Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 65(1) 2011 59-68

64th Annual Conference of Indian Society of Agricultural Statistics
Theme : Statistics and Informatics for Massive Data Sets

Sub-Theme 1
Dimension Reduction and Classification Procedures

Chairman : Kashinath Chatterjee, Visva Bharati University, Santiniketan
Conveners : A. Dhandapani, NAARM, Hyderabad
Alka Arora, IASRI, New Delhi
C. Pal, University of Kalyani, Kalyani

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

1. Asis Kr. Chattopadhyay, Calcutta University, Kolkata. Integrated Component Analysis and Robust Clustering related to Astronomical Data.
2. Rajni Jain, National Center for Agricultural Economics and Policy Research (NCAP), New Delhi. Dimensionality Reduction for Classification using Rough Sets
3. S. Ravichandran, Directorate of Rice Research, Hyderabad. Artificial Neural Network (ANN) Modelling - Status and Scope
4. Ranjana Agrawal, Indian Agricultural Statistics Research Institute (IASRI), New Delhi. Use of Data Reduction Techniques in Crop Yield Forecasts
5. Alka Arora, Indian Agricultural Statistics Research Institute (IASRI), New Delhi. Blending Rough Sets and Clustering Methods for Discovery of Patterns

Sub-Theme 2
Data Management and Statistical Analysis in Agriculture and Allied Fields

Chairman : A.K. Nigam, IASDS, Bangalore
Conveners : Rajender Parsad, IASRI, New Delhi
B.M.K. Raju, CRIDA, Hyderabad
P.K. Sahu, BCKV, Mohanpur

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

1. G.M. Saha, ISI, Kolkota. Data Management and Statistical Analysis in Agriculture and Allied Fields
2. A. Dhandapani, NAARM, Hyderabad. Data Management Issues in Developing Information Systems
3. J. Jayasankar, CMFRI, Kochi. National Marine Living Resources Data Centre & Data Analytics at CMFRI
4. Rajender Parsad, IASRI, New Delhi. Strengthening Statistical Computing for NARS

Sub-Theme 3
Estimation Problems in Multidimensional Data Sets: Challenges Ahead

Chairman : D.K. Jain, NDRI, Karnal
Conveners : U.C. Sud, IASRI, New Delhi
J. Jayasankar, CMFRI, Kochi
D. Mazumdar, BCKV, Mohanpur

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

1. Anup Dewanji, ISI, Kolkata. Non Parametric Estimation of Quality Adjusted Lifetime (QAL) Distribution in Some Illness Death Models
2. Amrender Kumar, IASRI, New Delhi. Application of Soft Computing Techniques in Identifying Relationship for Multidimensional Datasets
3. T.V. Sathianandan, CMFRI, Kochi. Estimation of Marine Fish Landings

After detailed discussions in the above themes, the following recommendations emerged out:

1. Integrated Component Analysis (ICA) method is a promising alternative way of dimension reduction for non-normal populations. Its use may be explored in analyzing massive data matrices in agricultural systems research with random cell observations.
2. Efforts may be made to explore the use of statistical data cloning in estimation problems of massive data sets.
3. Research in development of statistical techniques for Estimation Problems in Multidimensional Data Sets need to be undertaken.
4. Rigorous efforts may be made for sensitization of research personnel of NARS in the high end statistical computing and customized modules for online analysis need to be developed.
5. Use of Rough sets based tools such as reducts, core and approximate core may be explored for dimension reduction problems in large datasets.
6. Concerted efforts may be made to harness the symbiosis of forecast models and Artificial Neural Network (ANN) for providing reliable forecasts.
7. An important and integral component of databases is the generation of quality data. Sophisticated and efficient statistical techniques should be developed/ employed for generation of data so as to enable drawing appropriate and valid inferences.
8. Quality checks should be ensured while designing information systems for complex agricultural problems and PDA for data collection to provide reliable, timely and quality data.
9. Fisheries statistics being collected using scientific methodology should be utilized by the agencies reporting fisheries statistics.
10. An active collaboration between CMFRI and IASRI would be immensely useful in addressing statistical and data mining issues in fisheries research.
11. Efforts may be made to adapt the concept of Quality Adjusted Lifetime (QAL) in predicting the crop and pest life stages during critical phases of vulnerability.
12. Soft computing techniques may be developed/ utilized for drawing maximum information available in the research domain/databases to strengthen predictive modeling faculties in NARS.
13. Proper statistical techniques need to be used for evaluation of genetically modified crops, estimation of poverty, micro level planning impact assessment studies, etc. for ensuring inclusive growth.
14. Stochastic Programming Formulations may be employed for multi-response optimization in multi-response experiments.

