उच्च संकाय प्रशिक्षण केंद्र
Centre for Advanced Faculty Training

संदर्भ पुस्तिका-II
Reference Manual-II

कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें

(कृषि शिक्षा विभाग, भा.कृ.अनु.प. द्वारा प्रायोजित)

MODERN DATA ANALYTICS TECHNIQUES FOR AGRICULTURAL SURVEYS

(Sponsored by Agricultural Education Division, ICAR)

फरवरी 11- मार्च 03, 2025

February 11- March 03, 2025

पाठ्यक्रम समन्वयक : कौस्तब आदित्य
पाठ्यक्रम सह-समन्वयक : अंकुर बिश्वास
पाठ्यक्रम सह-समन्वयक : पंकज दास

Course Coordinator : Kaustav Aditya
Course Co-Coordinator : Ankur Biswas
Course Co-Coordinator : Pankaj Das

प्रतिदर्श सर्वेक्षण प्रभाग
भा.कृ.अनु.प. - भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, पूसा, नई दिल्ली-110012

DIVISION OF SAMPLE SURVEYS
ICAR-INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
LIBRARY AVENUE, PUSA, NEW DELHI-110012
https://iasri.icar.gov.in
2025

प्राक्कथन

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान (भा.कृ.सां.अ.सं.) ने 1930 में तत्कालीन इंपीरियल कृषि अनुसंधान परिषद में एक सांख्यिकी अनुभाग के रूप में अपनी यात्रा शुरू की और सांख्यिकी विज्ञान (सांख्यिकी, संगणक अनुप्रयोग और जैव सूचना विज्ञान) के क्षेत्र में अनुसंधान, शिक्षा और प्रशिक्षण आयोजित करने के लिए प्रासंगिकता के एक प्रमुख संस्थान के रूप में विकसित हुआ है। संस्थान मुख्य रूप से मौजूदा ज्ञान में अंतराल को पाटने के लिए कृषि सांख्यिकी और सूचना विज्ञान में अनुसंधान करने के लिए जिम्मेदार है। संस्थान, सांख्यिकी विज्ञान को, सूचना विज्ञान के साथ मिश्रित करके, कृषि विज्ञान में उनके विवेकपूर्ण सम्मिश्रण का उपयोग कर रहा है, ताकि कृषि के नए उभरते क्षेत्रों की चुनौतियों का सामना कर गुणवत्तापूर्ण कृषि अनुसंधान को बढ़ाया जा सके, और साक्ष्य आधारित नीति निर्णय लिए जा सके। संस्थान ग्रेजुएट स्कूल, भा.कृ.अनु.प.–भारतीय कृषि अनुसंधान संस्थान, नई दिल्ली के सहयोग से कृषि सांख्यिकी, संगणक अनुप्रयोग और जैव सूचना विज्ञान में एम.एस.सी. और पीएच.डी. डिग्री कार्यक्रम भी संचालित करता है। संस्थान राष्ट्रीय कृषि अनुसंधान एवं शिक्षा प्रणाली (एनएआरईएस) को सुदृढ़ बनाने के लिए सलाहकारी एवं परामर्श सेवाएं प्रदान करता है तथा राष्ट्रीय एवं अंतर्राष्ट्रीय संगठनों के लिए प्रायोजित अनुसंधान एवं परामर्श सेवाएं प्रदान करता है। राष्ट्रीय कृषि सांख्यिकी प्रणाली (एनएएसएस) को सुदृढ़ बनाने में पद्धतिगत सहायता भी प्रदान की जाती है। संस्थान एनएआरईएस के लिए मजबूत कृषि ज्ञान प्रबंधन प्रणालियों और कृत्रिम बुद्धिमत्ता आधारित अनुप्रयोगों के विकास में भी अग्रणी भूमिका निभा रहा है।

सांख्यिकीय रूप से मान्य एवं अर्थपूर्ण निष्कर्ष, जो अनुसंधान कार्यक्रमों से प्राप्त होते हैं, उच्च–गुणवत्ता वाले अनुसंधान का आधार बनते हैं एवं नीतिगत योजना, विशेष रूप से विकास संबंधी पहल एवं कार्यक्रम कार्यान्वयन के लिए अत्यंत महत्वपूर्ण होते हैं। इसलिए, डेटा संग्रह एवं विश्लेषण के लिए मजबूत सांख्यिकीय विधियों का उपयोग करना आवश्यक है। संस्थान द्वारा प्रदान किए जाने वाले प्रशिक्षण कार्यक्रम शोधकर्ताओं एवं योजनाकारों को नवीनतम सांख्यिकीय तकनीकों से परिचित कराने में अमूल्य सिद्ध होते हैं।

संस्थान कृषि सांख्यिकी एवं कंप्यूटर अनुप्रयोग में उन्नत संकाय प्रशिक्षण केंद्र (सीएएफटी) भी है। 11 फरवरी से 3 मार्च 2025 के दौरान आयोजित "कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें" नामक यह प्रशिक्षण कार्यक्रम कृषि शिक्षा विभाग, भा.कृ.अनु.प. द्वारा प्रायोजित है। इस प्रशिक्षण कार्यक्रम को संकाय सदस्यों/वैज्ञानिकों को विभिन्न प्रतिदर्श पद्धतियों की जानकारी प्रदान करने के उद्देश्य से तैयार किया गया है, जिसमें नवीनतम विकास एवं सर्वेक्षण डेटा विश्लेषण के लिए सॉफ्टवेयर पैकेजों के उपयोग पर विशेष ध्यान दिया गया है। साथ ही, भारत में कृषि एवं बागवानी सांख्यिकी के संग्रह प्रणाली की भी जानकारी दी जाएगी। इसके अतिरिक्त, फसल उत्पादन अनुमान/पूर्वानुमान के लिए एआई/एमएल, भा.कृ.अनु.प. के दृष्टिकोण से डिजिटल कृषि से संबंधित कुछ महत्वपूर्ण विषयों आदि को भी शामिल किया गया है। प्रशिक्षण कार्यक्रम व्यावहारिक रूप से उन्मुख है जिसमें अनुभव एवं क्षेत्र के दौरे पर जोर दिया गया है।

इस प्रशिक्षण कार्यक्रम के संकाय में वैज्ञानिकों एवं प्रतिष्ठित सांख्यिकीविदों को शामिल किया गया है, जो प्रतिदर्श सर्वेक्षण एवं संबंधित क्षेत्रों में व्यापक अनुभव रखते हैं। इस प्रशिक्षण कार्यक्रम की शुरुआत से पहले प्रकाशित एवं वितरित की जाने वाली संदर्भ प्रशिक्षण पुस्तिका प्रतिभागियों के कार्य कौशल को समृद्ध करने में बहुमूल्य ज्ञान प्रदान करेगी। यह आशा की जाती है कि इस प्रशिक्षण कार्यक्रम से प्राप्त अनुभव प्रतिभागियों को अपने कार्यस्थल पर इस ज्ञान का प्रभावी रूप से उपयोग करने में सक्षम बनाएगा। मैं इस संदर्भ प्रशिक्षण पुस्तिका को समय पर प्रस्तुत करने के लिए प्रभागाध्यक्ष (प्रतिदर्श सर्वेक्षण) एवं पाठ्यक्रम समन्वयकों को बधाई देता हूँ।

(राजेन्द्र प्रसाद)
निदेशक, भा.कृ.अनु.प.–भा.कृ.सां.अ.सं.

नई दिल्ली
11 फरवरी, 2025

FOREWORD

The ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI) began its journey as a Statistical Section in 1930 in the then Imperial Council of Agricultural Research. Over the years, it has evolved into a premier institute dedicated to research, education, and training in Statistical Sciences (Statistics, Computer Applications and Bioinformatics). ICAR-IASRI has played a key role in advancing research in Agricultural Statistics and Informatics, addressing knowledge gaps in these fields. The Institute offers M.Sc. and Ph.D. programs on Agricultural Statistics, Computer Applications and Agricultural Bioinformatics in collaboration with the Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi. Additionally, ICAR-IASRI provides customized and sponsored training courses at both national and international levels, aiming to be a center of excellence in Human Resource Development. The Institute also offers advisory and consultancy services to strengthen the National Agricultural Research and Education System (NARES), conducts sponsored research for national and international organizations, and supports the development of a robust Agricultural Knowledge Management System for NARES.

Statistically valid and meaningful inferences derived from research programmes form the basis of high-quality research and are crucial for policy planning, particularly in developmental initiatives and programme implementation. Consequently, it is vital to employ robust statistical methodologies for data collection and analysis. The training programs offered by the Institute are invaluable in familiarizing researchers and planners with the latest advancements in statistical techniques.

The Institute is also a Centre of Advanced Faculty Training in Agricultural Statistics and Computer Application. This training programme entitled "**Modern Data Analytics Techniques for Agricultural Surveys**" organized during February 11-March 3, 2025 is sponsored by Agricultural Education Division, ICAR. The training programme has been designed to provide exposure to faculty members/scientists on different sampling procedures with due emphasis on recent developments and use of software packages for survey data analysis as well as the system of collection of agricultural and horticultural statistics in India. In addition, some important topics related to AI/ML for crop yield estimation/forecasting in India, Digital agriculture - ICAR perspective etc. have also been included. The training programme is practical oriented with emphasis on hands on experience and field visits.

The faculty of this training programme comprises of scientists and eminent statisticians with vast experiences in the field of sample surveys and related areas. The training manual being brought out and distributed before the start of the training programme will provide a wealth of knowledge to the participants in enriching their work capabilities. It is expected that the experience gained from this training programme will enable the participants to use this knowledge in their respective work place. I wish to compliment Head, Division of Sample Surveys and Course Coordinators for bringing out this valuable document in time.

11/2/2025

(Rajender Parsad)
Director, ICAR-IASRI

New Delhi
11 February, 2025

vkeq[k

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकीय अनुसंधान संस्थान (भा.कृ.अनु.प.–भा.कृ.सां.अ.सं.) कृषि सांख्यिकी, कृषि जैवसूचना एवं संगणक अनुप्रयोग के क्षेत्र में अनुसंधान को बढावा देने तथा संचालित करने के लिये एक प्रमुख संस्थान है। यह संस्थान, भारतीय कृषि अनुसंधान परिषद् (भा.कृ.अनु.प.) के मानव संसाधन विकास कार्यक्रम के तत्वाधान में कृषि सांख्यिकी एवं संगणक अनुप्रयोग में उच्च संकाय प्रशिक्षण केन्द्र के रूप में भी कार्यरत है। कृषि फसलों, बागवानी फसलों, पशुधन एवं मत्स्य पालन के मामलों में विभिन्न प्राचलों के आकलन से सम्बन्धित प्रतिदर्श सर्वेक्षणों सहित कृषि सांख्यिकीय के विभिन्न क्षेत्रों में मौलिक एवं व्यवहारिक, दोनों प्रकार के अनुसंधान किये जा रहे हैं। प्रतिदर्श सर्वेक्षण प्रभाग के वैज्ञानिक प्रतिदर्श सर्वेक्षण के विभिन्न पहलुओं जैसे जटिल सर्वेक्षणों की अभिकल्पना और विश्लेषण, सर्वेक्षण आँकड़ों के विश्लेषण हेतु सॉफ्टवेयर विकसित करना, बूटस्ट्रैप विचरण आकलन तकनीक, अंशांकन और मॉडल अंशांकन अनुमानक, लघु क्षेत्र आकलन, रैंक सेट प्रतिचयन, अनुकूली क्लस्टर प्रतिचयन, एकाधिक फ्रेम सर्वेक्षण, स्थानिक नमूनाकरण एवं आकलन, भौगोलिक रूप से भारित प्रतिगमन आधारित आकलन, कृषि सर्वेक्षणों में भौगोलिक सूचना प्रणाली एवं सुदूर संवेदी तकनीकों का अनुप्रयोग इत्यादि के अनुसंधान में लगे हुए हैं। यह प्रभाग प्रतिदर्श सर्वेक्षण के क्षेत्र में कई अनुप्रयुक्त अनुसंधान गतिविधियों के लिए संयुक्त राष्ट्र के खाद्य एवं कृषि संगठन (एफएओ) के साथ अंतरराष्ट्रीय सहयोग में भी प्रवृत्त है।

''कृषि सर्वेक्षण के डेटा विश्लेषण की आधुनिक तकनीकें'' नामक इस प्रशिक्षण कार्यक्रम का मुख्य उद्देश्य कृषि विज्ञान के विभिन्न विषयों से सम्बन्धित प्रतिभागियों को प्रतिचयन की विभिन्न तकनीकों एवं आकलन विधियों, प्रतिदर्श सर्वेक्षणों में नवीनतम विकास एवं प्रतिदर्श आँकड़ों के विश्लेषण में प्रयोग होने वाले सॉफ्टवेयर पैकेज जैसे MS-Excel, R, SAS, Python एवं SPSS के प्रयोग, कृषि सर्वेक्षणों में भौगोलिक सूचना प्रणाली एवं सुदूर संवेदी तकनीकों का अनुप्रयोग इत्यादि की जानकारी प्रदान करना है। सैद्धांतिक के अपेक्षा व्यावहारिक पहलुओं पर अधिक जोर दिया गया है। प्रतिभागियों के उपयोग के लिए संदर्भ पुस्तिका सरल रूप में प्रस्तुत की गयी है।

