forewarning Powdery mildeW of Mango (PWM) disease using causal variables viz. temperature and humidity have been developed. While developing the models, the study also identifies best set of variables and the suitable algorithms for forewarning of PWM disease.

The purpose of this study has arisen out the need for developing crop disease forewarning systems which are evolved upon reliable, robust and improved soft computing methods. Various approaches are in vogue to build such early warning systems. Every approach has its own advantages and limitations. Soft computing techniques can be advantageously used in certain situations to convert abstract knowledge and heuristics into easily comprehensible rules. The expected gain in accuracy by using soft computing concepts such as rough sets or its hybridized model may justify the effort involved in using them in preference to the conventional models.

The rest of the paper is organised as follows. Section 2 deals with the preliminaries. Section 3 describes a case study. Section 4 presents the methodology used followed by results and discussion in Section 5. Finally, the conclusions are presented in the Section 6.

2. PRELIMINARIES

2.1 Logistic Regression

Let class variables are of 0-1 type. To handle the task of classification (Hastie *et al*. 2001) in Logistic Regression (LR) approach, the probability of membership in the first group, $p_1(x)$, is modelled directly as in equation (1) for the two categories problem where α and β are the parameters.

$$p_1(x) = \frac{e^{\alpha+\beta' x}}{1+e^{\alpha+\beta' x}} \quad (1)$$

2.2 Decision Tree

Decision tree induction represents a simple and powerful method of classification which generates a tree and a set of rules, representing the model of different classes, from a given dataset (Winston 1992). Decision Tree (DT) is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node represents the class (Hans and Kamber 2001). The top most node in a tree is the root node. For DT induction, ID3 algorithm and its successor C4.5 algorithm by Quinlan (1993) are widely used. Algorithm CJP (java implementation of C4.5) is used in this paper for DT induction. One of the strengths of decision trees compared to other methods of induction is the ease with which they can be used for numeric as well as non-numeric domains. Another advantage of decision tree is that it can be easily mapped to rules. The classical DT induction algorithm i.e. C4.5 by Quinlan (1993) is presented below for better understanding to the readers.

2.2.1 C4.5 algorithm

Let the training data is a set $S = s_1, s_2, ...$ of already classified samples. Each sample $s_i = x_1, x_2, ...$ is a vector where $x_1, x_2, ...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2, ...$ where $c_1, c_2, ...$ represent the class that each sample belongs to. C4.5 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. C4.5 examines the normalized Information Gain (difference in entropy) that results from choosing an attribute for splitting the data. Entropy(S) can be thought of as a measure of how random the class distribution is in S. Information gain is a measure given to an attribute a. Attribute a can separate S into subsets $S_{a1}, S_{a2}, S_{a3}, ..., S_{an}$. The information gain of a is then Entropy(S) – Entropy(S_{a1}) – Entropy(S_{a2}) – ... – Entropy(S_{an}). Information gain is then normalized by multiplying the entropy of each attribute choice by the proportion of attribute values that have that choice. The attribute with the highest normalized information gain is the one used to make the decision. The algorithm then recurs on the smaller sub lists. The pseudo code of the algorithm is as follows:

```
1. Check for base cases
2. For each attribute a
   2.1 Find the normalized information gain from splitting on a
3. Let a-best be the attribute with the highest normalized information gain
4. Create a decision node dnode that splits on a-best
5. Recur on the sub lists obtained by splitting on a-best and add those nodes as children of dnode
```

2.3 Rough Set Theory

RS theory was introduced in early 1980s by Z. Pawlak, a Polish mathematician and has been widely explored for pattern discovery since then. RS emerged as an important mathematical tool for managing uncertainty that arises in the indiscernibility between objects in a set, and has proved to be useful in a variety of knowledge discovery processes (Pawlak 1991, Komorowski 1977). Some of the basic terms and concepts pertaining to RS are discussed below.

2.3.1 Information system and decision table

In RS, knowledge is a collection of facts expressed in terms of the values of attributes that describe the objects. These facts are represented in the form of a data table. Entries in a row represent an object.

A data table described as above is called an information system. Formally, an information system S is a 4-tuple, $S = (U, Q, V, f)$ where, U a nonempty, finite set of objects is called the universe; Q a finite set of attributes; $V = \cup Vq$, $\forall q \in Q$ and Vq being the domain of attribute q; and $f : U \times Q \rightarrow V$, f be the information function assigning values from the universe U to each of the attributes q for every object in the set of examples.

2.3.2 Indiscernibility relation

For a subset $P \subseteq Q$ of attributes of an information system S, a relation called indiscernibility relation denoted by IND is defined in equation (2).

$$\mathrm{IND}_S(P) = \{ (x, y) \in U \times U : f(x, a) = f(y, a) \ \forall a \in P\} \tag{2}$$

The function $f(x, a)$ assigns the value of the attribute a for an object x. If $(x, y) \in \mathrm{IND}_S(P)$ then objects x and y are called indiscernible with respect to P. The subscript s may be omitted if information system is implied from the context. IND(P) is an equivalence relation that partitions universe U into equivalence classes, the sets of objects indiscernible with respect to P. Set of such partitions are denoted by U/IND(P).

2.3.3 Approximation of sets

Let $X \subseteq U$ be a subset of the universe. A description of X is desired that can determine the membership status of each object in U with respect to X. Indiscernibility relation is used for this purpose. If a partition defined by IND(P) (denoted by Y in Equation 3) partially overlaps with the set X, the objects in such an equivalence class can not be determined without ambiguity. The description of such a set X is defined in terms of P-lower approximation (denoted as $\underline{P}$) and P-upper approximation (denoted as $\overline{P}$) where for $P \subseteq Q$:

$$\begin{aligned} \underline{P}X &= \cup\{Y \in U/\mathrm{IND}(P) : Y \subseteq X\} \\ \overline{P}X &= \cup\{\mathrm{Y} \in U/\mathrm{IND}(P) : Y \cap X \neq \emptyset\} \end{aligned} \tag{3}$$

A set X for which $\underline{P}X = \overline{P}X$ is called an exact set otherwise it is called rough set with respect to P.