ABSTRACTS

Sub-theme 1 : Dimension-reduction and Classification Procedures

1. Independent Component Analysis and Robust Clustering related to Astronomical Data

Asis Kumar Chattopadhyay

Independent Component Analysis (ICA) is closely related to Principal Component Analysis (PCA). Whereas ICA finds a set of source data that are mutually independent, PCA finds a set of data that are mutually uncorrelated. The assumption that data from different physical processes are uncorrelated does not always imply the reverse case that uncorrelated data are coming from different physical processes. This is because lack of correlation is a weaker property than independence.

In the present case an objective classification of the globular clusters of NGC 5128 has been carried out. Components responsible for significant variation have been obtained through both Principal Component Analysis (PCA) and Independent Component Analysis (ICA) and the classification has been done by K-Means clustering. The set of observable parameter includes structural parameters, spectroscopically determined Lick indices and radial velocities from the literature.

A robust clustering technique applicable to very large data sets has also been discussed.

2. Dimensionality Reduction for Classification using Rough Sets

Rajni Jain

Classification is an important research topic in the field of data mining and knowledge discovery. There have been many data classification methods including decision tree methods, statistical methods, neural networks, rough sets, etc. One of the main obstacles while doing classification is dataset dimensionality. Dimensionality reduction methods try to find a reduced number of dimensions to account for the original data. Principal Component Analysis is the best known of these techniques. However, it offers some limitations to the use of induced model for interpretation and understanding. Recently, there has been rapid growth of interest in rough set theory and its applications for classification and feature selection.

Rough set theory was developed by Zdzislaw Pawlak in the early 1980's [9]. The main goal of the rough set analysis is to synthesize approximation of concepts using the acquired data. Rough set theory is based on the concept of indiscernibility relation and offers a simplified search for dominant attributes called reducts, in datasets. Reducts are the optimal number of attributes which preserve the indiscernibility between the objects in the dataset. A dataset may have zero, one or multiple reducts. An approximate core is proposed as an important tool to deal with the datasets which are having multiple reducts.

The Forest cover type data from the UCI repository is one the large datasets containing 581012 examples with 54 attributes representing 12 features. The classification task is to predict the Forest cover type (7 classes) given only cartographic data. The performance parameters - accuracy, complexity, number of rules and number of attributes in the resulting classifiers are compared to identify a suitable set of attributes for classification. The results using approximate core are comparable with the other published results for this dataset.

Validity of approximate core on Forest cover type dataset suggests that the tool is useful for classification problems involving real time large datasets. In future, there is a need for the software for computation of approximate core and its integration with classification algorithms to encourage the application of approximate core in real time problems.