हम संस्थान एवं अतिथि संकाय के सभी संकाय सदस्यों का धन्यवाद करते है जिन्होंने इस कार्यक्रम को सार्थक एवं सफल बनाने में अपना बहुमुल्य समय लगा कर सहयोग दिया है। हम प्रशिक्षण कार्यक्रम के आयोजन के लिए विभिन्न व्यवस्थाएं करने के लिए शामिल विभिन्न समितियों के अध्यक्षों एवं सदस्यों के भी आभारी हैं। उनके अथक प्रयासो से इस संदर्भ पुस्तिका को समय से तैयार करने में मदद मिली है। हम इस प्रशिक्षण कार्यक्रम में प्रतिभागियों को नामित करने के लिए भारतीय कृषि अनुसंधान परिषद् के विभिन्न संस्थानों, राज्य कृषि विश्वविद्यालयों आदि के आभारी हैं। इस प्रशिक्षण कार्यक्रम के आयोजन का दायित्व हमें सौंपने के लिए हम भारतीय कृषि अनुसंधान परिषद् के शिक्षा प्रभाग के आभारी हैं। हम डॉ. राजेंद्र प्रसाद, निदेशक, भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान एवं डॉ. तौकीर अहमद, प्रभागाध्यक्ष, प्रतिदर्श सर्वेक्षण प्रभाग का इस कार्यक्रम में मार्गदर्शन एवं निरंतर सहयोगएवं प्रशिक्षण कार्यक्रम को सुचारू संचालन के लिए सभी आवश्यक सुविधाएं उपलब्ध कराने के लिए आभारी हैं। अंत में, हम उन सभी का आभार प्रकट करते हैं जिन्होंने, इस संदर्भ पुस्तिका को तैयार करने में सहयोग दिया है।

ub fnYyh **y[kdx.k**
11 Qjojh] 2025

PREFACE

The ICAR-Indian Agricultural Statistics Research Institute, New Delhi is a premier Institute for promoting and conducting research in the field of Agricultural Statistics, Agricultural Bioinformatics and Computer Applications. The Institute is also functioning as a Centre of Advanced Faculty Training (CAFT) in Agricultural Statistics and Computer Application under the aegis of Human Resource Development Programme of the Indian Council of Agricultural Research (ICAR). Both basic and applied research are being carried out in various areas of Agricultural Statistics including Sample Surveys relating to estimation of different parameters of interest in case of field crops, horticulture crops, livestock and fisheries etc. Scientists of the Division of Sample Surveys are engaged in research on various aspects of sample surveys like design and analysis of complex surveys, application of statistical softwares for survey data analysis, bootstrap variance estimation techniques, calibration and model calibration estimators, small area estimation, ranked set sampling, adaptive cluster sampling, multiple frame surveys, spatial sampling and estimation, geographically weighted regression based estimation approaches, application of GIS and remote sensing techniques in agricultural surveys etc. The division is also engaged in international collaborations with Food and Agriculture Organization of the United Nations (FAO) for several applied research activities in the field of sample surveys.

The broader objective of this training programme on "**Modern Data Analytics Techniques for Agricultural Surveys**" is to provide exposure to the participants belonging to different disciplines of agricultural sciences in proper understanding of various sampling techniques and estimation procedures, some recent developments in sample surveys, use of software packages for survey data analysis like MS-Excel, R, SAS, Python and SPSS, application of remote sensing and GIS techniques in agricultural surveys etc. More emphasis is given on the applied aspects rather than theoretical. The reference manual is presented in a simplified and comprehensive manner for better usage by the participants.

We take this opportunity to thank all the faculty members from the institute and the guest faculties who have devoted their valuable time and energy in making this training program successful. Their sincere efforts helped in bringing out this lecture manual on time. We are also thankful to the Chairman and members of various committees involved in successful organization of this training programme. We are also thankful to various ICAR Institutes, State Agricultural Universities etc. for nominating participants to this training programme. We are indebted to the Agricultural Education Division of ICAR for entrusting the responsibility of organizing this training programme. We are also grateful to Dr. Rajender Parsad, Director, ICAR-IASRI and Dr. Tauqueer Ahmad, Head, Division of Sample Surveys for their guidance and continuous support in this training programme and providing all the necessary facilities for smooth conduct of this training programme. In the end, we are thankful to one and all who helped in preparing this reference manual.

New Delhi **Authors**
February 11, 2025

विषय-सूची

CONTENTS

19.	Crop Yield Forecasting in India - **K. N. Singh**	**19.1-19.6**
20.	Overview of Bioinformatics and Applications in Agriculture - **Mir Asif Iquebal, Sarika and Dinesh Kumar**	**20.1-20.12**
21.	Overview of Calibration Estimation in Survey Sampling - **Kaustav Aditya**	**21.1-21.6**
22.	Statistical Data Integration in Survey Sampling - **Rahul Banerjee**	**22.1-22.10**
23.	Model-Based and Model-Assisted Approaches in Survey Sampling - **Rahul Banerjee, Ankur Biswas, Kaustav Aditya and Tauqueer Ahmad**	**23.1-23.8**
24.	Development of Indices with Survey Data and Applications - **Deepak Singh**	**24.1-24.6**
25.	Overview of Remote Sensing and Applications in Agricultural Surveys - **Prachi Misra Sahoo**	**25.1-25.10**
26.	Overview of GIS and Applications in Agricultural Surveys - **Prachi Misra Sahoo**	**26.1-26.20**
27.	Generation of Livestock Statistics in India eLISS Portal & App - **Prachi Misra Sahoo**	**27.1-27.22**
28.	Surveys for Estimation of Area and Production of Horticultural Crops - **Tauqueer Ahmad**	**28.1-28.12**
29.	Crop Yield estimation initiative for Crop Insurance under PMFBY - **Sunil Dubey and S. Bandyopadhyay**	**29.1-29.4**
30.	FASAL2.0 - An Overview - **Karan Choudhary, Preeti Tahlani and S. Bandyopadhyay**	**30.1-30.6**
31.	Energy Audit Surveys in Agriculture - **Kastav Aditya and Bharti**	**31.1-31.4**
32.	Digital Agriculture: ICAR Perspective - **Anil Rai**	**32.1-32.16**

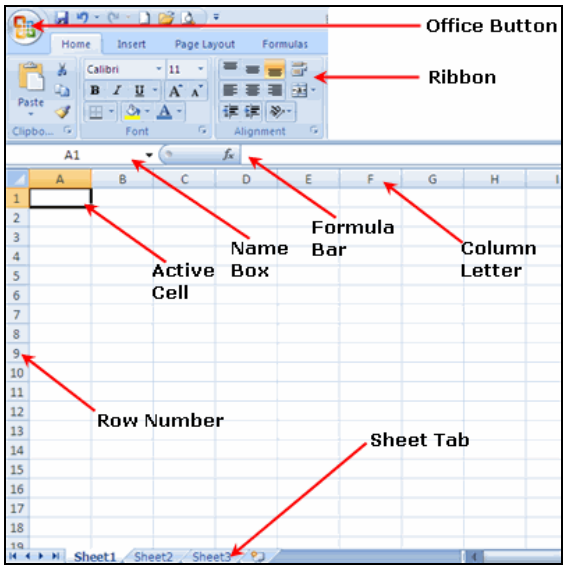
MS-EXCEL: STATISTICAL PROCEDURES

Cini Varghese

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

Introduction:

Microsoft (MS) Excel ( ) is a powerful spreadsheet that is easy to use and allows you to store, manipulate, analyze, and visualize data. It also supports databases, graphic and presentation features. It is a powerful research tool and needs a minimum of teaching. Spreadsheets offer the potential to bring the real numerical work alive and make statistics enjoyable. But the main disadvantage is that some advanced statistical functions are not available and it takes a longer computing time as compared to other specialized software.



Data Entry in Spreadsheets

- Data entry should be started soon after data collection in the field
- The raw data collected should be entered directly into computer. Calculations (e.g. % dry matter) or conversions (e.g. kg/ha to t/ha) by hand will very likely result in errors and therefore require more data checking once the data are in MS-Excel. Calculations can be written in MS-Excel using formulae (e.g. sum of wood biomass and leaf biomass to give total biomass).

Data Checking

One can use calculations and conversions for data checking. For example, if the collected data is grain yield per plot it may be difficult to see whether the values are

reasonable. However, if these are converted to yield per hectare then one can compare the numbers with our scientific knowledge of grain yields. Simple formulae can be written to check for consistency in the data. For example, if tree height is measured 3 times in the year, a simple formula that subtracts 'tree height 1' from 'tree height 2'can be used to check the correctness of the data. The numbers in the resulting column should all be positive. We cannot have a shrinking tree! For new columns of calculated or converted data suitable header information (what the new column is, units and short name) at the top of the data should be included.

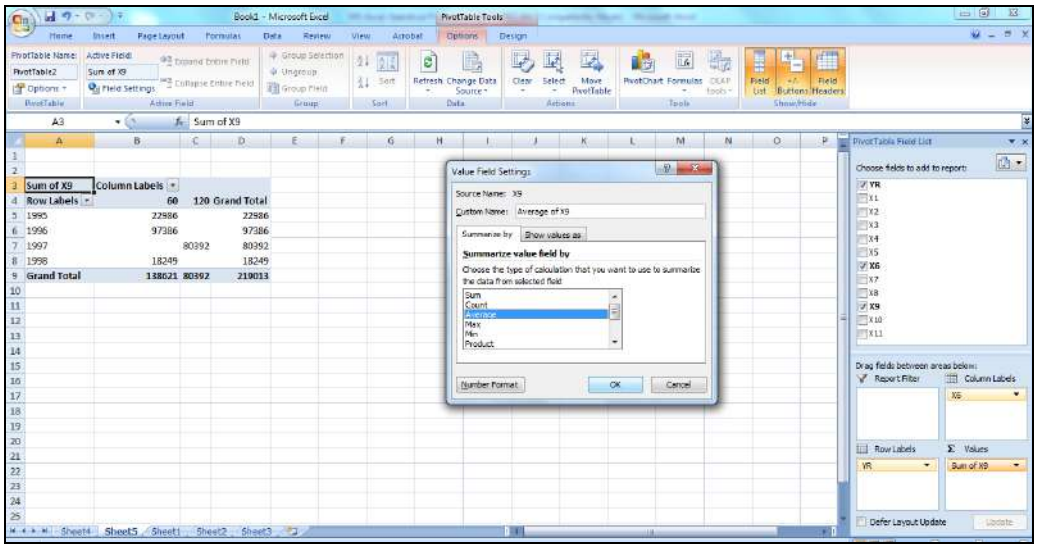
Missing Values

In MS-Excel the missing values are BLANK cells. It is useful to know this when calculating formulae and summaries of the data. For example, when calculating the average of a number of cells, if one cell is blank MS-Excel ignores this as an observation (i.e., the average is the sum/number of non-blank cells). But if the cell contains a '0' then this is included in the calculation (i.e., the average is the sum/no. of cells). In a column of 'number of fruit per plot', a missing value could signify zero (tree is there but no fruit), dead (tree was there but died so no fruit), lost (measurement was lost, illegible.) or not representative (tree had been browsed severely by goats). In this example, depending on the objectives of the trial, the scientist might choose to put a '0' in the cells of trees with no fruit and leave blank (but add comments) for the other 'missing values'.

Pivot Tables (to check consistency between replicates)

Variation between replicates is expected, but some level of consistency is also usual. We can use pivot tables to look at the data. A pivot table is an interactive worksheet table that quickly summarizes large amount of data using a format and calculation methods you choose. It is called pivot table because you can rotate its row and column heading around the core data area to give you different views of the source data. A pivot table provides an easy way for you to display and analyze summary information about data already created in MS-Excel or other application.

- Keep the cursor anywhere within the data range
- Choose “Insert” “Pivot Table” then “OK”
- From the “Pivot table Field List” drag and drop the respective fields under “Column Labels” , “Row Labels” and “Σ Values”
- Select “Value Field Settings” by clicking on the down arrow in “Σ Values” and choose the appropriate option and then click “OK”



Scatter Plots (to check consistency between variates)

We can often expect two measured variables to have a fairly consistent relationship with each other. For example, 'number of fruits' with 'weight of fruits' or Stover yield plotted against grain yield. To look for odd values we could plot one against the other in a scatter plot. Scatter plots are useful tools for helping to spot outliers. This option is available under "Insert" menu.

Line Plots (to examine changes over time)

Where measurements on a 'unit' are taken on several occasions over a period of time it may be possible to check that the changes are realistic. A check back at the problematic data which is not in the usual trend can be made. . This option is available under "Insert" menu.

Double Data Entry

One effective, although not always practical, way of checking for errors caused by data entry mistakes is double entry. The data are entered by two individuals onto separate sheets that have the same design structure. The sheets are then compared and any inconsistencies are checked with the original data. It is assumed that the two data entry operators will not make the same errors. There is no 'built-in' system for double entry in MS-Excel. However, there are some functions that can be used to compare the two copies. An example is the DELTA function that compares two values and returns a 1 if they are the same and a 0 otherwise. To use this function we would set up a third worksheet and input a formula into each cell that compares the two identical cells in the other two worksheets. The 0's on the third worksheet will therefore identify the contradictions between the two sets of data. This method can also be used to check survey data but for the process to work the records must be entered in exactly the same order in both sheets. If a section at the bottom of the third worksheet contains mostly 0's, this could indicate that you have omitted a record in one of the other sheets.