2.3.4 Dependency of attributes

RS introduces a measure of dependency of two subsets of attributes P, $R \subseteq Q$. The measure is called a degree of dependency of P on R, denoted by $\gamma_R(P)$. It is defined as

$$\begin{aligned} \gamma_R(P) &= \frac{\mathrm{Card}(\mathrm{POS}_R(P))}{\mathrm{Card}(U)} \text{ where } \mathrm{POS}_R(P) \\ &= \underset{X \in U/\mathrm{IND}(P)}{\square} \underline{R}X \end{aligned} \tag{4}$$

The set POS_R(P), positive region, is the set of all the elements of U that can be uniquely classified into partitions U/IND(P) by R. Here, Card refers to the cardinality of the set included in the parenthesis. Thus, numerator and denominator are the number of objects in the positive region denoted by $\mathrm{POS}_R(P)$ and the universe U respectively. Coefficient γ_R (P) represents the fraction of the number of objects in the universe which can be properly classified. If P totally depends on R then $\gamma_R(P) = 1$, else $\gamma_R(P) < 1$.

2.3.5 Reduct

The minimum set of attributes that preserves the indiscernibility relation is called a reduct. The relative reduct of the attribute set P, $P \subseteq Q$, with respect to the dependency $\gamma_P(Q)$ is defined as a subset RED(P, Q) $\subseteq$ P such that:

1. $\gamma_{\text{RED}(P,\,Q)}(Q) = \gamma_P(Q)$, i.e. relative reduct preserves the degree of inter attribute dependency
2. For any attribute $a \in \text{RED}(P, Q)$, $\gamma_{\text{RED}(P,Q)-\{a\}}(Q) < \gamma_P(Q)$ i.e. the relative reduct is a minimal subset with respect to property 1.

Computation of a minimal optimum reduct is a NP hard problem. However a single relative reduct can be computed using efficient heuristics. Johnson's algorithm is one such method which is available in Rosetta software (http://www.idi.ntnu.no/~aleks/rosetta/).

2.3.6 Rule discovery

Rules can be perceived as data patterns that represent the relationships between attribute values. RS theory provides mechanism to generate rules directly from the dataset by reading the values of the attributes present in reduct from the given decision table.

2.4 Proposed Model – Rough Set based Decision Tree (RDT)

RDT model as proposed by Minz and Jain (2003a) combines merits of both RS and DT induction algorithm. It aims to improve efficiency, simplicity and generalization capability of both the base algorithms as shown by Minz and Jain (2003b). In the present study, a variant of RDT called RJP (Table 3) is used as a representative of RDT approach. Algorithm RJP for the induction of rough decision tree is presented below (Jain and Minz 2003).

Algorithm RJP

1. Input the training dataset say T1.
2. Discretize the continuous attributes if any, and label the modified dataset as T2.
3. Obtain the minimal decision relative reduct of T2, say R.
4. Reduce T2 based on reduct R and label the reduced dataset as T3.
5. Apply C4.5 algorithm on T3 to induce decision tree.
6. Convert the decision tree to rules (if needed) by traversing all the possible paths from root to each leaf node.

The training data-T1 is a collection of examples used for supervised learning to develop the classification model. In step 2, continuous attributes of the dataset (if any) are discretized. The next step involves computation of a reduct R. The reduct helps in reducing the training data, which is finally used for decision tree induction. Algorithms like Boolean reasoning algorithm, Johnson's algorithm or Genetic Algorithms can be used for the computation of the optimal reduct. In this paper, Johnson's algorithm based on efficient heuristics (implemented in Rosetta software), is used for the computation of a single reduct. More details pertaining to RDT model are available in Minz and Jain (2005).

3. CASE STUDY

Powdery Mildew of Mango (PWM) caused by Oidium mangiferae Berthet is responsible for foliar as well as inflorescence infection in mango. Generally, PWM epidemic occurs in the third and fourth week of March when the inflorescences are of the age of 6-7 weeks. The spread of the disease is greatly manifested by factors such as temperature, humidity, wind velocity, dews, wind direction etc. because it is an airborne disease.

The PWM dataset for the study has been taken from the project "Epidemiology and forecasting of PWM" undertaken at Central Institute for Subtropical Horticulture, Uttar Pradesh. From the original data, the attributes relative humidity and maximum temperature are selected because of the prior information available about contribution of these factors to the occurrence of PWM (Misra *et al*. 2004). As repeated life cycles of PWM are around 4-7 days, periods from 8th of March up to 14th of March i.e. one day a prior to the start of possible occurrence of epidemic (3rd week of March) were taken for developing forewarning models. Moving averages of maximum temperature and relative humidity are computed for March 8th –11th , 8th –12th , 8th –13th and 8th –14th and are referred by the set of corresponding pair of variables as{(T811, H811), (T812, H812), (T813,

Table 1. Pre-processed dataset for Powdery Mildew of mango

Year	T811	H811	T812	H812	T813	H813	T814	H814	STATUS
1987	28.20	91.25	28.48	88.60	29.50	85.50	30.14	82.86	1
1988	32.05	75.75	31.64	80.60	31.27	77.83	30.66	79.57	0
1989	26.60	86.25	26.28	87.20	26.47	87.33	26.31	89.14	0
1990	27.50	91.25	28.12	91.20	28.17	92.00	28.43	91.00	1
1991	28.43	87.00	28.70	86.20	29.00	83.50	29.57	80.57	0
1992	30.12	69.23	30.45	68.58	30.80	68.31	31.25	67.82	1
1993	30.50	61.75	30.48	61.13	30.37	60.56	30.33	61.76	0
1994	30.45	89.25	30.56	85.80	30.63	83.17	30.71	81.14	1
1995	28.63	61.38	29.10	61.20	29.58	61.17	30.71	61.57	0
1996	31.63	60.33	31.90	60.87	32.67	60.89	33.07	59.76	1
1997	32.13	71.00	32.20	69.40	31.67	69.00	31.50	68.29	0
2000	29.00	78.33	29.23	78.60	29.36	78.83	29.52	79.14	0

H813), (T814, H814)} in Table 1. Data is partitioned into train and test pairs as shown in Table 2. For example, the entry 1987-94 under the first column called MODEL means the data for the years 1987-94 is used for learning the model while the data for the years 1995-97 and 2000 is used for the model validation.