3. Artificial Neural Network (ANN) Modeling - Status and Scope

S. Ravichandran

Rice production in India is an important part of the national economy. India is the world's second largest producer of white rice, accounting for 80% of all world rice production. Rice is India's preeminent crop, and is the staple food of the people of the eastern and southern parts of the country. Modelling and forecasting all-India rice production is carried out by utilising data on all-India rice area, production and yield for the period 1950-51 to 2008-09 along with all-India rainfall data from June to September for the corresponding period using various time series modelling methodologies such as Autoregressive Integrated Moving Average (ARIMA) and Artificial

Neural Network (ANN). Autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). They are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to remove the non-stationarity where as ANN is an information processing system that roughly replicates the behaviour of a human brain by emulating the operations and connectivity of biological neurons. ANN modelling methodology performs better for modelling and forecasting time-series data. Modelling rice production data is carried out using ANN methodology by making use of time series production data of the country as a whole. Time series data is obtained from the Directorate of Economic and Statistics, Government of India along with rainfall data obtained from Indian Meteorological Department (IMD). Developed model is utilized for forecasting kharif 2010 rice production. Modelling and forecasting data on all-India rice production is carried out by making time-series data on all-India rice area, production and yield during 1950-51 to 2009-10 along with all-India rainfall data for the corresponding period including the rainfall received by the country during July 2010. India received good monsoon rainfall in 2010. As per Indian Meteorological Department, monsoon during 2010 is normal. Based on the data on rice production and rainfall for the period 1950-51 to 2009-10, ANN models were developed using various ANN algorithms by making use of 70% of the data for training the model, 20% for testing and remaining 10% of data is utilised for validating the developed model. Forecasting was carried out by making use of the most efficient model which satisfied the goodness of fit of statistical modelling. Based on the best developed ANN model, prediction of rice production for the year 2010-11 was worked out separately for kharif and rabi. Rice production for 2010-11 would be 98.7 million tons out of which kharif rice production would be around 80 million tons.

4. Use of Data Reduction Techniques in Crop Yield Forecasts

Ranjana Agrawal and Chandrahas

Weather based modeling is one of the major approaches for forecasting crop yields. The approach utilizes data on various weather variables affecting crop yield. Weather affects crop differently during different stages of crop growth. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into fine intervals. This will increase number of variables in the model and in turn a large number of parameters will have to be evaluated from the data. This will require a long series of data for precise estimation of the parameters which may not be available in practice. Thus, a suitable data reduction technique is required which converts weather variables in different periods during the crop season to relatively smaller number of manageable variables which could be used in the model. Four approaches have been studied for the purpose viz. weighted weather indices, discriminant function, principal component analysis and factor analysis using data of wheat yield in various districts of Uttar Pradesh. Results revealed that principal component analysis and factor analysis were not found appropriate in most of the cases. Out of the remaining two, in some cases discriminant function approach performed better than weather indices based approach whereas in some cases, reverse trend was observed. As such, out of these two, none of the method revealed uniform superiority over the other.

5. Blending Rough Sets and Clustering Methods for Discovery of Patterns

Alka Arora, Rajni Jain and Shuchita Upadhyaya

Clustering is an exploratory data mining technique which deals with grouping of objects, such that objects within a single cluster have similar characteristics, hence objects in different clusters are dissimilar. From data mining perspective, the underlying objective of applying clustering technique on dataset is to discover the concept and patterns within the data which can be revealed by grouping the objects into clusters. Wealth of clustering algorithms is available in literature, but majority of them lacks in producing cluster description in the form of pattern. At times post processing of obtained clusters is essential in order to understand the pattern of obtained clusters. Approach presented in this paper is based on integration of Reduct from Rough Set Theory (RST) with obtained clusters and results in pattern extraction for better understanding of clusters. Approach is broadly categorized into four steps.

Step 1 : Partitioning clustering algorithm is applied on dataset in order to obtain non overlapping distinct clusters.

Step 2 : Clustering algorithm is intended to form clusters having most attribute values common to their members (cohesion) and few values common to members of other clusters. Intuitively, attributes which have similar value for majority of objects in the cluster are considered significant and rest are non significant in generating pattern for that cluster. Post processing of individual clusters is carried out using reduct from RST. Computation of reduct on individual clusters provides the set of attributes which distinguishes objects in a cluster and is considered non significant in generating pattern for that cluster.