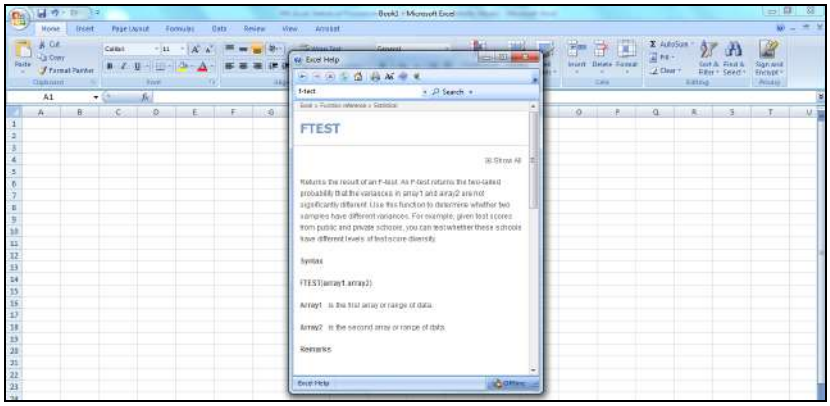
Preparing Data for Export to a Statistical Package

Statistical analysis of research data usually involves exporting the data into a statistical package such as GENSTAT, SAS or SPSS. These packages require you to give the MS-Excel cell range from which data are to be taken. In the latest editions of MS-Excel we can mark these ranges within MS-Excel and then transfer them directly into the statistical packages.

- Highlight the data you require including the column titles (the codes which have been used to label the factors and variables).
- Go to the Name Box, an empty white box at the top left of the spreadsheet. Click in this box and type a name for the highlighted range (e.g., Data). Press Enter.
- From now on, when you want to select your data to export go to the Name Box and select that name (e.g. Data). The relevant data will then be highlighted.

MS-Excel Help

If you get stuck on any aspect of MS-Excel then use the Help facility by clicking "F1" key. It contains extensive topics and by typing in a question you can extract the required information. See the snapshot below for an example:



FEATURES OF MS-EXCEL

Analytic Features

- The windows interface includes windows, pull down menus, dialog boxes and mouse support
- Repetitive tasks can be automated with MS-Excel. Easy to use macros and user defined functions
- Full featured graphing and charting facilities
- Supports on screen databases with querying, extracting and sorting functions
- Permits the user to add, edit, delete and find database records

Presentation Features

- Individual cells and chart text can be formatted to any font and font size
- Variations in font size, style and alignment control can be determined
- The user can add legends, text, pattern, scaling and symbols to charts.

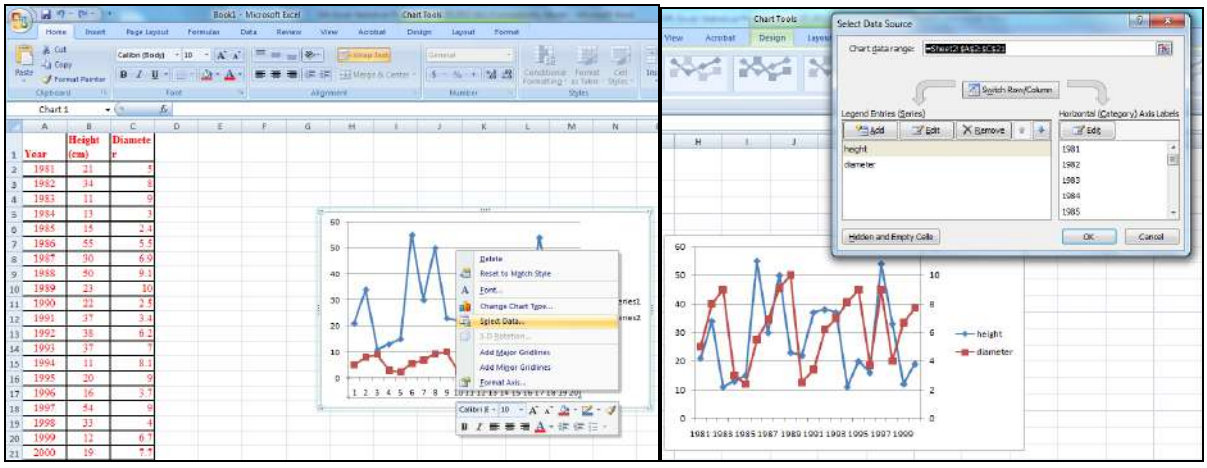
Charts and Graphs

A chart is a graphic representation of worksheet data. The dimension of a chart depends upon the range of the data selected. Charts are created on a worksheet or as a separate document that is saved with an extension xlsx. MS-Excel automatically scales the axes, creates columns categories and labels the columns. Values from worksheet cells or data points are displayed as bars, lines, columns, pie slices, or other shapes in the chart. Showing a data in a chart can make it clearer, interesting and easier to understand. Charts can also help the user to evaluate his/her data and make comparisons between different worksheet values.

Creating Line Chart

- Select relevant part of data
- Choose “Insert” “line”
- Select an appropriate option of line chart and click

Necessary changes in the chart can be done by clicking the right button of the mouse and choosing appropriate options.



Sorting and Filtering

MS-Excel makes it easy to organize, find and create report from data stored in a list.

Sort: To organize data in a list alphabetically, numerically or chronologically.

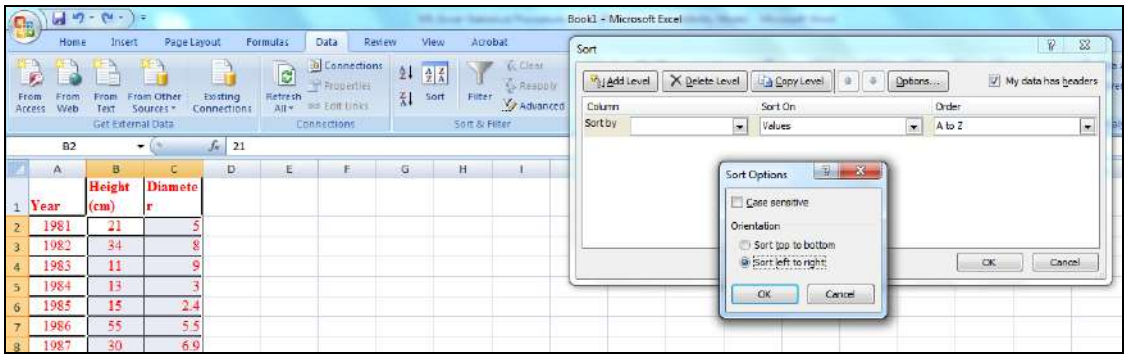
(i) To sort entire list

- Select a single cell in the list

- Choose “data” “sort”

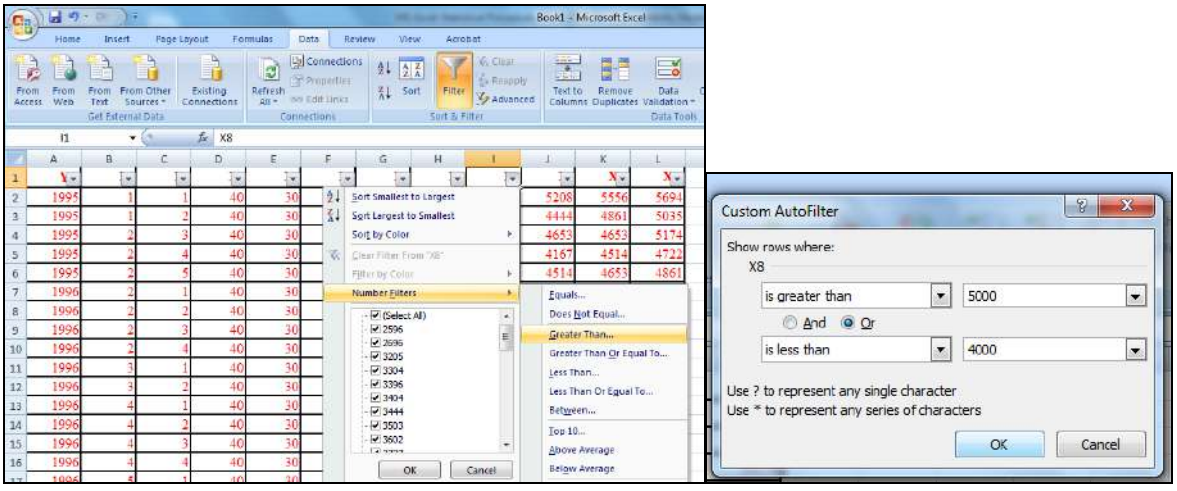
(ii) Sorting column from left to right

- Choose the “option” button in the sort dialog box
- In the sort option dialog box, select “sort left to right”
- Choose “OK”



Filter: To quickly find and work with a subset of your data without moving or sorting it.

- Choose “Data” and click on “Filter”
- MS-Excel place a drop down arrow directly on the column labels of the list
- Choose the column based on which the data has to be filtered. Clicking on the arrow displays a list of all the unique items in the column. Choose “Number Filter” option and define the required conditions.



STATISTICAL FUNCTIONS

Excel’s statistical functions are quite powerful. In general, statistical functions take lists as arguments rather than single numerical values or text. A list could be a group of numbers

separated by commas, such as (3,5,1,12,15,16), or a specified range of cells, such as (A1:A6), which is the equivalent of typing out the list (A1,A2,A3,A4,A5,A6). The function COUNT(list) counts the number of values in a list, ignoring empty or nonnumeric cells, whereas COUNTA(list) counts the number of values in the list that have any entry at all. MIN(list) returns a list's smallest value, whereas MAX(list) returns a list's largest value. The functions AVERAGE(list), MEDIAN(list), MODE(list), STDEV(list) all carry out the statistical operations you would expect (STDEV stands for standard deviation), when you pass a list of values as an argument.

Create a Formula

Formulas are equations that perform calculations on values in your worksheet. A formula starts with an equal sign (=). For example, the following formula multiplies 2 by 3 and then adds 5 to the result: =5+2*3. The following formulas contain operators and constants:

Example formula	What it does
=128+345	Adds 128 and 345
=5^2	Squares 5

- Click the cell in which you want to enter the formula.
- Type = (an equal sign).
- Enter the formula.
- Press ENTER.

Create a Formula that Contains References or Names: A1+23

The following formulas contain relative references and names of other cells. The cell that contains the formula is known as a dependent cell when its value depends on the values in other cells. For example, cell B2 is a dependent cell if it contains the formula =C2.

Example formula	What it does
=C2	Uses the value in the cell C2
=Sheet2!B2	Uses the value in cell B2 on Sheet2
=Asset-Liability	Subtracts a cell named Liability from a cell named Asset

- Click the cell in which the formula enter has to be entered.
- In the formula bar, type = (equal sign).
- To create a reference, select a cell, a range of cells, a location in another worksheet, or a location in another workbook. One can drag the border of the cell selection to move the selection, or drag the corner of the border to expand the selection.
- Press ENTER.

Create a Formula that Contains a Function: =AVERAGE(A1:B4)

The following formulas contain functions:

Example formula	What it does
=SUM(A:A)	Adds all numbers in column A
=AVERAGE(A1:B4)	Averages all numbers in the range

- Click the cell in which the formula enter has to be entered.
- To start the formula with the function, click “insert function” on the formula bar.
- Select the function.
- Enter the arguments. When the formula is completed, press ENTER.

Create a Formula with Nested Functions: =IF(AVERAGE(F2:F5)>50, SUM(G2:G5),0)

Nested functions use a function as one of the arguments of another function. The following formula sums a set of numbers (G2:G5) only if the average of another set of numbers (F2:F5) is greater than 50. Otherwise it returns 0.

STATISTICAL ANALYSIS TOOLS

Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that one can use to save steps when you develop complex statistical or engineering analyses. Provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Accessing the Data Analysis Tools: To access various tools included in the Analysis ToolPak click on “Data” menu, then click “Data Analysis” and select the appropriate analysis option. If the “Data Analysis” command is not available, we need to load the Analysis ToolPak “select and run the “Analysis ToolPack” from the “Add-Ins”.

Correlation

The “Correlation” analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. It can be used to determine whether two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

If the experimenter had measured two variables in a group of individuals, such as foot-length and height, he/she can calculate how closely the variables are correlated with each other. Select “Data”, “Data Analysis”. Scroll down the list, select “Correlation” and click OK. A new window will appear where the following information needs to be entered:

Input range. Highlight the two columns of data that are the paired values for the two variables. The cell range will automatically appear in the box. If column headings are included in this range, tick the Labels box.

Output range. Click in this box then select a region on the worksheet where the user want the data table displayed. It can be done by clicking on a single cell, which will become the top left cell of the table.

Click OK and a table will be displayed showing the correlation coefficient (r) for the data.

CORREL(array1, array2) also returns the correlation coefficient between two data sets.

Covariance

Covariance is a measure of the relationship between two ranges of data. The “covariance” tool can be used to determine whether two ranges of data move together, *i.e*., whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

To return the covariance for individual data point pairs, use the COVAR worksheet function.

Regression

The “Regression” analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables. For example, one can analyze how grain yield of barley is affected by factors like ears per plant, ear length (in cms), 100 grain weight (in gms) and number of grains per ear.

Descriptive Statistics

The “Descriptive Statistics” analysis tool generates a report of univariate statistics for data in the input range, which includes information about the central tendency and variability of the entered data.

Sampling

The “Sampling” analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, a representative sample

can be used. One can also create a sample that contains only values from a particular part of a cycle if you believe that the input data is periodic. For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

Random Number Generation

The "Random Number Generation" analysis tool fills a range with independent random numbers drawn from one of several distributions. We can characterize subjects in a population with a probability distribution. For example, you might use a normal distribution to characterize the population of individuals' heights.