4. METHODOLOGY

All the eight independent variables as shown in Table 1 along with STATUS as dependent variable were used as input for the machine learning algorithms. However in case of LR, only two variables can be taken at a time as the number of observations in the dataset (Table 1) is less. Machine learning algorithms as well as traditional logistic regression method are employed using the training and test data pairs as identified in Table 2. The algorithms and the corresponding software that are used in this paper for forewarning of PWM disease are presented in Table 3. The Logistic Regression (LR) model has already been applied to the dataset by Misra *et al*. (2004). The redundant variables (if any) are filtered using concepts of information theory in CJP algorithm and using concept of reducts in RS and RJP algorithm. In this section, we demonstrate the characteristics of the output from each of the algorithms (Table 3) with the help of an example using the train data for the years1987-97 and the test data for the year 2000 (Table 2). The overall mean accuracy of the models for each algorithm is presented and discussed in Section 5.

Table 2. Train and test data pairs for different models

Model	Train Data	Test Data
1987-94	1987-94	1995, 1996, 1997, 2000
1987-95	1987-95	1996, 1997, 2000
1987-96	1987-96	1997, 2000
1987-97	1987-97	2000

Table 3. Learning algorithms used for PWM dataset

Id	Algo	Description	Software	Model
1	LR	Logistic Regression	SAS	Coeffi-cients
2	RS	Rough Set reducts (decision relative full discernibility global)	Rosetta	Rules
3	CJP	Java Implementation of C4.5 Pruned	Weka	DT
4	RJP	Rough set based DT induction embedding J4.8 for DT induction with Pruned tree	Rosetta, Weka, C++ programs	DT

4.1 LR

Consider the LR model given in equation (1) noting that for the two variables case, the expression $\alpha + \beta' x$ would be $a + bT(.) + cH(.)$. Table 4 shows estimates of parameters *a*, *b* and *c* for the model using the attribute pairs i.e. (*T*811, *H*811), (*T*812, *H*812), (*T*813, *H*813) or (*T*814, *H*814) separately for the train data 1987-97 (Table 2). Using these parameter estimates, the outcomes for the training set and the test set can be predicted by plugging in the corresponding parameter estimates from Table 4 in equation (1).

To decide whether the status of the disease is of epidemic nature, it is necessary to have a cut off value beyond which the probability value lies. It is fairly realistic to keep as a thumb rule the cut off value of probability as 0.5. Then if probability is less than this value then the event that epidemic will occur will be minimal, otherwise there is more chance of occurrence of disease in epidemic proportions. It is emphasized here that there is no objective procedure to be considered as a general rule. If one wants to be more stringent, then the cut off value can be increased as per requirement. Statistically speaking, depending upon the problem under consideration there is always a possibility of error because we deal with sample data for model development. Thus, the consideration of 0.5 as a cut off value in the present study is to a greater extent appropriate.

For the test data 2000, using (*T*811, *H*811), (*T*812, *H*812), (*T*813, *H*813) or (*T*814, *H*814), the corresponding $p(x)$ values as defined in equation (1) are 0.44, 0.44, 0.39 and 0.29. All these probabilities being less than 0.5 imply that the predicted STATUS is 0 which is same as the observed STATUS (Misra *et al.* 2004).

Table 4. Parameters of the LR model developed for the PWM prediction

Model	Years	1987-97
8th to 11th day	*a*	–10.79
	b	0.19
	c	0.06
8th to 12th day	*a*	–13.97
	b	0.3
	c	0.06
8th to 13th day	*a*	–36.95
	b	0.88
	c	0.13
8th to14th day	*a*	–70.43
	b	1.73
	c	0.24

4.2 RS

Employing RS approach for the different train data (Table 2), the set {*H*811, *T*814} is a computed reduct (Section 2.3.5). By using the discretized train data of the years 1987-97, the following three rules are generated. The rules are simple to comprehend for applying to the unseen dataset.

1. If (*H*811> = 88.2) AND (*T*814 <31.0) => then STATUS =1
2. If (*H*811 < 88.2)) AND (*T*814<31.0) => then STATUS = 0
3. If (*H*811<88.2)) AND (*T*814> = 31.0) => then STATUS = 1

The rules when applied to the test data i.e. the year 2000, correct prediction is obtained as illustrated in the following example.

Example 1: For the year 2000, *H*811 = 78.33 and *T* 814 = 30.71. Observing the three rules, we can identify that the Rule 2 is applicable to this dataset. Therefore, predicted value of the STATUS is 0. This is verified by the observed value of the STATUS (Table 1).

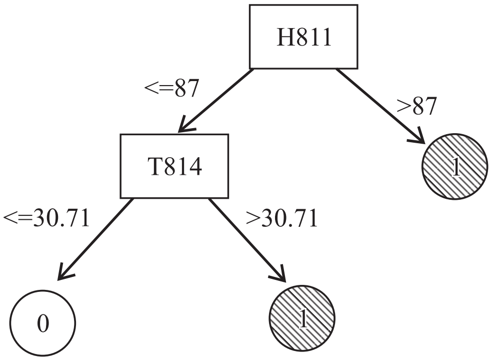
4.3 CJP

The prediction model which is obtained by employing the CJP algorithm using data of 1987-97 as the train data, is represented as decision tree in Figure 1. The corresponding rules are obtained by following the path from the root of the decision tree towards its leaf (Fig. 1).