Step 3 : Non significant (reduct) attributes are removed. Then in order to generate concise cluster pattern, descriptors (attribute value pair) are evaluated on Precision Coverage Coefficient (PCC) score which is formed using the generalized concept of PE and Coverage Ratio in Cluster for descriptors.

$$PE(a = v) = \frac{Support_U\ (a = v) - Support_{C_i}(a = v)}{card\ U - card\ C_i}$$

$$CRC(a = v) = \frac{card_{Ci}(a = v)}{card\ C_i}$$

$Support_{C_i}\ (a = v) = card_{C_i}\ (a = v)$ = number of objects satisfying $(a = v)$ in clusterC_i.

$Support_U\ (a = v)$ = number of objects satisfying $(a = v)$ in universe set.

$$PCC(a = v) = \sqrt{(1 - PE)(a = v)) * CRC(a = v)}$$

PCC score is a real number in the interval [0, 1] and reflects the significance of descriptor in a cluster. If PCC score is 1, then single descriptor is sufficient to describe the cluster. User can select the descriptors with greater PCC threshold λ (user defined threshold), for pattern formulation.

Step 4 : If PCC score is less than 1 then concatenate the descriptor with next descriptor from array. Every descriptor is added to the pattern only if there is decrease in the value of PE; Evaluate the pattern on PE and CRC; If PE and CRC are in specified threshold limits then output pattern. This Procedure needs to be repeated with every descriptor in the array. Keep on concatenating the descriptors in descending order of their PCC score to obtain multiple pattern satisfying threshold criteria of PE and CRC. Carry out the process of concatenation of descriptors till array is exhausted or desired numbers of patterns are obtained.

Sub-theme 2 : Data Management and Statistical Analysis in Agriculture and Allied Fields

1. Data Management and Statistical Analysis in Agriculture and Allied Fields

G.M. Saha

The increasing complexity of agricultural data and their management in finding solutions for certain farmers' problems requires adequate attention and efforts of research workers in these areas to develop appropriate tools. Fortunately, developments in computer technology continually expand the possibilities in agricultural data analysis and processing. Some researchers, therefore, tackle some of the important issues in agricultural data management and processing, using an integration of statistical tools and qualitative modeling techniques, in order to describe the complex structure of agricultural processes running in specific situations. Specific methodologies are, in fact, needed to make more efficient use of data collected by providing a means of effectively analyzing the data.

Analysis of data may be described as a process of inspecting, cleaning, transforming and modeling data with the goal of highlighting useful information, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science and social science domains.

In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA) and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical or structural models for predictive forecasting or classification, while text analytics applies statistical, linguistic and structural techniques to extract and classify information from textual sources. Data integration is a precursor to data analysis and data analysis is closely linked to data visualization and data dissemination. The term data analysis is sometimes used as a synonym for data modeling.

Broadly, we have (1) Qualitative data analysis, and (2) Quantitative data analysis. The steps of data analysis consist of (a) Data cleaning, (b) initial data analysis, and (c) main or final data analysis. The initial data analysis has the following steps: Quality of data/ measurements, initial transformations if necessary, characteristics of data sample (descriptive statistics) and final stage of the initial data analysis consisting of actual statistical data analysis.

Qualitative research uses quantitative data analysis (QDA) to analyze text, interview transcripts, photographs, art, field notes of (ethnographic) observations, etc.

Data cleaning is an important procedure during which the data are inspected and erroneous data are, if necessary, preferably and possibly corrected. Data cleaning can be done during the stage of data entry. If this is done, it is important that no subjective decisions are made.

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that are aimed at answering the original research question. The initial data analysis phase is guided by the following three questions: Quality of data, Quality of measurements, and Initial Transformations.

The quality of the data should be checked as early as possible. Data quality can be assessed in several ways, using different types of analyses: frequency counts, descriptive statistics (mean, standard deviation, median). Normality (skewness, kurtosis, frequency histograms, normal probability plots), associations (correlations, scatter plots) etc.