ANOVA: Single Factor

"ANOVA: Single Factor" option can be used for analysis of one-way classified data or data obtained from a completely randomized design. In this option, the data is given either in rows or columns such that observations in a row or column belong to one treatment only. Accordingly, define the input data range. Then specify whether, treatments are in rows or columns. Give the identification of upper most left corner cell in output range and click OK. In output, we get replication number of treatments, treatment totals, treatment means and treatment variances. In the ANOVA table besides usual sum of squares, Mean Square, F-calculated and P-value, it also gives the F-value at the pre-defined level of significance.

ANOVA: Two Factors with Replication

This option can be used for analysis of two-way classified data with m-observations per cell or for analysis of data obtained from a factorial CRD with two factors with same or different levels with same replications.

ANOVA: Two Factors without Replication

This option can be utilized for the analysis of two-way classified data with single observation per cell or the data obtained from a randomized complete block design. Suppose that there are 'v' treatments and 'r' replications and then prepare a v × r data sheet. Define it in input range, define alpha and output range.

t-Test: Two-Sample Assuming Equal Variances:

This analysis tool performs a two-sample student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal. TTEST(array1,array2,tails,type) returns the probability associated with a student's t test.

t-Test: Two-Sample Assuming Unequal Variances:

This t-test form assumes that the variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. Use this test when the groups under study are distinct.

t-Test: Paired Two Sample For Means:

This analysis tool performs a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. One can use this test when there is a natural pairing of observations in the samples, like a sample group is tested twice - before and after an experiment.

F-Test Two-Sample for Variances

The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances. For example, you can use an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams. FTEST(array1, array2) returns the result of an F-test, the one tailed probability that the variances of Array1 and array 2 are not significantly different.

Transformation of Data

The validity of analysis of variance depends on certain important assumptions like normality of errors and random effects, independence of errors, homoscedasticity of errors and effects are additive. The analysis is likely to lead to faulty conclusions when some of these assumptions are violated. A very common case of violation is the assumption regarding the constancy of variance of errors. One of the alternatives in such cases is to go for a weighted analysis of variance wherein each observation is weighted by the inverse of its variance. For this, an estimate of the variance of each observation is to be obtained which may not be feasible always. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Major aims of applying transformations are to bring data closer to normal distribution, to reduce relationship between mean and variance, to reduce the influence of outliers, to improve linearity in regression, to reduce interaction effects, to reduce skewness and kurtosis. Certain methods are available for identifying the transformation needed for any particular data set but one may also resort to certain standard forms of transformations depending on the nature of the data. Most commonly used transformations in the analysis of experimental data are Arcsine, Logarithmic and Square root. These transformations of data can be carried out using the following options.

Arcsine (ASIN): In the case of proportions, derived from frequency data, the observed proportion p can be changed to a new form $\theta = \sin^{-1}(\sqrt{p})$. This type of transformation is known as angular or arcsine transformation. However, when nearly all values in the data lie

between 0.3 and 0.7, there is no need for such transformation. It may be noted that the angular transformation is not applicable to proportion or percentage data which are not derived from counts. For example, percentage of marks, percentage of profit, percentage of protein in grains, oil content in seeds, etc., can not be subjected to angular transformation. The angular transformation is not good when the data contain 0 or 1 values for p. The transformation in such cases is improved by replacing 0 with (1/4n) and 1 with [1-(1/4n)], before taking angular values, where *n* is the number of observations based on which p is estimated for each group.

ASIN gives the arcsine of a number. The arcsine is the angle whose sine is number and this number must be from -1 to 1. The returned angle is given in radians in the range $-\pi/2$ to $\pi/2$. To express the arcsine in degrees, multiply the result by $180/\pi$. For this go to the CELL where the transformation is required and write =ASIN (Give Cell identification for which transformation to be done)* 180*7/22 and press ENTER. Then copy it for all observations.

Example: ASIN (0.5) equals 0.5236 ($\pi/6$ radians) and ASIN (0.5)* 180/PI equals 30 (degrees).

Logarithmic (LN): When the data are in whole numbers representing counts with a wide range, the variances of observations within each group are usually proportional to the squares of the group means. For data of this nature, logarithmic transformation is recommended. It squeezes the bigger values and stretches smaller values. A simple plot of group means against the group standard deviation will show linearity in such cases. A good example is data from an experiment involving various types of insecticides. For the effective insecticide, insect counts on the treated experimental unit may be small while for the ineffective ones, the counts may range from 100 to several thousands. When zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective in normalizing positively skewed distributions. It is also used to achieve additivity of effects in certain cases.

LN gives the natural logarithm of a positive number. Natural logarithms are based on the constant e (2.718281828845904). For this go the CELL where the transformation is required and write = LN(Give Cell Number for which transformation to be done) and press ENTER. Then copy it for all observations.

Example: LN(86) equals 4.454347, LN(2.7182818) equals 1, LN(EXP(3)) Equals 3 and EXP(LN(4)) equals 4. Further, EXP returns e raised to the power of a given number, LOG returns the logarithm of a number to a specified base and LOG 10 returns the base-10 logarithm of a number.

Square Root (SQRT): If the original observations are brought to square root scale by taking the square root of each observation, it is known as square root transformation. This is appropriate when the variance is proportional to the mean as discernible from a graph of group variances against group means. Linear relationship between mean and variance is

commonly observed when the data are in the form of small whole numbers (*e.g.*, counts of wildlings per quadrat, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.). When the observed values fall within the range of 1 to 10 and especially when zeros are present, the transformation should be, $\sqrt{(y + 0.5)}$.

SQRT gives square root of a positive number. For this go to the CELL where the transformation is required and write = SQRT (Give Cell No. for which transformation to be done = 0.5) and press ENTER. Then copy it for all observations. However, if number is negative, SQRT return the #NUM ! error value.

Example: SQRT(16) equals 4, SQRT(-16) equals #NUM! and SQRT(ABS(-16)) equals 4.

Once the transformation has been made, the analysis is carried out with the transformed data and all the conclusions are drawn in the transformed scale. However, while presenting the results, the means and their standard errors are transformed back into original units. While transforming back into the original units, certain corrections have to be made for the means. In the case of log transformed data, if the mean value is $\bar{y}$, the mean value of the original units will be antilog ($\bar{y}$ + 1.15 $\bar{y}$) instead of antilog ($\bar{y}$). If the square root transformation had been used, then the mean in the original scale would be antilog $((\bar{y} + V(\bar{y}))^2$ instead of $(\bar{y})^2$ where $V(\bar{y})$ represents the variance of $\bar{y}$. No such correction is generally made in the case of angular transformation. The inverse transformation for angular transformation would be $p = (\sin q)^2$.

Sum(SUM): It gives the sum of all the numbers in the list of arguments. For this go to the CELL where the sum of observations is required and write = SUM (define data range for which the sum is required) and press ENTER. Instead of defining the data range, the exact numerical values to be added can also be given in the argument viz. SUM (Number1, number2,…), number1, number2,… are 1 to 30 arguments for which you want the sum.

Example: If cells A2:E2 contain 5, 15,30,40 and 50; SUM(A2:C2) equals 50, SUM(B2:E2,15) equals 150 and SUM(5,15) equals 20.

Some other related functions with this option are:

AVERAGE returns the average of its arguments, PRODUCT multiplies its arguments and SUMPRODUCT returns the sum of the products of corresponding array components.

Sum of Squares (SUMSQ): This gives the sum of the squares of the list of arguments. For this go to the CELL where the sum of squares of observations is required and write = SUMSQ (define data range for which the sum of squares is required) and press ENTER.

Example: If cells A2:E2 contain 5, 15, 30, 40 and 50; SUMSQ(A2:C2) equals 1150 and SUMSQ(3,4) equals 25.

Matrix Multiplication (MMULT): It gives the matrix product of two arrays, say array 1 and array 2. The result is an array with the same number of rows as array1, say a and the same number of columns as array2, say b. For getting this mark the a × b cells on the spread sheet. Write =MMULT (array 1, array 2) and press Control +Shift+ Enter. The number of columns in array1 must be the same as the number of rows in array2, and both arrays must contain only numbers. Array1 and array2 can be given as cell ranges, array constants, or references. If any cells are empty or contain text, or if the number of columns in array1 is different from the number of rows in array2, MMULT returns the ≠VALUE! error value.

Determinant of a Matrix (MDETERM): It gives the value of the determinant associated with the matrix. Write = MDETERM(array) and press Control + Shift + Enter.

Matrix Inverse (MINVERSE): It gives the inverse matrix for the non-singular matrix stored in a square array, say of order p. i.e., an array with equal number of rows and columns. For getting this mark the p × p cells on the spread sheet where the inverse of the array is required and write = MINVERSE(array) and press Control + Shift + Enter. Array can be given as a cell range, such as A1:C3; as an array constant, such as {1,2,3;4,5,6;7,8,8}; or as a name for either of these. If any cells in array are empty or contain text, MINVERSE returns the ≠VALUE! error value.

Example: MINVERSE ({4,-1;2,0}) equals {0,0.5;-1,2}and MINVERSE ({1,2,1;3,4,-1;0,2,0}) equals {0.25, 0.25,-0.75;0,0,0.5;0.75,-0.25,-0.25}.

Transpose (TRANSPOSE): For getting the transpose of an array mark the array and then select copy from the EDIT menu. Go to the left corner of the array where the transpose is required. Select the EDIT menu and then paste special and under paste special select the TRANSPOSE option.

EXERCISES ON MS-EXCEL

1. Table below contains values of pH and organic carbon content observed in soil samples collected from natural forest. Compute mean, median, standard deviation, range and skewness of the data.
2.

Soil pit	pH (x)	Organic carbon (%) (y)		Soil pit	pH (x)	Organic carbon (%) (y)
1	5.7	2.10		9	5.4	2.09
2	6.1	2.17		10	5.9	1.01
3	5.2	1.97		11	5.3	0.89
4	5.7	1.39		12	5.4	1.60

5	5.6	2.26		13	5.1	0.90
6	5.1	1.29		14	5.1	1.01
7	5.8	1.17		15	5.2	1.21
8	5.5	1.14				

3. Consider the following data on various characteristics of a crop:

pp	**ph**	**ngl**	**yield**
142	0.525	8.2	2.47
143	0.64	9.5	4.76
107	0.66	9.3	3.31
78	0.66	7.5	1.97
100	0.46	5.9	1.34
86.5	0.345	6.4	1.14
103.5	0.86	6.4	1.5
155.99	0.33	7.5	2.03
80.88	0.285	8.4	2.54
109.77	0.59	10.6	4.9
61.77	0.265	8.3	2.91
79.11	0.66	11.6	2.76
155.99	0.42	8.1	0.59
61.81	0.34	9.4	0.84
74.5	0.63	8.4	3.87
97	0.705	7.2	4.47
93.14	0.68	6.4	3.31
37.43	0.665	8.4	1.57
36.44	0.275	7.4	0.53
51	0.28	7.4	1.15
104	0.28	9.8	1.08
49	0.49	4.8	1.83
54.66	0.385	5.5	0.76
55.55	0.265	5	0.43
88.44	0.98	5	4.08
99.55	0.645	9.6	2.83
63.99	0.635	5.6	2.57
101.77	0.29	8.2	7.42
138.66	0.72	9.9	2.62
90.22	0.63	8.4	2

(i) Sort yield in ascending order and filter the data ph less than 0.3 or greater than 0.6 from the data.

(ii) Find the correlation coefficient and fit the multiple regression equation by taking yield as dependent variable.

4. Let **A**, **B** and **C** be three matrices as follows:

$$\mathbf{A}=\begin{bmatrix}2 & 4 & 6 & 1 & 9\\ 3 & 5 & 6 & 7 & 2\\ 8 & 3 & 9 & 1 & 5\\ 3 & 1 & 1 & 1 & 3\end{bmatrix} \quad \mathbf{B}=\begin{bmatrix}1 & 3\\ 5 & 7\\ 2 & 4\\ 1 & 9\\ 8 & 1\end{bmatrix} \quad \mathbf{C}=\begin{bmatrix}2 & 3 & 1 & 8 & 4\\ 3 & 6 & 7 & 8 & 8\\ 2 & 3 & 5 & 5 & 7\\ 2 & 3 & 6 & 6 & 1\\ 1 & 2 & 8 & 5 & 5\end{bmatrix}.$$

Find (i) **AB** (ii) $\mathbf{C}^{-1}$ (iii) $|\mathbf{A}|$ (iv) $\mathbf{A}^{T}$.

5. Draw line graph for the following data on a tree species:

Year	Height (cm)	Diameter
1981	21	5.0
1982	34	8.0
1983	11	9.0
1984	13	3.0
1985	15	2.4
1986	55	5.5
1987	30	6.9
1988	50	9.1
1989	23	10.0
1990	22	2.5
1991	37	3.4
1992	38	6.2
1993	37	7.0
1994	11	8.1
1995	20	9.0
1996	16	3.7
1997	54	9.0
1998	33	4.0
1999	12	6.7
2000	19	7.7

Also draw a bar diagram using the above data.

6. The table below lists plant height in cm of seedlings of rice belonging to the two varieties. Examine whether the two samples are coming from populations having equal variance, using F-test. Further, test whether the average height of the two groups are the same, using appropriate t-test.