Example 2: For the year 2000, we observe from Table 1 that *H*811 = 78.33 and *T*814 = 30.71. Consider the decision tree (output of CJP) in Figure 1. Starting from the root node and following the tree as per matching of the conditions in each branch, we reach the final node, also called leaf, having value 0. Therefore the predicted STATUS = 0 which is same as the observed value of the STATUS (Table 1). The prediction method using the rules from the decision tree is similar to Example 1.

4.4 RJP

Like CJP, the model obtained from RJP algorithm is a decision tree which can be mapped to rules (Fig. 2). However, it is observed that the branches representing the conditions are different from the branches of the decision tree from CJP algorithm.

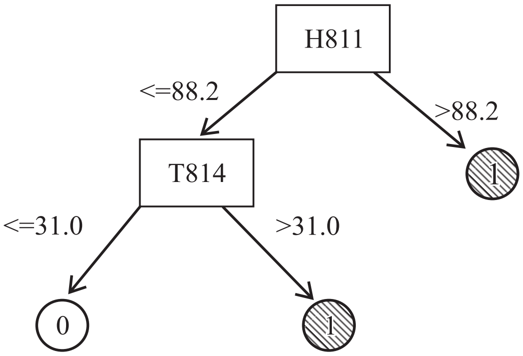


Rules:

1. If (H811>87) then Status = 1
2. If (H811<=87) and (T814<=30.71) then Status = 0
3. If (H811<=87) and (T814>30.71) then Status = 1

Note:
The Status = 1 means that epidemic will occur while Status=0 means no epidemic.

Fig. 1. The Prediction model for PWM Epidemic as obtained using CJP Algorithm on 1987-97 data as the training dataset



Rules:

1. If (H811>88.2) then Status = 1
2. If (H811<=88.2) and (T814<=31.0) then Status = 0
3. If (H811<=88.2) and (T814>31) then Status = 1

Note:
The Status = 1 means that epidemic will occur while Status=0 means no epidemic.

Fig. 2: The Prediction model for PWM Epidemic as obtained using RJP Algorithm on 1987-97 data as the training dataset

4.5 Performance Evaluation

Considering the costs associated with the wrong prediction of the PWM disease, accuracy is considered as the most important evaluation measure. In order to estimate the average accuracy initially the models are developed using the algorithms listed in Table 3. The average accuracy of the corresponding models was computed for each algorithm using training data and test data pairs (Table 2) and by considering each of the attribute sets - {*T*811, *H*811}, {*T*812, *H*812}, {*T*813, *H*813}, {*T*814, *H*814}. To determine average accuracy using all variables attribute set {*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814} is used for each algorithm. STATUS is used as the decision variable for each run of the algorithm. The superset of all the attributes, for example {*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814}, has not been used for LR because of its limitation in handling datasets when the order of the number of attributes is same as the number of observations. However, all the variables along with the pairs of variables are used for the RS, CJP and RJP algorithms to investigate whether the accuracy estimates improve by including all the variables for the training. Inclusion of all variables helps to identify the best set of input attributes for learning the model.

5. RESULTS AND DISCUSSION

The comparative discussion of the algorithms is done using the Fig. 3 and Fig. 4. Fig. 3(i) shows the overall mean accuracy of the models using two variables at a time. For example, mean accuracy of RJP algorithm in Fig. 3 (i) is computed using the formula:

$$\text{Mean Accuracy} = \frac{\sum_{v}\sum_{m}\text{Accuracy}_{vm}}{n} \qquad (5)$$

where each *v* i.e. variable pair in {{*T*811, *H*811}, {*T*812, *H*812},{*T*813, *H*813},{*T*814, *H*814}}; each *m* in {1987-94, 1987-95, 1987-96, 1987-97}; and *n* = number of observations.

Further, part (ii) of Fig. 3 presents the estimated mean accuracy using the variable set (*T*811, *H*811, *T*812, *H*812, *T*813, *H*813, *T*814, *H*814). As LR does not permit use of all variable pairs together, the results are not shown for LR in this figure.

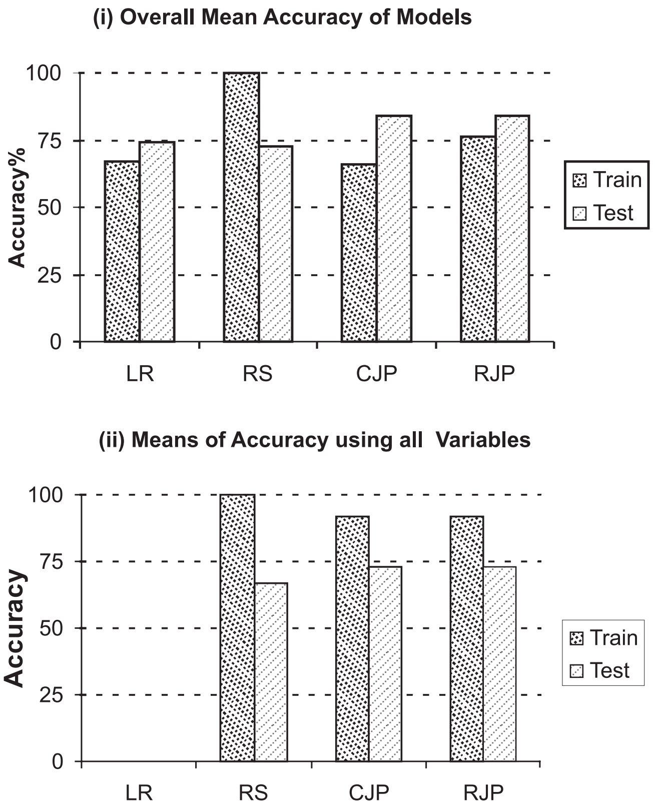


Fig. 3. (i)The Mean of accuracy with two variables at a time

(ii) The Mean of accuracy from with all variables at a time. As LR does not allow using all variables, the bar is not shown for LR in this graph

Fig. 4 helps in comparing and selecting the appropriate pair of variables for disease prediction. Using the results for the test data, the variables *T*811, *H*811 can be selected for better predicting capabilities in all machine learning algorithms. Although the results for the training data exhibited different scenario, yet the average accuracy estimates for the test data are considered more realistic. The behaviour of each of the algorithm as observed from the Fig. 3 and Fig. 4 is discussed separately.