Comparison and correction of differences in coding schemes: variables are compared with coding schemes of variables external to the data set and possibly corrected if coding schemes are not comparable. The choice of analyses to assess the data quality during the initial data analysis phase depends on the analyses that will be conducted in the main analysis phase.

The quality of the measurement instruments should only be checked during the initial data analysis phase when this is not the focus or research question of the study. There are two ways to assess measurement quality (i) confirmatory factor analysis (ii) analysis of homogeneity (internal consistency).

After assessing the quality of the data and of the measurements, one might decide to impute missing data, or to perform initial transformations of one or more variables, although this can also be done during the main analysis phase.

Characteristics of data sample can be assessed by looking at (i) basic statistics of important variables (ii) scatter plots (iii) correlations and (iv) cross-tabulations.

During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken. The original plan for the main data analyses can and should be specified in more detail and/or rewritten. In order to do this, several decisions about the main data analyses can and should be made.

2. Database Management Issues in Developing Information Systems

A. Dhandapani

In this talk, issues relating to data management that arise while developing information systems are discussed. The discussion revolves around the author’s experience while developing “e-Pest Surveillence” Information Systems at National Centre for Integrated Pest Management, IARI, New Delhi.

Agriculturally important pests (insects, diseases, nematodes, weeds) cause huge loss every year and the need for effective pest surveillance was long felt one and various efforts were initiated with varying amount of success. However, pest out breaks in cotton growing areas of Punjab, Haryana and Rajasthan has resulted in creation of nation-wide pest surveillance system with emphasis on regular monitoring, reporting using scientifically designed surveys coupled with Information Communication Technologies.

The e-Pest surveillance system developed consists of (i) Use of hand-held devices with a customized application written in Windows CE for data entry; (ii) an intermediate desktop application for converting the data entered in the hand-held device into XML file; (iii) a web-interface for authorized users to upload the XML file generated by the hand-held device and generation of reports (Village-wise/District-wise) etc.; and (iv) District-wise Pest incidence maps using GIS. The author was involved in developing the web-interface of e-Pest surveillance system.

At the core of the web-interface is a relational database developed to store pest monitoring data designed and implemented in Microsoft® SQL Server 2000. Relational Database Management Systems (RDBMS) allow efficient way of retrieving the data. A windows service was designed to convert XML file into relevant tables within the database. This windows service watches a particular folder for any uploaded XML file at specified time interval and start moving the data in the XML file into corresponding table whenever a new XML file was found.

Converting raw data into useful reports involve developing relevant queries. As the type of data collected vary depending on the type of pest, it was necessary to develop various queries to compute different type of summary measures and collect them into a single report. Some of the summary measures developed include weighted averages, percentage besides usual averages. All the queries developed were parameterized queries and implemented in SQL Server as Stored Procedures so that they are cached and the performance could be improved. Stored Procedures are also useful in implementing business logic at a central place for easy maintenance.

Using the queries created, reports were created in ASP.NET. The interface provides a convenient way of getting relevant parameters such as District/Village, pest name(s), report period (starting and ending dates) etc. from the user and invokes stored procedures to produce formatted reports.

Summary reports of pest levels would be of limited use as they could be understood only by experts and could not be useful for advisory purposes. For converting the summary details into more readily useful forms, decision rules were implemented in the system. Some of the decision rules implemented include (i) Converting the summary information into 4 classes of pest level viz. absent; low; medium and high. The conversion depends on pest, (ii) Use of other information to increase or decrease the category of the pest depending on stage of the crop, presence/level of Natural enemies, weather conditions etc.

The major database management issues which came up during the development of e-pest surveillance issues include (i) Design issues – Making sure that the database design is same both in hand-held as well as centralized database; Identifying previously visited fields so that consistency is maintained (it may be possible that different hand-held device was used in the previous visit); (ii) Developing XML structure for data-exchange between hand-held device and centralized database, size of XML file; (iii) Creating SQL Stored Procedures in T-SQL – Optimizing Queries, implementing complex rules for decision making (iv) Maintaining Stored Procedures and Web-application.