Plot	1	2	3	4	5	6	7	8	9	10
Group I	23	17.4	17	20.5	22.7	24	22.5	22.7	19.4	18.8
Group II	8.5	9.6	7.7	10.1	9.7	13.2	10.3	9.1	10.5	7.4

7. Examine whether the average organic carbon content measured from two layers of a set of soil pits from a pasture are same using paired t-test from the data given below:

Soil pit		1	2	3	4	5	6	7	8	9	10
Organic carbon (%)	**Layer 1 (x)**	1.59	1.39	1.64	1.17	1.27	1.58	1.64	1.53	1.21	1.48
	Layer 2 (y)	1.21	0.92	1.31	1.52	1.62	0.91	1.23	1.21	1.58	1.18

8. Mycelial growth in terms of diameter of the colony (mm) of *R. solani* isolates on PDA medium after 14 hours of incubation is given in the table below. Carry out the CRD analysis for the data. And draw your inferences.

R. solani isolates	**Mycelial growth**		
	Repl. 1	**Repl. 2**	**Repl. 3**
RS 1	29.0	28.0	29.0
RS 2	33.5	31.5	29.0
RS 3	26.5	30.0	
RS 4	48.5	46.5	49.0
RS 5	34.5	31.0	

9. Following is the data on mean yield in kg per plot of an experiment conducted to compare the performance of 8 treatments using a Randomized Complete Block design with 3 replications. Perform the analysis of variance.

Treatment (Provenance)	**Replication**		
	I	**II**	**III**
1	30.85	38.01	35.10
2	30.24	28.43	35.93
3	30.94	31.64	34.95
4	29.89	29.12	36.75
5	21.52	24.07	20.76
6	25.38	32.14	32.19
7	22.89	19.66	26.92
8	29.44	24.95	37.99

10. From the following data make a summary table for finding out the average of X_9 for various years and various levels of X_6 using pivot table and pivot chart report option of MS-Excel.

YR	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
1995	1	1	40	30	0	60	40	4861	5208	5556	5694
1995	1	2	40	30	0	60	40	4167	4444	4861	5035
1995	2	3	40	30	0	60	40	4618	4653	4653	5174
1995	2	4	40	30	0	60	40	4028	4167	4514	4722

1995	2	5	40	30	0	60	40	4306	4514	4653	4861
1996	2	1	40	30	0	60	40	6000	5750	5499	6250
1996	2	2	40	30	0	60	40	5646	5000	5250	5444
1996	2	3	40	30	0	60	40	4799	5097	4896	5299
1996	2	4	40	30	0	60	40	5250	5299	4194	4847
1996	3	1	40	30	0	60	40	5139	5417	5764	5903
1996	3	2	40	30	0	60	40	5417	5694	6007	6111
1996	4	1	40	30	0	60	40	6300	7450	7750	8000
1996	4	2	40	30	0	60	40	6350	7850	7988	8200
1996	4	3	40	30	0	60	40	5750	6400	6600	6700
1996	4	4	40	30	0	60	40	6000	7250	7450	7681
1996	5	1	40	30	0	60	40	3396	4090	5056	5403
1996	5	2	40	30	0	60	40	5194	5000	6000	6500
1996	5	3	40	30	0	60	40	4299	4250	4750	5250
1996	6	1	40	30	0	60	40	4944	5194	5000	5097
1996	6	2	40	30	0	60	40	5395	5499	5499	5597
1996	6	3	40	30	0	60	40	3444	5646	5000	5000
1996	6	4	40	30	0	60	40	6250	6500	6646	6750
1997	1	1	120	30	30	120	60	5839	6248	6199	6335
1997	1	2	120	30	30	120	60	5590	5652	5702	5851
1997	2	1	120	30	30	120	60	4497	4794	4894	5205
1997	2	2	120	30	30	120	60	4696	5006	5304	5702
1997	2	3	120	30	30	120	60	4398	4596	4894	5304
1997	2	4	120	30	30	120	60	4497	5503	5702	6099
1997	3	1	120	30	30	120	60	4199	5602	5801	6000
1997	3	2	120	30	30	120	60	3404	3901	4199	4497
1997	3	3	120	30	30	120	60	3602	5404	5503	5801
1997	3	4	120	30	30	120	60	3602	4297	4497	4696
1997	4	1	120	30	30	120	60	3205	3801	4199	4894
1997	4	2	120	30	30	120	60	3801	4794	6099	6298
1997	4	3	120	30	30	120	60	3503	5205	6298	6795
1997	4	4	120	30	30	120	60	3205	4894	5503	6199
1997	5	1	120	30	30	120	60	4199	4099	4199	4297
1997	5	2	120	30	30	120	60	3304	3702	3602	3801
1997	5	3	120	30	30	120	60	2596	2894	3106	3205
1998	1	1	40	30	0	60	40	3727	3106	3404	3503
1998	1	2	40	30	0	60	40	4894	4348	4447	4534
1998	1	3	40	30	0	60	40	2696	2795	3056	3205
1998	2	2	40	30	0	60	40	5503	4298	4497	4795
1998	2	3	40	30	0	60	40	5006	3702	3702	3901

11. From the data given in problem 10, sort X_{10} in ascending order. Also, filter the data for $X_{11} < 4200$ or $X_{11} > 5000$.

HANDS-ON EXERCISE ON SAMPLING SCHEMES USING MS EXCEL

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Sample surveys are a cost-effective method for data collection, allowing for accurate and reliable inferences about population parameters. These surveys involve selecting a representative subset of the population, from which conclusions about the entire target population can be drawn. Analyzing survey data is crucial for extracting meaningful insights from collected responses. Microsoft Excel offers a range of powerful tools for cleaning, organizing, and analyzing survey data efficiently. Key advantages of using Excel for survey analysis include:

- Easy data entry and organization
- Built-in statistical and analytical functions
- Visual representation through charts and graphs
- PivotTables for summarizing large datasets

2. Importing and Organizing Survey Data

- Importing Data
  - If survey data is in CSV or Excel format, open the file in Excel
  - If using Google Forms, download responses as a CSV file and open it in Excel

- Cleaning Data
  - Remove blank rows/columns
  - Ensure uniform responses (e.g., "Yes" vs. "yes")
  - Remove duplicates. Use Data > Remove Duplicates
  - Handle missing data

- Formatting Data for Analysis
  - Convert responses into numerical values where necessary (e.g., Yes = 1, No = 0)
  - Use Data Validation for consistency in data entry

3. Exploratory Data Analysis (EDA) of Survey Data

Basic statistical summaries help understand the central tendency and dispersion of survey data. Excel functions to compute these include:

3.1 Basic Statistical Functions:

- **Count**: =COUNT(cell range) for numerical data and =COUNTA(range) for categorical data
- **Minimum and Maximum**: =MIN(cell range), =MAX(cell range)
- **Mean (Average)**: =AVERAGE(cell range)
- **Median**: =MEDIAN(cell range)

- **Mode**: =MODE(cell range)
- **Standard Deviation**: =STDEV.P(range) or =STDEV.S(cell range)
- **Variance**: =VAR.P(cell range) or =VAR.S(cell range)

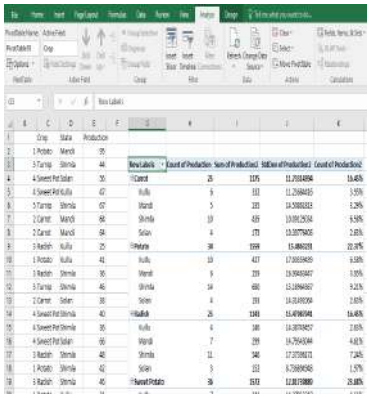
	Formula		
Sum	SUM(G2:G20)	COUNT	COUNT(G2:G20)
Count	COUNT(G2:G20)		COUNTA(G2:G20)
			COUNTBLANK(G2:G20)
			COUNTIF(G2:G20,">50")
Measure of Central Tendency			COUNTIFS(G2:G20,">50",G2:G20,"<70")
Mean	AVERAGE(G2:G20)		
Median	MEDIAN(G2:G20)	AVERAGE	AVERAGE(G2:G20)
Mode	MODE(G2:G20)		AVERAGEA(G2:G20)
GM	GEOMEAN(G2:G20)		AVERAGEIF(G2:G20,">45")
HM	HARMEAN(G2:G20)		AVERAGEIFS(G2:G20,F2:F20,"Onion")
Measure of Dispersion			
Standard Deviation	STDEV.P(G2:G20)		
Variance	VAR.P(G2:G20)		
Average Deviation	AVEDEV(G2:G20)		
Skewness	SKEW(G2:G20)		
Kurtosis	KURT(G2:G20)		

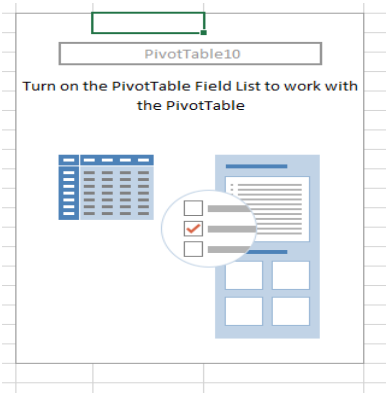
3.2 Data Visualization: Visualizing survey data helps identify trends and relationships. Some key chart types in Excel include:

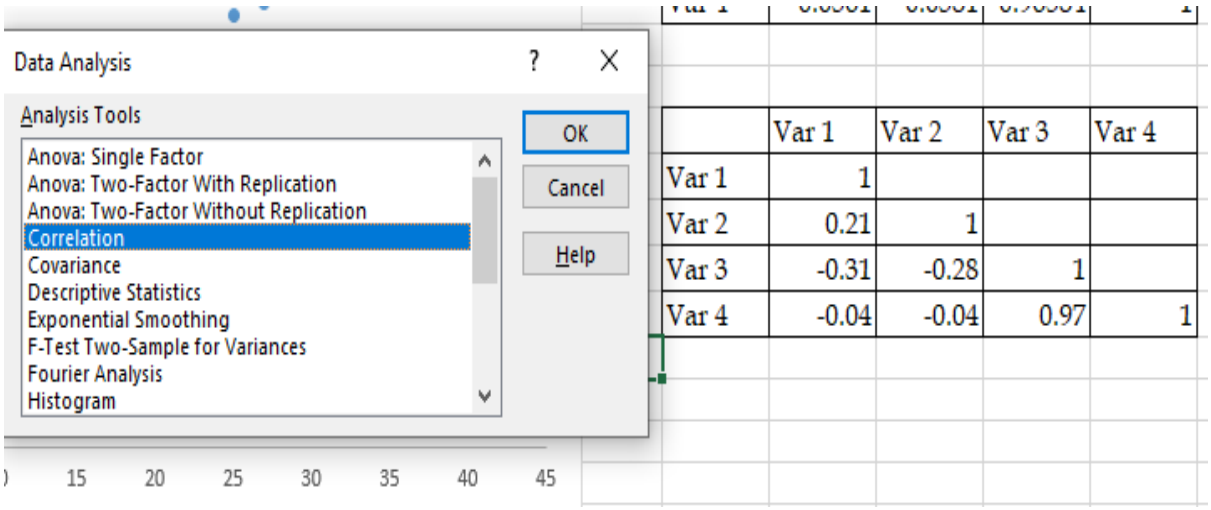
- **Bar Charts**: Useful for categorical data comparisons (Insert > Bar Chart)
- **Histograms**: Shows frequency distributions (Insert > Histogram)
- **Pie Charts**: Displays proportions (Insert > Pie Chart)
- **Box Plots**: Highlights outliers and spread (Insert > Box and Whisker)
- **Scatter Plots**: Shows relationships between numerical variables (Insert > Scatter Plot)
- **Pivot Charts**: Dynamic visual representation of aggregated data (Insert > PivotChart)

Pivot Tables for Survey Data Exploration: Pivot tables are one of the most powerful tools in MS Excel for summarizing and analyzing survey data. Steps to create a pivot table includes:

- Select the dataset and go to Insert > PivotTable.
- Choose whether to place the PivotTable in a new or existing worksheet.
- Drag fields into the Rows, Columns, Values, and Filters sections.
- Use Value Field Settings to apply functions like Count, Sum, Average, etc.
- Apply filters and sorting for deeper insights.







	Var 1	Var 2	Var 3	Var 4
Var 1	1			
Var 2	0.21	1		
Var 3	-0.31	-0.28	1	
Var 4	-0.04	-0.04	0.97	1

3.3 Identifying Trends and Patterns

- Use **Conditional Formatting** (Home > Conditional Formatting) to highlight trends.
- Apply **Filters and Sorting** to isolate specific groups.
- Utilize **Moving Averages** (=AVERAGE(cell range)) for trend analysis.
- Analyze **Correlations** using =CORREL(cell range1, cell range2) to measure relationships between numerical survey responses.

3.4 Detecting Outliers: Outliers can distort results and should be detected and handled appropriately.

- Use **Box Plots** to visually identify outliers.
- Apply the **Interquartile Range (IQR) method**:
  - Find Q1 (=QUARTILE.INC(range,1)) and Q3 (=QUARTILE.INC(range,3)).
  - Compute IQR: =Q3 - Q1.
  - Identify outliers as values below Q1 - 1.5*IQR or above Q3 + 1.5*IQR.
- Highlight outliers using **Conditional Formatting**.

