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.

$$y = b_0 + b_1 x + u$$

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.