The e-Pest surveillance system was successfully tested in 3 districts of Andhra Pradesh and it was demonstrated from uploading of XML file to report generation. The same database design with minor modifications is now in use at National Centre for Integrated Pest Management and was successfully used in Maharashtra during 2009-10 for monitoring pests of soybean and cotton and during the current year, rice pest monitoring has been taken up in different districts of Orissa.

3. National Marine Living Resources Data Centre (NMLRDC) and Resource Analytics

J. Jayasankar

Fish is one such resource which hoodwinks even the most seasoned of observers. Though some leverage can be gained while watching fish which are grown under controlled or quasi-controlled environments, marine fisheries have been often at their deceptive best when it comes to assessment. The time tested method of fish stock assessment which reveals the best of past and forewarns a bit of future, has a over bearing demand of torrential data input. As the adage goes- “Model is as good as the data...”, fish stock assessment extends the same to mystique levels adding to uncertainty even in which way one has to proceed. Hilburn (2006) puts the classical paradox in perspective when he states that the naive-most of all methods viz, growth rate as compared the previous year would be the best bet for a robust prediction of what is in store next. As is evident fishery resource assessment has the minimum requirement of a continuous, systematic and accessible collation of data on the resources.

Central Marine Fisheries Research Institute (CMFRI), functioning since 1947 and based at Kochi, Kerala, India, is a research institution of repute under the Indian Council of Agricultural Research (ICAR), has marine fish stock assessment as one of its major mandates and has attuned its research commitments concomitant with the growth of marine fisheries in India, which has lead to progressive refinement in stock assessment techniques adopted. Owing to the uniqueness of the Marine Capture Fisheries scenario witnessed off Indian waters, a strong data support system which is solidly interwoven with a statistical analytical paradigm has been foreseen by the founding fathers of CMFRI and a perpetual data deposition mechanism had been thrashed out as early as in the first decade of its inception.

Towards monitoring the exploited marine resources of India, CMFRI had venture into a twin pronged approach of collecting data as well as honing up the collection mechanism vis-a-vis methodology since 1950's. Starting from the first survey conducted along Malabar coast in 1950-51, CMFRI has been regularly conducting surveys while refining the sampling designs as and when required. Since early 60's a near permanent setup was established which had a mandate of collecting catch and effort details of marine resources on weekly basis. As a result in 1983, National Marine Living Resources Data Centre (NMLRDC) was established. The latter half of 1981 marked an epoch in the history of the Institute with the advent of electronic computational facilities and installation of Unix based multi-user and multi-tasking networked environment. The team of statisticians at the Institute prepared the grounds for a scalable electronic data repository by way of preparing codes for the species and gears with which the species are caught, which went a long way in unification of data pertaining to different regions. A sort of homogeneity which was hitherto missing was achieved and that ensured evenness in the reporting of catch and effort and comparison across geographic boundaries. A full fledged computer based analysis of estimates has been in vogue since 1989.

It can be stated that NMLRDC has been one such unique databases which bootstraps onto its own honing by the type of information processed using it at regular intervals. It showcases a system where research and information get wound into a mutually rewarding eternal spiral, which immensely benefits a state and a class of citizens for whom fisheries is a livelihood issue.

4. Strengthening Statistical Computing for NARS

Rajender Parsad

For providing a healthy and enabling statistical computing environment, a general purpose high end Statistical package has been procured with 151 licenses for perpetual use with 03 years updates and upgrades by 151 National Agricultural Research System (NARS) Organizations. It can be installed on multiple official machines both in standalone as well as intranet mode. The goal of the project is to provide research guidance in statistical computing and computational statistics so as to provide enabling statistical computing facilities to the researchers of NARS. The efforts would not merely be focused on an interface of statistics, computer science and numerical analysis, but it would also involve designing of intelligent algorithms for implementing statistical techniques particularly for analyzing massive data sets, simulation, bootstrap, etc.).