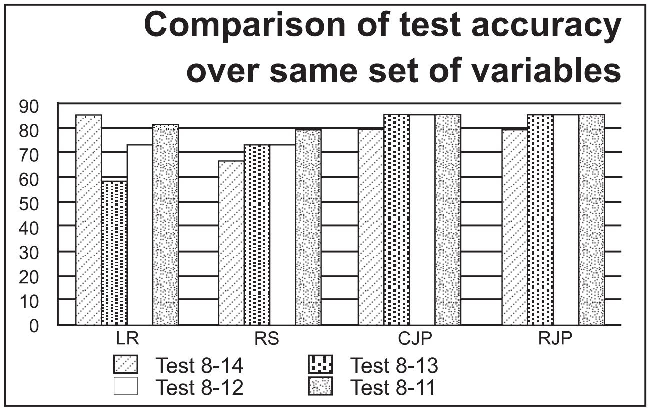


Fig. 4. Comparison of accuracy estimated by using data over the same set of variables

Note: LR gives its best performance on the test data using moving averages of max. temperature and humidity covering 8-14 days while other algorithms can give best accuracy estimates using 8-11 days of temperature and humidity information.

5.1 LR

Trends pertaining to the mean of computed accuracies are shown in Fig. 3(i). Training accuracy is observed to be lesser as compared to test accuracy (75%) for this method. For the purpose of identifying the best pair of variables, mean accuracies are presented for each set of the variables in Fig.4. When the attributes corresponding to the moving averages of more number of days are used the test accuracy estimate worsens except for the *T*814 and *H*814 (Fig. 4). It is in contrast to the general opinion that more data, either in terms of the size of the training data or in terms of more days in computing the moving averages would result in improved performance of the model (Fig. 4). As per the test data estimates, it is suggested to use *T*811, *H*811 variables for forewarning of PWM disease using LR algorithm because this results in forewarning well ahead in time without much loss of accuracy as compared to *T*814, *H*814 variables.

5.2 RS

Attribute pairs used for LR were also used for RS for the sake of comparison between the two. The mean accuracy on the training set is observed as 100 per cent for each case irrespective of the size of the training data or the set of attributes (Fig. 3(i) and Fig. 3(ii)). Mean test accuracy for RS is observed to be much less as compared to training accuracy for paired variables as well as while using all the 8 variables together. When performance of RS is compared with LR, it is observed that RS performs well on training data while LR is better for test data (Fig. 3). To identify the best pair of variables, it is observed from Figure 4 that with addition of one day in computation of moving averages, test accuracy deteriorates from 79.2 per cent for the variables (*T*811, *H*811) to 66.7 per cent for the variables (*T*814, *H*814). Thus, the pair (*T*811, *H*811) is recommended for forewarning of PWM disease using RS algorithm. However, use of all the 8 attributes as input to the RS algorithm has resulted in identification of *H*811 and *T*814 as relevant attributes with mean accuracy shown in Fig. 3(ii).

5.3 CJP

Comparison of overall mean accuracies of CJP with LR and RS shows that CJP performs well on the test data unlike training data (Fig. 3). As the aim of any forewarning model is to have better prediction for the test cases which are unseen as well, CJP is preferable in comparison to LR and RS approaches for disease prediction. Figure 4 shows that increasing one day at a time for computation of moving averages of attributes does not affect the test accuracy except for the case of *T*814, *H*814 showing little decrease in test accuracy. Thus for the reasons mentioned as above for LR algorithm, *T*811, *H*811 is recommended for predicting PWM disease. Parallel use of all attributes results in the selection of the attribute set {*H*811} or {*H*811, *T*814} as the most relevant attributes. However, it has not resulted in improvement of the test accuracy over the pair of attributes.

5.4 RJP

Test accuracy is improved for RJP algorithm as per expectations even though it does not show 100 per cent accuracy on the training data unlike RS (Fig. 3). Like CJP, increasing the number of days in computation of attributes does not show any impact on test accuracy but decrease the test accuracy for the case of *T*814, *H*814 (Fig. 4). Hence variables *T*811, *H*811 are recommended as best set of variables for predicting PWM. Here, we would also like to mention that slight decrease in test accuracy on adding an extra day for computing moving average is not a strange behaviour because biological cycle of a pathogen depends more on the weather conditions as compared to the exact date of a calendar month. Use of all variables as input to RJP identifies *H*811 and *T*814 as the most significant attributes. However, it has not resulted in improvement of the test accuracy over the pair of attributes.

In Fig. 3, we observe that for three algorithms namely LR, CJP and RJP, test accuracy is more than the training set accuracy. Although this behaviour is not commonly observed, yet it is not unusual. There have been a number of published reported results (Table 5) in the literature on different datasets using different models and algorithms where the training set accuracy is observed to be lesser as compared to the test set accuracy [Clark et al (1989), Mitra *et al.* (1997), Duch (2001)].

Table 5. Some reported cases where training set accuracy is less than test set accuracy

S.No.	Model	Training	Test	Dataset	References
1	NN	76.9	80.4	Hepatobiary disorder	Duch *et al.* 2001
2	AN	87.5	88.69	Vowel data	Mitra *et al.* 1997
3	SN	98.11	100	Hepato	Mitra *et al*. 1997
4	C4.5	89.0	89.8	Quadrant-200	Klaus *et al*. 1995
5	Default Rule	54	56	Lymphography	Clark *et al*. 1989
6	Default Rule	70	71	Breast Cancer	Clark *et al.* 1989
7	Default Rule	23	26	Primary Tumor	Clark *et al.* 1989

A special mention is also needed regarding 100 per cent accuracy of prediction in some cases (Fig. 3). It is emphasised that100 per cent accuracy for the training data may occur due to over fitting. But, whether the high accuracy over the training data holds good for future prediction will be substantiated if similar performance is observed for the test data as well. For example, in the present analysis, RS exhibits 100 per cent accuracy over the training data. But this can not be substantiated because RS performs badly over the test data. Thus, performance of RS on the training data is attributed to over fitting as is evident from its relatively worse performance on the test data.