3.5 **Descriptive Statistics Using Data Analysis ToolPak:** The Data Analysis ToolPak is an Excel add-in that provides a collection of analysis tools, including statistical, financial analysis, etc. This add-in simplifies tasks that would otherwise require complex formulas or external tools. It's especially useful for analysts, researchers, and students who need to conduct quick data processing, hypothesis testing, and advanced statistical analysis. Go to File > Options > Add-ins > Analysis ToolPak > Enable

Overview of the Data Analysis Tools: The ToolPak includes a variety of tools for:

- **Descriptive Statistics**: Measures of central tendency, dispersion, and distribution.
- **Regression Analysis**: Linear regression, multiple regression, and other statistical models.

- **ANOVA**: One-way and two-way analysis of variance.
- **Correlation**: Calculating the correlation coefficient between two variables.
- **t-Test**: Paired and two-sample tests for hypothesis testing.
- **Histograms**: Creating and analyzing frequency distributions.
- **F-Test**: Testing equality of variances between two datasets.
- **Moving Averages**: Smoothing out data to identify trends.
- **Sampling**: Creating random samples and conducting analysis on them.

Production	
Mean	46.54605
Standard Error	1.097378
Median	49
Mode	51
Standard Deviation	13.52938
Sample Variance	183.0442
Kurtosis	-1.24167
Skewness	-0.22964
Range	44
Minimum	23
Maximum	67
Sum	7075
Count	152
Largest(1)	67
Smallest(1)	23
Confidence Level(95.0%)	2.168198

4. **Sampling Schemes:** Sampling is the process of selecting a subset of individuals from a population to estimate characteristics of the entire population. Sampling schemes can be broadly categorized into two types, such as: Probability sampling (Simple Random Sampling, Systematic Sampling, Stratified Sampling, Cluster Sampling, Multi Stage Sampling etc.) and Non Probability Sampling (Quota Sampling, Judgement Sampling, Snow Ball Sampling etc.)

   4.1 Simple Random Sampling (SRS): In Simple Random Sampling (SRS), each element of the population has an equal chance of being selected. Steps involved to select simple random sample in MS Excel:

   - Create a list of items (or population). For example, assume a population of 100 people, numbered from 1 to 100 in column A.
   - In the next column (Column B), generate random numbers using function =RANDBETWEEN() function, which generates a random number between given range.
     - **Formula**: =RANDBETWEEN()
     - Drag down the formula to select the units in the sample.

   4.2 Systematic Sampling: In Systematic Sampling, first unit is selected with the help of random numbers, and the remaining units are selected automatically according to a predetermined pattern. The systematic sampling technique is operationally more

convenient than simple random sampling. It also ensures, at the same time that each unit has an equal probability of inclusion in the sample. Suppose there are N units in a population, which are numbered from 1 to N. Let N = nk, (n = sample size and k = an integer), $k=N/n$ and if a random number less than or equal to k is selected and every kth unit thereafter. The resulting samples are called kth systematic sampling and such a process is called Linear Systematic sampling. Steps involved to select systematic sample in MS Excel:

- Prepare the data Like in SRS, create a population list (1 to 100)
- Choose the Sampling Interval (k):
  Decide the sample size (for example, 10) and the population size (100). The sampling interval k is calculated as: N/n = 100/ 10 = 10. So, the interval is every 10th person.
- Select the First Element Randomly: Use the =RANDBETWEEN() function to select a random starting point between 1 and k, for example, if the random number between 1 and 10 is 3, then the starting point is the 3rd person.
- Select Every kth Person: From the randomly chosen starting point, select every 10th person in the list. If the random starting point is 4, select the 4th, 14th, 24th, and so on.

4.3 Stratified Sampling

In stratified random sampling, population is divided into non-overlapping groups based on specific characteristics, such as age, gender, income level, or education. Each group is called a stratum. After dividing the population into homogeneous strata, a random sample is taken from each stratum. The sample size taken from each group can be proportional to the stratum size or equal across all strata, depending on the objective. Then, these samples are used to give conclusion about the whole target population. By ensuring that each subgroup is represented, stratified random sampling helps reduce sampling error and increases the precision of the overall estimate. Steps involved to select stratified random sample in MS Excel:

- Prepare the data: Create a list of individuals and assign them to different homogeneous strata.
- Sample within Each Stratum: For each subgroup, randomly select a subset of individuals. The allocation of sample size within each stratum can be done using different methods depending on the research objectives and the relative importance of each stratum. The common methods for allocating the sample size across strata are equal allocation, proportional allocation, optimum allocation.
- Combine the analysis to get final results.

4.4 Cluster Sampling: In Cluster Sampling, the population is divided into clusters (groups), and entire clusters are randomly selected for inclusion in the sample. Steps involved to select cluster in MS Excel:

- Prepare the data: Divide the population into clusters. For example, group individuals by location or department.
- Select Clusters Randomly: Use a random number generator to select a certain number of clusters.
- Include All Members of the Selected Clusters: Once clusters are selected, include all individuals within those clusters in the final sample. If there are 10 clusters, randomly select 2 clusters and include all individuals in those clusters.

5. Conclusion

This chapter explored how to implement various sampling schemes in MS Excel, including Simple Random Sampling, Systematic Sampling, Stratified Sampling, and Cluster Sampling. Excel's built-in functions, such as =RAND(), sorting tools, and basic arithmetic operations, make it an excellent tool for performing sampling in both straightforward and more complex scenarios. The flexibility of Excel allows users to efficiently manage and analyze data, providing a practical solution for a wide range of sampling techniques.

R SOFTWARE - AN OVERVIEW

Kaustav Aditya and Hukum Chandra

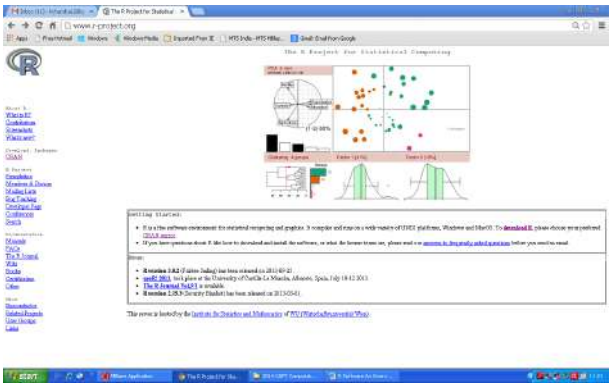
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

R is a free software environment for statistical computing and graphics. It is almost perfectly compatible with S-plus. The only thing you need to do is download the software from the internet and use an editor to write your program (e.g. Notepad). It contains most standard methods of statistics as well as lot of less commonly used methods and can be used for programming and to construct your own functions. It is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. It is available for down load from **http://www.r-project.org**/. The primary purpose of this lecture is to introduce R.

2. To Download R Software

- In any web browser (e.g. Microsoft Internet Explorer), go to: **http://www.r-project.org**



- Downloads: **CRAN**

- Set your Mirror: Anyone in the **India or any other country** is fine.

- On your right hand side you will see Download R for Windows. Click there
- Click on **base**

- Click on **R-3.0.2.exe** (52 megabytes, 32/64 bit) and save it to your hard disc.

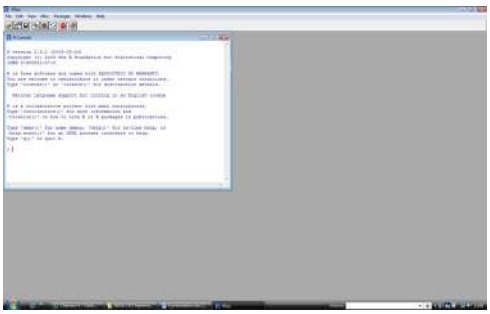
- This is the latest available version of the software. It is an '.exe' file, which you can save in your hard disc. By double clicking on the name of this file, R is automatically installed. All you need to do is follow the installation process.

3. To Open R Software

The installation process automatically creates a shortcut for R. Double click this icon to open the R environment. R will open up with the appearance of a standard Windows implementation (i.e. various windows and pull-down menus). Note that R is an interpreted language and processes commands on a line by line basis. Consequently it is necessary to hit **ENTER** after typing in (or pasting) a line of R code in order to get R to implement it.

4. To Run R Program Code

The main active window within the R environment is the **R Console**. This is a line editor and output viewer combined into one window. Here at the command **prompt** (the symbol >), we can enter R commands which run instantly upon pressing the carriage return key. This sign (>) is called prompt, since it prompts the user to write something, see below.



We can also run blocks of code which we have copied into the paste buffer from another source. In this session we shall use the Windows-supplied editor **Notepad** to display and edit our R program code. If we were to write some R of code, then simply copy it from the editor and paste it into the R Console, then the code would run in real time.

To Open The Editor

Here we are using the Windows-supplied editor **Notepad** to display and edit our R program code, although any general-purpose editor will suffice. Open Notepad by going to the Start button and clicking on:

Start > All Programs > Accessories > Notepad

Having opened Notepad, open the file, for example, **Intro_to_R.txt** (containing the program code, assume that it is copied in C: / derive) by selecting the following option from the pull-down menu:

File > Open

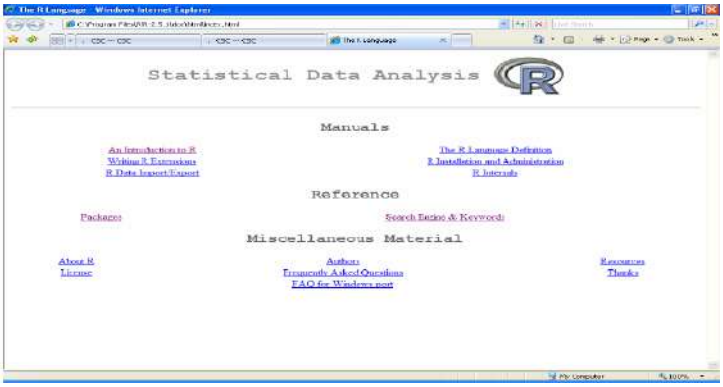
Click on the down-arrow at the top of the "Open" dialog box and change the selection to "Look in" **C:**. You should now see the filename **Intro_to_R.txt** among a list of files. Double-click on the filename to open it.

A Couple Of Other Useful Things

- Please remember that R is **case-sensitive** so we need to be consistent in our use of lower and upper case letters, both for commands and for objects.
- When the program has finished, we should see the **red** command prompt (>) pop up in the R Console window. This indicates that control is returned to the user, so that you can now type more R commands if you wish.
- A **comment** in R code begins with a hash symbol (#). Whole lines may be commented or just the tail-end of a line. Examples are:

Help

Html-help can be invoked from the Help-menu. From the opening webpage, you can access manuals, frequently asked questions, references to help for individual packages, and most importantly, Search Engine. Help is the best place to find out new functions, and get descriptions on how to use them.



Getting Started With R

Commands in R are given at the **command prompt**.

Simple calculations, vectors and graphics

To begin with, we'll use R as a calculator. Try the following commands:

> 2+7

> 2/(3+5)

> sqrt(9)+5^2

> sin(pi/2)-log(exp(1))

Help about a specific command can be had by writing a question mark before the command, for instance:

> ?log

As an alternative, help can be used; in this case, help (log). The help files are a great resource and you will soon find yourself using them frequently.

Comments can be written using the #-symbol as follows:

> 2+3 # The answer should be 5

Vectors and matrices

Vectors and matrices are of great importance in many numerical problems. To create a vector named mydata and assign the values 7, −2, 5 to it, we write as follows:

> mydata <- c(7,-2,5)

The symbol **<-** (or alternatively use =) should be read as "**assigns**". The command c can be interpreted (by you, the user) as column or combine. The second element of the vector can be referred to by the command

> mydata[2]

and elements between 2 and 3 (i.e. elements 2 and 3) by

> mydata[2:3]

Vectors can be manipulated, for instance by adding a constant to all elements, as follows.

> myconst <- 100; mydata + myconst

Using the **semicolon** allows us to write multiple commands on a single line

A vector x consisting of the integers between 1 and 10; 1, 2, . . . , 10; can be created by writing

```
> x <- c(1:10)
```

Vectors with sequences of numbers with particular increments can be created with the seq command:

```
> mydata1 <- seq(0,10,2) # integers between 0 and 10, with increment 2
```

Read x and y

```
x<- c(2,3,1,5,4,6,5,7,6,8)
y<- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)
```

Read two vectors

```
weight<- c(60, 72, 57,90)
height<-c(1.75, 1.80, 1.65, 1.90)
bmi<- weight/height^2                    # Compute body mass index (BMI)
```

Functions on vectors

```
length(x)          #To compute length of data in x.
[1] 10
sum(x)             #To compute sum of data in x.
[1] 47
sum(x^2)
[1] 265
mean(x)            #To compute mean of data in x.
[1] 4.7
mean(y)
[1] 22
var(x)             #To compute variance of x.
[1] 4.9
sqrt(var(x))       # To compute standard deviation of x.
[1] 2.213594
sum((x-mean(x))^2)
[1] 44.1
sqrt(var(x))/mean(x)*100    #To compute coefficient of variation
```

To compute summary features of data in x

```
summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	3.25	5.00	4.70	6.00	8.00

To compute summary features of data in x^2

```
summary(x^2)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	10.75	25.00	26.50	36.00	64.00

Some calculations

```
sum(weight)
mean(weight)  or sum(weight)/ length(weight)
Denote by  xbar= mean(weight)  then
sqrt(sum((weight- xbar)^2))/ length(weight))
sd(weight)
cor(x,y)        #To compute correlation coefficient between x and y.
var(x,y)        #To compute covariance between x and y.
```

Slightly more complicated example …

The rule of thumb is that the BMI for a normal weight individual should be between 20 and 25, and we want to know if our data deviate systematically from that.