The availability of healthy statistical computing environment would enable the researchers in NARS to undertake probing, in-depth, appropriate, intractable analysis of data generated from agricultural research including those in advanced research areas like biotechnology, genomics, micro-arrays, forecasting, agricultural field experiments, surveys, microarrays, massive data sets such as climate change, biodiversity, market intelligence, etc. It would also facilitate data sharing over web and creation of analytics over the web useful for All India Co-ordinated Research Projects and other Network Projects of NARS.

For providing a service oriented computing, a Portal is being established which will be available to NARS users through IP Authentication at http://stat.iasri.res.in:8080/sscnarsportal. User name and password can be obtained from Nodal Officers available at www.iasri.res.in/sscnars by any researcher from Indian NARS. In the beginning, plan is to make available the analysis of data generated from any block design and split plot design on this portal.

This project has brought all 151 NARS organizations in a closed network. The training component of the project is also very exhaustive and targets at training 200 trainers and 1400 agricultural research scientists in the country in the usage of high-end statistical package. These would then train other agricultural research scientists. Such an effort would have a multiplier effect.

Sub-theme 3 : Estimation Problems in Multi-dimensional Data Sets: Challenges Ahead

1. Nonparametric Estimation of Quality Adjusted Lifetime (QAL) Distribution in Some Illness-Death Models

Biswabrata Pradhan, Anup Dewanji and Alok Goswami

The concept of Quality Adjusted Lifetime (QAL) was developed in the context, where patients may experience several health states which differ in their quality of life measured by a utility coefficient ranging from zero to unity. The QAL is defined as

$$Q = \int_O^T W(u)du$$

where Q denotes the QAL, T the lifetime and W(u) the utility coefficient (represents the ‘quality of life’) at time *u*.

The space of the health state is usually assumed to be discrete. Then the QAL reduces to a weighted sum of the time spent (sojourn time) in each health state.

While dealing with censored data, there is informative censoring when transformed into QAL scale (Gelber *et al.* 1989; Glasziou *et al.* 1990; Lin *et al.* 1997; Huang and Louis 1999, among many others).

That is, even if the original lifetime *T* and the censoring time *C* are independent, *A* and the corresponding quality adjusted censoring time do not remain independent. This, in the literature, is known as induced dependent censoring.

Although it might seem natural to undertake a standard survival analysis with the observed QAL values (censored and uncensored), this approach leads to bias due to this induced dependent censoring.

Glasziou *et al.* (1990, StatMed) suggested partitioned survival analysis for progressive state models to estimate mean QAL restricted to a certain time, which is usually determined by follow-up time of clinical trials.

Zhao and Tsiatis (2000, Sank); Huang and Louis (1999, LDA); Chen and Sen (2001, Bmtc; 2004; CSTM) developed nonparametric techniques for estimating mean QAL.

Zhao and Tsiatis (1997, Bmtk; 1999, Bmtc); Huang and Louis (1998, Bmtk); Van der Laan and Hubbard (1999; Bmtc) developed nonparametric techniques for estimating the distribution of QAL.

Drawbacks

Note that all the previous work first convert the lifetime data into QAL scale and then adjust the estimation to account for the bias due to dependent censoring in QAL scale.

Not applicable when some of the transition times are not observed. Monotonicity is not guaranteed. The existing methods do not consider the structure of the illness-death models involving different health states and the relationship between the different sojourn times, making the estimates less efficient.

First derive the theoretical expression for the distribution of QAL as a function of the joint distribution of the sojourn times in all the states. Estimate the relevant distributions from the observed lifetime (or, sojourn time) data in each state. This is possible even when some of the transition times are not observed. Substitute the estimated distributions in the theoretical expression for QAL distribution (Pradhan and Dewanji, 2009, Stat Med; Pradhan, Dewanji and Sengupta, 2010, CSTM). This approach takes care of the bias since the estimation is carried out without transition forming into QAL scale. It also has some other advantages.