As training set accuracy is not considerably important for the purpose of final comparison of algorithms, mean accuracies of the test data as obtained from all the algorithms is compared in Table 6. It is evident that test performance of CJP and RJP are comparable. Hence, among the algorithms CJP and RJP are recommended for prediction of PWM.

Table 6. Comparison of average of test accuracy (in per cent) of various algorithms

Variables used	LR	RS	CJP	RJP
Pairwise	75	74	83	84
All	*	62	74	74

Note: '*' indicates all variables were not used together in LR because of its limitation in handling all 8 variables together.

5.5 Differential Behaviour of Various Techniques and Contribution of the Study

The behaviour of the various techniques on training data and test data and their pair wise comparison can be explained by putting them into four categories (Table 7). For example, If we compare any two algorithms say A1 and A2 using the training and the test data, then the behaviour of the algorithm A1 over the algorithm A2 would belong to one of the four categories say 1, 2, 3 or 4 (see column Catg in Table 7). In this table, the entry 'worse' under the column 'Training Data' means that algorithm A1 (the first in the pair (A1,A2)) is shown to perform worse than the algorithm A2 for the training data. Similarly the entry 'better' under the column 'Test Data' means that algorithm A1 is shown to perform better than or equal to the algorithm A2 for the test data. The overall preference of the algorithm for the corresponding category is known by the comment on the overall performance of algorithm A1 over algorithm A2 (Table 7).

Table 7. The categories of different behaviour on pair wise comparison of prediction accuracy of an algorithm A_1 with algorithm A_2

(Algo A1, Algo A2)	Accuracy of algo A1 over algo A2 on		CatG	Comment on Performance of A1 with respect to A2
	Training data	Test data		
(LR, RJP)	worse	worse	1	worse
(LR, CJP), (RS, LR), (RS, CJP), (RS, RJP)	better	worse	2	worse
(LR, RS), (CJP, RS), (CJP, RJP), (RJP, RS)	worse	better	3	better
(CJP, LR), (RJP, LR), (RJP, CJP)	better	better	4	better

Category 1 includes the situation where algorithm A1 performs worse than the algorithm A2 over training as well as test data. Under this situation A2 is recommended over A1 for prediction. Comparison of LR with RJP denoted by (LR, RJP) belongs to this category (Table 7, Fig. 3). Thus, RJP emerges as the better performer than LR in this comparison.

Category 2 includes the behaviour that perform exclusively better for training data but contrastingly worse for the test data e.g. RS approach gives better accuracy over training but worse for test when compared with LR, CJP or RJP algorithms (Table 7, Fig.3,). The good performance over the training data is not important but test data performance is certainly important while comparing the algorithms. Hence in this category, algorithm 2 is considered better over algorithm 1. Comparisons of (RS, LR), (LR, CJP), (RS, CJP), (RS, RJP) fall under category 2. Here, it is observed that LR performs better than RS and CJP performs better than LR. This implies CJP is better than RS as well as LR. Further, (RS, RJP) implies that RJP is better than RS.

Category 3 includes the situation where algorithm A1 is worse than the algorithm A2 on training data but better than the algorithm A2 on the test data. In such cases algorithm A1 is to be selected for forewarning because they have shown better performance on the test data due to their least tendency towards over fitting during model learning. The cases of (LR, RS), (CJP, RS) and (CJP, RJP) and (RJP, RS) were observed to belong into this category (Table 7, Fig. 4). Here, the algorithms LR, CJP and RJP emerge better in pair wise comparison, but LR is being rejected in its comparison with other algorithms falling under category 4.

Category 4 includes the behaviour where the algorithm A1 performs better over algorithm A2 on training data as well as test data. Whenever any algorithm is able to achieve this, it means the model is perfect, model has truly captured the causing agents of the disease. Naturally, algorithm A1 is considered better over A2 in this category. The cases of (CJP, LR), (RJP, LR) and (RJP, CJP) were observed to belong into this category (Table 7, Fig.3). Consequently, CJP and RJP are recommended for prediction of the PWM disease.

Based on the discussion in this section, contributions of this study in predicting PWM disease are

1. CJP and RJP model are recommended for forewarning PWM because of better predicting accuracy over conventional method namely LR.
2. Temperature and humidity values pertaining to 8-11 days is found more appropriate for predicting PWM disease.
3. The underlying assumption regarding normal distribution of the values of the variables is not necessary to have better prediction.

4. The resulting model i.e. rules and DT are easy to interpret as well as easy to apply in comparison to classical method of LR.

6. CONCLUSIONS

Powdery Mildew of Mango (PWM) is a devastating disease and a prediction model to forewarn the epidemic outbreak of PWM using data from historical years is required. Predictive models are developed using the algorithms LR, RS, CJP and RJP by using different training-test pairs and attributes representing weather parameters. The results support the recommendation of CJP and RJP for prediction in crop diseases as it performs better than LR and RS in terms of performance parameters. The resulting models are easy to understand and implement without much technical expertise. The temperature and humidity variables relating to 8th-11th days of month of March are recommended for predicting PWM disease.

ACKNOWLEDGEMENTS

The authors are thankful to the referee for the very useful comments that helped in improving the quality of the manuscript.

REFERENCES

Clark, P., Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning,* **3(4),** 261-283.

Duch,W., Adamczak, R. and Grabczewski K. (2001). A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks,* **12(2),** 277-305.

Han, J. and Kamber, M. (2001). *Data Mining Concepts and Techniques.* Morgan Kaufmann Publisher.

Hastie, T., Tibshirani, R. and Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer.

Jain, R. and Minz, S. (2003). Should decision trees be learned using rough sets? In *Proc. 1st Indian International Conference on Artifical Intelligence* (IICAI-03), 1466-1479, Hyderabad.