- We can use a one sample t test to assess whether the 6 persons' BMI can be assumed to have mean 22.5 given that they come from a normal distribution.
- We can use function t.test
- Although you might not be knowing about t test but example is just to give some indication of what real statistical output look like

```
t test        (see ? t.test)
t.test (bmi, mu=22.5)
        One Sample t-test
        data:  bmi
        t = -0.5093,        df = 3,        p-value = 0.6456
        alternative hypothesis: true mean is not equal to 22.5
        95 percent confidence interval:
                                18.29842        25.54231
        sample estimates:
        mean of x
        21.92036
```

If mu is not given then t.test would use default mu=0

The p value is not small, indicating that it is not at all unlikely to get data like those observed if the mean were in fact 22.5

Classical Tests

To load the library of classical tests statistics available with R software use

library(stats)

#To get results of t-test for comparing population means of x and y when variances are not equal.

```
t.test(x,y)
```

To get results for usual t-test when variances are equal. If T is replaced by F then it is equal to t.test(x, y)

```
t.test(x,y,var.equal=T)
```

```
?t.test
library(stats)
x<- c(2,3,1,5,4,6,5,7,6,8)
y<- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)
mean(x)
mean(y)
var(x,y)
cor(x,y)
t.test(x)
t.test(x,y)
t.test(x,y,var.equal=T)
var.test(x,y)          #To compare variances of x and y.
```

The commands **rbind** and **cbind** can be used to merge row or column vectors to matrices. Try the following:

```
x <- c(1,2,3)
y <- c(4,5,6)
A = cbind(x,y)
B = rbind(x,y)
C = t(B)
```

The last command gives the matrix transpose of B. Now type A, B or C to see what the different matrices look like.

5. Simple Graphics

Graphics - one of the most important aspects of presentation and analysis of data is generation of proper graphics. Graphic features of a data can be viewed very effectively using R. R is capable of creating high quality graphics. Graphs are typically created using a series of high-level and low-level plotting commands. High-level functions create new plots and low-level functions add information to an existing plot. Customize graphs (line style, symbols, color, etc) by specifying graphical parameters. Specify graphic options using the par() function. The function par() is used to set or get graphical parameters. This function contains 70 possible settings and allows you to adjust almost any feature of a graph. Graphic parameters are reset to the defaults with each new graphic

device. Most elements of par() can be set as additional arguments to a plot command, however there are some that can only be set by a call to par(), mfrow, mfcol see the documentation for others.

Scatterplot And Line Graphs

Scatter plots: are useful for studying dependencies between variables.

- The **plot()** function is used for producing scatterplots and line graphs

See ? **plot**

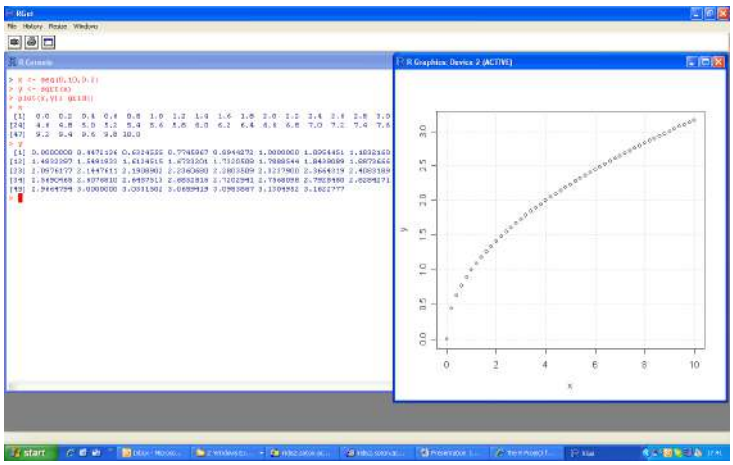
- Using the **plot command**

  x <- seq(0,10,0.2)

  y <- sqrt(x)

  plot(x,y); grid()

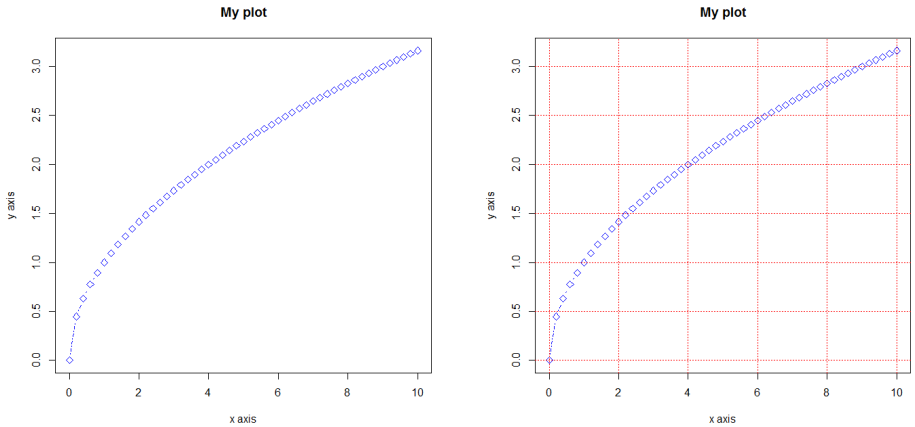
- As one might guess, the last command adds a grid to the plot.



plot(x,y); grid()

plot(x,y, **type="b"**, col="blue", lwd=1, lty=4, pch=5, main="My plot", xlab="x axis", ylab="y axis")

grid(col="red")

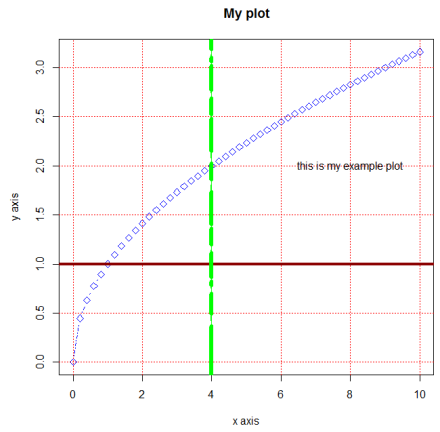


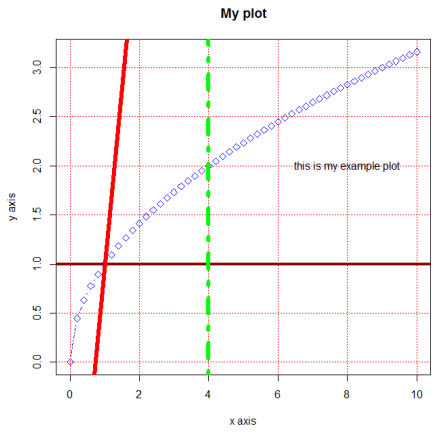
Common arguments for plot()

type	1-character string denoting the plot type

xlim	x limits, c(x1, x2)
ylim	y limits, c(y1, y2)
main	Main title for the plot
sub	Sub title for the plot
xlab	x-axis label
ylab	y-axis label
col	Color for lines and points
pch	Number referencing a plotting symbol or a character string
cex	A number giving the character expansion of the plot symbols
lty	Number referencing a line type
lwd	Line width

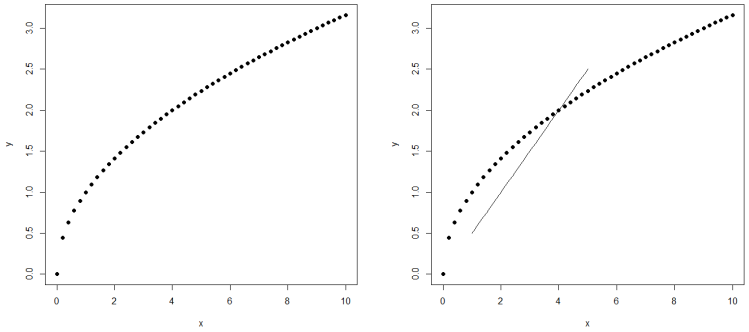
```
plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis", ylab="y axis")
grid(col="red")
text(8,2,"this is my example plot")
abline(h=1,v=4, col=c("darkred","green"), lty=c(1,4), lwd=c(4,6))
reg.lm=lm(x~y)
abline(reg.lm, col="red",lwd=6)                    #To add the regression line
```





There is wealth of plotting parameters you can set

```
plot(x,y)
plot(x,y, pch=16) : plot with new mark with dark circle
x1<- seq(1,5,0.1)
lines(x1,.5*x1)  #lines will add (x,y) values
```



dx<- rnorm(20,5,5) ## generate 100 random number from standard normal distribution

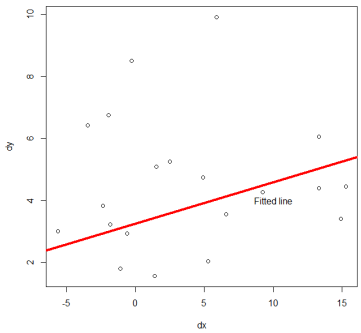
dy<- rchisq(20,5) ## generate 100 random number from chisq distribution with mean 5

plot(dx,dy,pch=1)

fit<-lm(dx~dy)

abline(fit,col="red",lwd=4)

text(10,4,"Fitted line")



See ? plot

See ? points

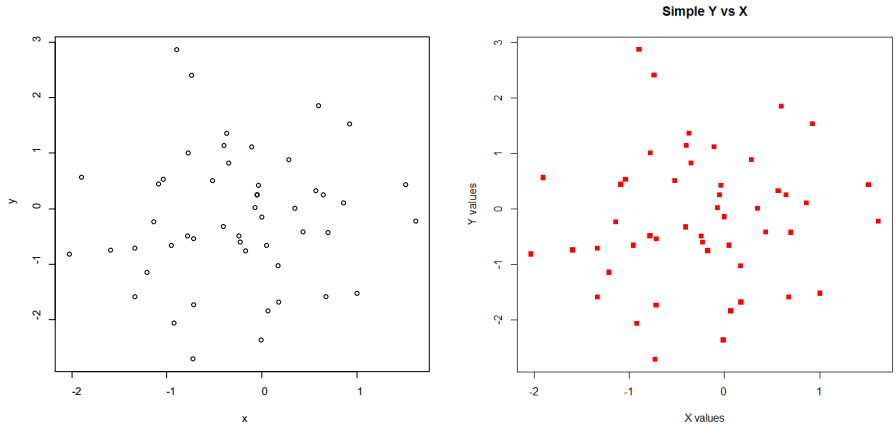
x <- rnorm(50) ;y <- rnorm(50)

group <- rbinom(50, size=1, prob=.5)

Basic Scatterplot

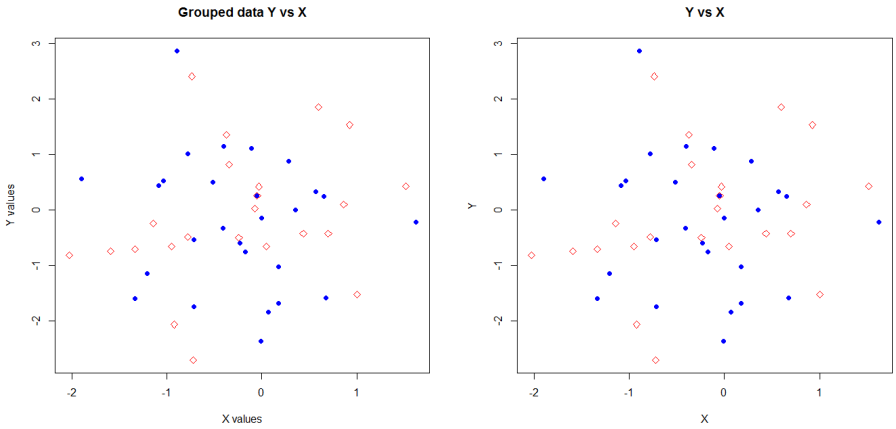
plot(x, y)

plot(x, y, xlab="X values", ylab="Y values", main="Simple Y vs X", pch=15, col="red")



Distinguish between two separate groups

plot(x, y, xlab="X values", ylab="Y values", main="Grouped data Y vs X", pch=ifelse(group==1, 5, 19), col=ifelse(group==1, "red", "blue"))



plot(x, y, xlab="X", ylab="Y", main="Y vs X", type="n")

points(x[group==1], y[group==1], pch=5, col="red")

points(x[group==0], y[group==0], pch=19, col="blue")

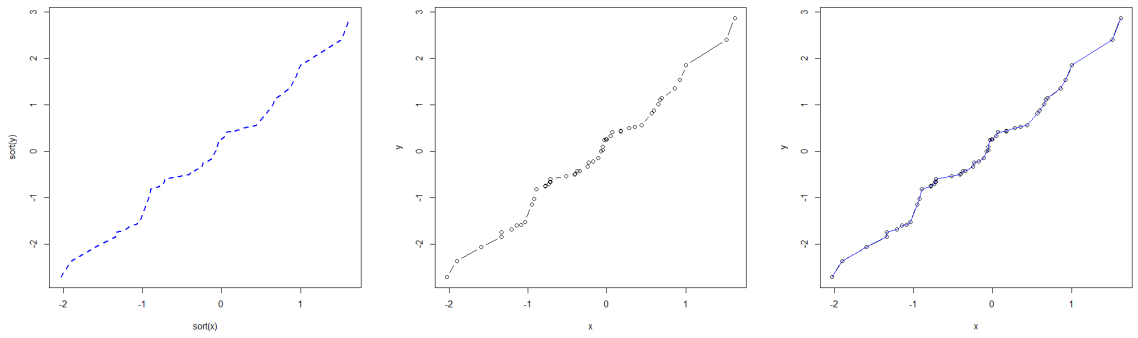
plot(x, y, xlab="X", ylab="Y", main="Y vs X", type="n")

points(cbind(x,y)[group==1,], pch=5, col="red")