We investigate how the different estimates of QAL distribution behave with respect to bias and precision. We consider the proposed nonparametric (NP) estimate, Kaplan-Meier (KM) estimate and Zhao-Tsiatis (ZT) estimate. The consistency of the NP and ZT estimates is evident with increasing sample size. The standard error decreases with sample size. In contrast, the KM estimate is biased. The NP estimate performs better than

ZT estimate, especially when sample size is small (n = 50).

We consider different dependent models to describe the dependence between T_0 and T_{12}.

1. Semi-parametric Dependent Model
2. Markov Dependence
3. Arbitrary Dependence

Derivation of distribution of QAL can be in closed form for many situations, especially when the number of states are few. Estimation of QAL distribution is simple and natural leading to monotonic estimates. The structure of the illness-death model involving different health states and the relationship between the different sojourn times are explicitly used in the derivation of QAL distribution, making the estimate more efficient. This method can deal with some missingness of transition times. Incorporates dependence between different sojourn times. Incorporation of covariate effect is simple.

2. Application of Soft Computing Technique in Identifying Relationship in Multidimensional Datasets

Amrender Kumar and Ratna Raj Laxmi

Multi-dimensional data set is highly skewed and non-normally distributed. Recent developments in the field of non-parametric statistical analysis establishes that the soft computing technique which combines the ability of fuzzy approach to represent and manage imprecise data and knowledge together with the accepted capacities for learning and heuristic computation of neural networks in identifying complicated relationships in multidimensional datasets, without making a priori assumptions regarding the nature of these relationships. The uses of soft computing technique for identifying the relationship in multidimensional data sets are examined.

3. Estimation of Marine Fish Landings in India

T.V. Sathianandan

The main land of India has a total coast line of about 8129 km distributed along 9 maritime states and union territories of Pondicherry, Daman and Diu. As per 2005 marine fisheries census conducted by the Central Maine Fisheries Research Institute (CMFRI) there are about 3000 marine fishing villages and about 1400 marine fish landings centres along the coast line. More than 2000 species of marine living organisms belonging to 83 commercially important species groups land in the landing centres and fisheries harbours through out the year by large number of crafts and gears. Estimation of species wise, gear wise and region wise marine fish landings is a challenging task.

Monitoring and assessment of exploited marine fishery resources of the country is one of the important mandates of CMFRI. Catch and effort statistics along with biological data on fish caught by various gears form the basis of fish stock assessment. Thus, marine fish landings are estimated by CMFRI from commercial landings through sampling. The sampling methodology evolved over time and used by CMFRI for estimation of marine fish landings is a stratified multi-stage random sampling design. Vast experience that the Institute have in collection of marine fish catch statistics from 1950 onwards and the results of different pilot surveys conducted by the Indian Council of Agricultural Research on different occasions resulted in the development of the sampling design.

In the sampling design, stratification is done over space and time. Over space each maritime state is divided into suitable number of non-overlapping fishing zones on the basis of intensity of fishing and also based on geographical considerations. To maintain homogeneity within strata the fishing zones are further divided into sub-strata for some of the fishing zones. Stratification over time is a calendar month.

The collection of data from the landings centres adopting the sampling design is carried out by a group of 80 well qualified, trained and dedicated filed staff posted in 25 stations located along the coastal belt. The field staffs are trained in collection of data on catches and can identify fished organisms to the species level. Out of the 25 stations 10 are research/regional stations with full scientific and infrastructure backup. The overall operation of this uninterrupted exercise is coordinated and monitored by Fishery Resources Assessment Division of CMFRI with the support of the 10 research/regional centres. Information on species landed, quantity landed, crafts and gears used, effort in hours, distance from lading centre, wind direction etc are collected using different schedules prepared for the survey. The data is centrally processed at the headquarters and the estimates are made and added to a database.