Komorowski, J., Pawlak, Z., Polkowki, L. and Skowron, A. (1999). Rough Sets: A Tutorial. In : *Rough Fuzzy Hybridization, Pal S.K. and Skowron, A.(eds.),* Springer, 3-99.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technomettrics,* **19(2)**, 191-200.

Minz, S. and Jain, R. (2005). Refining decision tree classifiers using rough set tools. *Int. J. Hybrid Intell. Sys.,* **2(2),** 133-148.

Minz, S. and Jain, R. (2003a). Rough set based decision tree model for classification. *Proc. 5th International Conference on Data Warehousing and Knowledge Discovery, (DaWaK 2003) Prague,* Czech Republic, September 3-5, 2003, LNCS 2737, 172-181.

Minz, S. and Jain, R. (2003b). Hybridizing rough set framework for classification: An experimental view. In : *Design and Application of Hybrid Intelligent Systems, A. Abraham et al.* (eds.), IOS Press, 631-640.

Misra, A.K., Prakash, O. and Ramasubramanian V. (2004). Forewarning powdery mildew caused by oidium mangiferae in mango *(Mangifera Indica)* using logistic regression models. *Ind. J. Agric. Sci.,* **74(2)**, 84-87.

Mitra, S., De, R.K. and Pal, S.K. (1997). Knowledge-based fuzzy MLP for classification and rule generation. *IEEE Trans. Neural Networks,* **8(6)**, 1338-1350.

Pawlak, Z. (1991). *Rough Sets-Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Dordecht.

Pujari, A.K. (2000). *Data Mining Techniques.* Universities Press.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kauffman.

Rosetta, Rough set toolkit for analysis of data available at http://www.idi.ntnu.no/~aleks/rosetta/

Winston, P.H. (1992). *Artificial Intelligence.* Addison-Wesley.

Witten, I.H. and Frank E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann Publishers.

Monograph on α-designs

Rajender Parsad, V.K. Gupta, P.K. Batra, S.K. Satpati and P. Biswas
Indian Agricultural Statistics Research Institute, New Delhi

by
Kishore Sinha
Birsa Agricultural University, Ranchi

α-designs are resolvable incomplete block designs, introduced by Patterson and Williams (1976). Its parameters are $v = ks$, $b = r, s, k, r$ and s. The dual of these designs are also α-designs. a-designs are useful as field experiment with large number of varieties. Most of the times, it may not be feasible to run the entire experiment in one season or location. In resolvable block designs locations or seasons are taken care of by replications and variation within a location or season is taken care of by blocking.

The monograph contains a-arrays for generating a-designs along with the layout plan of a-designs with $6 \leq v \leq 150$, $3 \leq k \leq 10$, $2 \leq r \leq 5$.

Comparisons of a-designs are made with corresponding square lattice designs, rectangular lattice designs, resolvable PBIB (2) designs given in Clatworthy (1973) and the α-designs obtainable from arrays given by Patterson *et al*. (1978) and from dualization of these basic arrays. Quite a good number of the designs perform better.

This is a well prepared and useful monograph for variety trials where the experiment is to be spread at different locations or seasons. I must congratulate the authors for the commendable work. I am tempted to anticipate a similar endeavour on factorial experiments.

However, I would like to mention that if and when a revision of the monograph is made, α-designs with (i) $k = 2$ (ii) 6 £ r £ 10, and an example showing analysis of data obtained from an experiment conducted in a-designs may be considered for inclusion.

REFERENCES

Patterson, H.D. and Williams, E. R. (1976). A new class of resolvable incomplete block designs. *Biometrika*, **63(1),** 83-92.

Patterson, H.D., Williams, E.R. and Huneter, J.S. (1978). Block designs for variety trials. *J. Agric. Sci.*, **90,** 395-400.

Author Index Volume 63, No. 1, April 2009

भारतीय कृषि सांख्यिकी संस्था

(हिन्दी परिशिष्ट, इस अंक में प्रकाशित शोध–पत्रों के सारांश)

सुरेश चन्द्र राय

खंड 63 अप्रैल 2009 अंक 1

अनुक्रमणिका

आन्ध्र प्रदेश के जनपदों के मध्य सामाजिक-आर्थिक विकास में विविधता

प्रेम नारायण, एस.डी. शर्मा, एस.सी. राय एवं वी.के. भाटिया
भारतीय कृषि सांज़्यिकी अनुसंधान संस्थान, नई दिल्ली

आन्ध्र प्रदेश के विभिन्न जनपदों के सामाजिक-आर्थिक विकास का आकलन 50 संकेतकों के इष्टतम संयुज़्त सूचकों के आधार पर किया गया है। संकेतकों के आंकड़े 22 जिलों के लिए 2001-2002 वर्ष के लिए प्रयोग में लाए गए। विकास स्तर का आकलन कृषि क्षेत्र, में गुंटूर जिला प्रथम स्थान पर पाया गया। विभिन्न जिलों की अवसंरचना सुविधाएँ, कृषि तथा कुल सामाजिक-आर्थिक विकास को धनात्मक रूप से प्रभावित करती हैं तथा कृषि विकास, कुल सामाजिक-आर्थिक विकास से धनात्मक रूप से संबंधित पाया गया।

इष्टतम तथा इष्टमेतर दशाओं के अन्तर्गत पश्च-स्तरण में दक्ष आकलन

एम.सी. अग्रवाल एवं एस.सी. सेनापति*
दिल्ली विश्वविद्यालय, दिल्ली

स्थिर समष्टि में समष्टि योग तथा समष्टि माध्य के आकलन के लिए जैसा बसु (1971), स्मिथ (1976) तथा कई अन्य व्यज़्तियों द्वारा दर्शाया गया है, उसी प्रकार इस लेख में लेखकों द्वारा दक्ष अनभिनत पश्च-स्तरण पर आधारित आकलकों का वर्णन किया गया है। इष्टतम तथा इष्टमेतर दशाओं में प्रस्तावित आकलकों का समूह सामान्य पश्च-स्तरण आकलक तथा सामान्य सरल माध्य से अधिक दक्ष पाया गया। प्रस्तावित आकलकों के समूह का परीक्षण प्रतिबंधी यादृच्छिकीकरण अनुमिति की दृष्टि से किया गया है।

* रावेन्शा विश्वविद्यालय, कटक

रैंज़्ड सेट प्रतिचयन द्वारा मारगेनिस्टर्न प्रारूप द्विचरीय सुप्राचलिक बंटन के प्राचलों का आकलन

मनोज चाको एवं पी. यागीन थॉमस
केरल विश्वविद्यालय, तिरूअनन्तपुरम

रैंज़्ड सेट प्रतिचयन पद्धति का उपयोग वहाँ पर किया जाता है जहाँ प्रतिचयन इकाइयों की कोटि संज़्या निरीक्षण अथवा किसी चयनित सहायक चर के आधार पर हो सके। इस लेख में अध्ययन चर Y से सज़्बद्ध प्राचलों के विभिन्न आकलकों को दर्शाया गया है जो सहायक चर X के रैंज़्ड सेट प्रतिदर्श पर आधारित है। यहाँ पर चर X चर Y से सहसंबंधित है और (X, Y) प्रतिदर्श मारगेनिस्टर्न प्रारूप द्विचरीय सुप्राचलिक बंटन में है। इस लेख में प्रतिपादित सिद्धान्त आँकड़ों के द्वारा समझाया गया है। आकलकों की दक्षता की परस्पर तुलना भी की गई है।

एक चक्रीय तथा बहु-चक्रीय मत्स्य पालन पद्धति में टोर पुटीटोरा (हैमिल्टन) की लज़्बाई एवं भार के संबंध तथा उनकी वृद्धि - एक वस्तुस्थिति अध्ययन

एन. ओकेन्द्रो सिंह, मो. वसी आलम*, अमृत कुमार पाल*, एवं सुरेन्द्र कुमार**
शीतजल मात्स्यिकी अनुसंधान निदेशालय, भीमताल, नैनीताल

टोर पुटीटोरा जो शीतजल की एक मत्स्य प्रजाति है, उनकी संज़्या विगत कुछ वर्षों में अत्यन्त कम हो गई है तथा यह बहुमुखी खतरों की द्योतक है। इस अन्वेषण में एक चक्रीय बहुचक्रीय मत्स्य पालन पद्धति के अन्तर्गत टोर-पुटीटोरा प्रजाति की मछलियों की लज़्बाई तथा भार के संबंधों और उनकी वृद्धि के आकलन की एक विधि विकसित की गई है जिसका प्रयोग मछलियों के मूल्यांकन के लिए किया जाता है। उनकी वृद्धि के आकलन के संबंध में वान-वेर्टालोफी निदर्श सवोज़्म पाया गया।

* भा०कृ०सां०अ०सं०, नई दिल्ली

** कुमाँऊ विश्वविद्यालय, नैनीताल

फसलों में पोषक प्रबन्धन के लिए निर्णयाधार प्रणाली

एस. पाल, आई.सी. सेठी एवं अल्का अरोड़ा
भारतीय कृषि सांज़्यिकी अनुसंधान संस्थान, नई दिल्ली

फसल उत्पादन में वृद्धि तथा मृदा–उत्प्रवणता के लिए पोषक प्रबन्धन मुज़्य भूमिका निभाता है। इन बातों को ध्यान में रखकर भारतीय कृषि सांज़्यिकी अनुसंधान संस्थान ने एक निर्णयाधार प्रणाली (DSSNMC) का निर्माण तथा उसका विकास किया है। DSSNMC वेब पर आधारित होती है और यह कृषकों की फसलों के पोषक प्रबन्धन के लिए निर्णयाधार प्रणाली पर सलाह देती है। इस प्रणाली का महत्व बहुत अधिक है ज़्योंकि कृषि विशेषज्ञ सदैव उपलज़्ध नहीं होते। DSSNMC तीन विभिन्न दशाओं में सलाह देते हैं। प्रथम मॉड्यूल मृदा परीक्षण पर आधारित होता है। जहाँ मृदा परीक्षण नहीं होता वहाँ द्वितीय और तृतीय मॉड्यूल का उपयोग किया जाता है जो कृषि भूमि की स्थिति (ज़िला अथवा क्षेत्र) पर आधारित होता है।

फ़सल बीमारियों की अग्रिम चेतावनी के लिए यन्त्र शिक्षण

रजनी जैन, सोनाझरिया मिन्ज़* एवं रामासुब्रमनियन वी.**
राष्ट्रीय कृषि आर्थिकी एवं नीति अनुसंधान केन्द्र, नई दिल्ली

संगणक के उपयोग में लाने के साथ फसल बीमारियों की अग्रिम चेतावनी का महत्व बहुत बढ़ गया है। फसल की बीमारियों की समय से अग्रिम चेतावनी न केवल उपज की क्षति को कम करेगी बल्कि इसे प्रभावशाली ढंग से रोकने में सहायक होगी। अग्रिम चेतावनी के लिए परज़्परागत रूप से लॉजिस्टिक समाश्रयण (LR) तथा विविज़कर विश्लेषण पद्धतियों का प्रयोग किया जाता था। अभी हाल में यन्त्र शिक्षण (Machine Learning) तकनीकियों जैसे – DT, RS आदि अनेक नवीन पद्धतियों का प्रयोग होना प्रारंभ हो गया है। इस लेख में यन्त्र शिक्षण की तीन प्रमुख पद्धतियों की क्षमता का आकलन किया गया है। उदाहरणस्वरूप आम में अग्रिम चेतावनी के लिए यथोचित निदर्श का निर्माण तापक्रम तथा आद्रता के आधार पर किया गया है।

* जवाहरलाल नेहरू विश्वविद्यालय, नई दिल्ली

** भारतीय कृषि सांज़्यिकी अनुसंधान संस्थान, नई दिल्ली