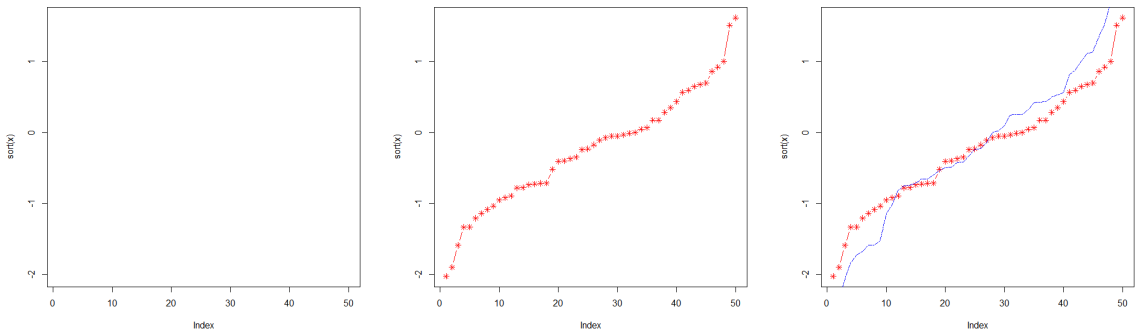
points(cbind(x,y)[group==0,], pch=19, col="blue")

Line Graphs

plot(sort(x), sort(y), type="l", lty=2, lwd=2, col="blue")




```
plot(x, y, type="n")
lines(sort(x), sort(y), type="b")
lines(cbind(sort(x),sort(y)), type="l", lty=1, col="blue")
plot(sort(x), type="n")
lines(sort(x), type="b", pch=8, col="red")
lines(sort(y), type="l", lty=6, col="blue")
```

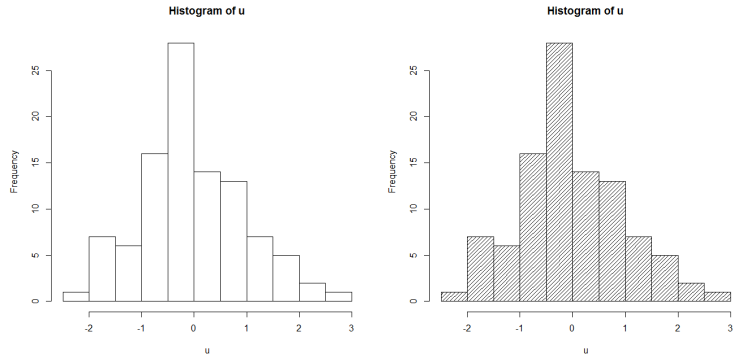


Histogram

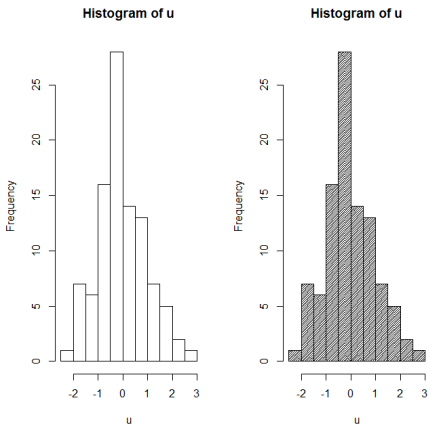
Histograms: used to study the distribution of continuous data, use command **hist**.

hist: function to plot histogram

```
u<- rnorm(100)        # generate 100 random numbers from SND
hist(u)               #default histogram
hist(u, density=20)   #with shading
```

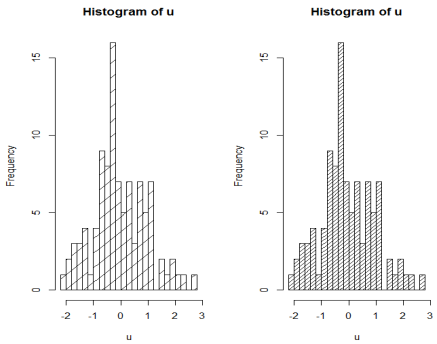


The sequence of commands below plots two histograms in one window
par(mfrow=c(a,b)) gives a rows with b plots on each row. Try
par(mfrow=c(1,2)); hist(u);hist(u, density=50)



#with specific number of bins

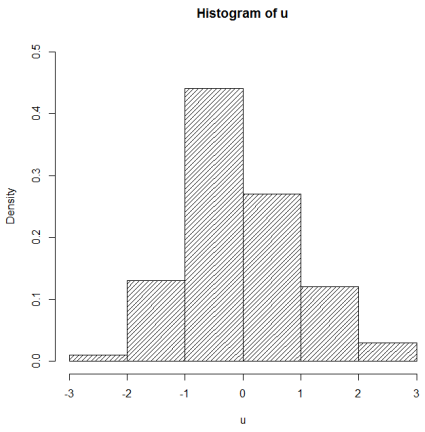
par(mfrow=c(1,2)); hist(u, density=5, breaks=20); hist(u, density=20, breaks=20)



Read in the help file about hist- **help(hist)**

Proportion, instead of frequency also specifying y-axis

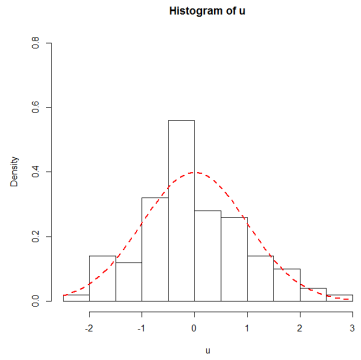
hist(u, density=20, breaks=-3:3, ylim=c(0,.5), prob=TRUE)



hist(u,freq=F,ylim = c(0,0.8))

curve(dnorm(x), col = 2, lty = 2, lwd = 2, add = TRUE)

The freq=F argument to hist ensures that the histogram is in terms of densities rather than absolute counts



overlay normal curve with x-lab and ylim

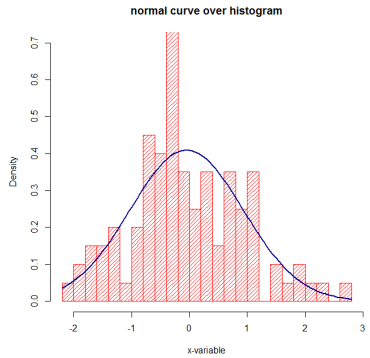
colored normal curve

Uses the observed mean and standard deviation for plotting the normal curve

m<-mean(u) ;std<-sqrt(var(u))

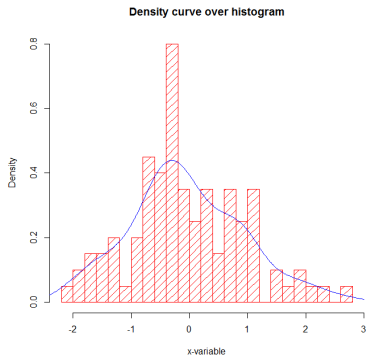
hist(u, density=20, breaks=20, prob=TRUE, xlab="x-variable", col="red", ylim=c(0, 0.7), main="normal curve over histogram")

curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)



hist(u, density=10, breaks=20, col="red", prob=TRUE, xlab="x-variable", ylim=c(0,0.8),main="Density curve over histogram")

lines(density(u),col = "blue")

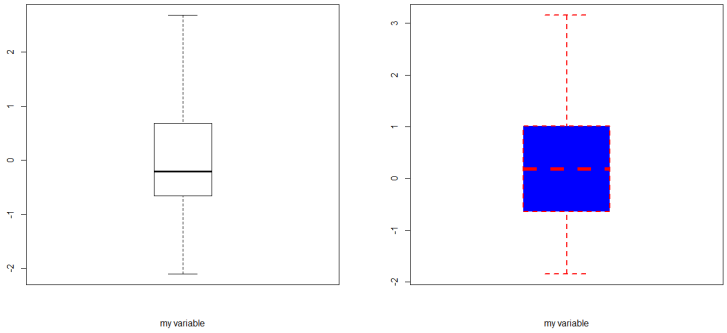


Boxplots

Boxplots: also a useful tool for studying data. It shows the median, quartiles and possible outliers. The R command is **boxplot**, which we use on the same variables as the histogram:

boxplot(u, xlab="my variable", boxwex=.4) # Basic boxplot

boxplot(u, xlab="my variable", boxwex=.6, col="blue", border= "red", lty=2, lwd=2)



we create data: three variables

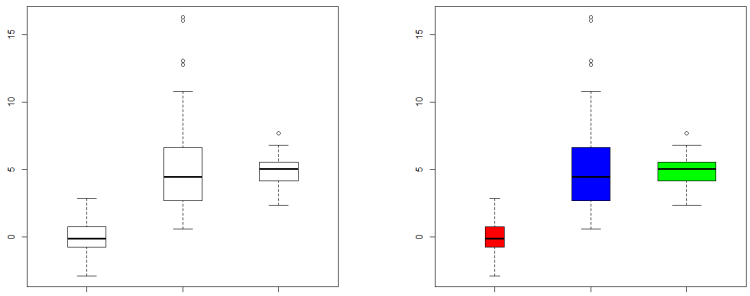
u1<- rnorm(100) ## generate 100 random number from standard normal distribution

u2<- rchisq(100,5) ## generate 100 random number from chisq distribution with mean 5

u3<- rnorm(100,5,1) ## generate 100 random number from normal distribution with mean 5, sd 1

boxplot(u1,u2,u3, boxwex=.4)

boxplot(u1,u2,u3, boxwex=c(.2,.4,.6),col=c("red","blue","green"))

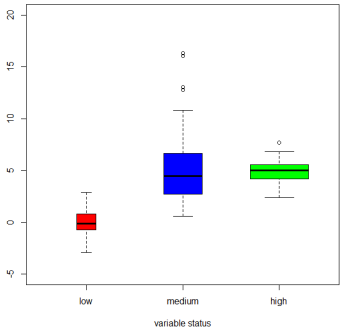


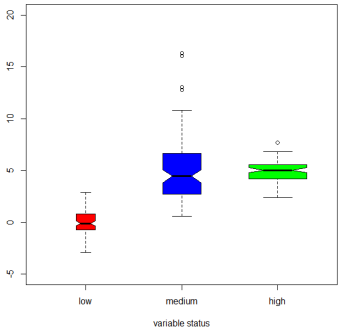
variablename<-c("low","medium", "high")

boxplot(u1,u2,u3,names=variablename,boxwex=c(.2,.4,.6), col=c("red","blue","green"),

ylim=c(-5, 20), xlab="variable status")

boxplot(u1,u2,u3,names=variablename,
boxwex=c(.2,.4,.6),col=c("red","blue","green"),ylim=c(-5, 20),xlab="variable status",
notch = TRUE)





try

boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot = F)

boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot = T)

?boxplot

Barchart (Or Barplot)

The R command is barplot

MPCE <- c(400, 300,600,550,425)

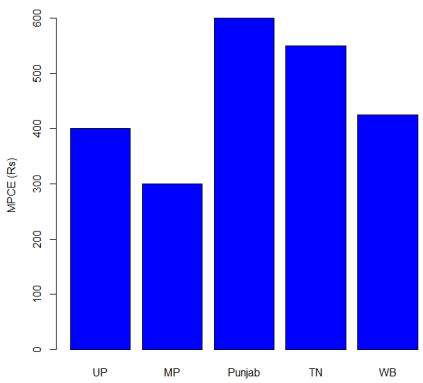
Suppose data in MPCE are average MPCE of some states whose names are to be assigned against their value. Following commands are required:

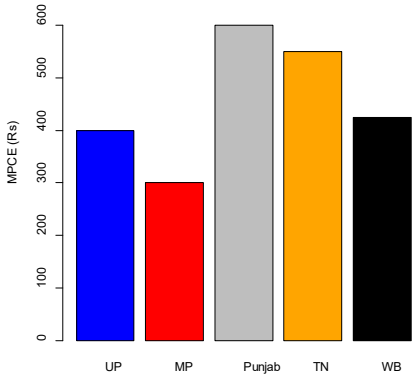
names(MPCE)<-c("UP","MP","Punjab","TN","WB")

To assign names of states. Double quotation mark " " means that names are characters not numeric.

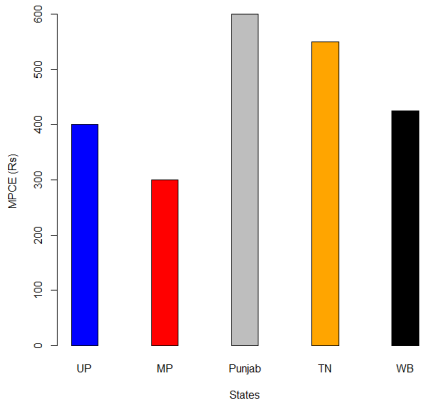
barplot(MPCE, names=names(MPCE), ylab="MPCE (Rs)",col="blue")

barplot(MPCE, names=names(MPCE),ylab="MPCE (Rs)", col = c("blue","red","gray","orange","black"))





barplot(MPCE, space=2,names=names(MPCE),xlab="States", ylab="MPCE (Rs)", col = c("blue","red","gray","orange","black"))

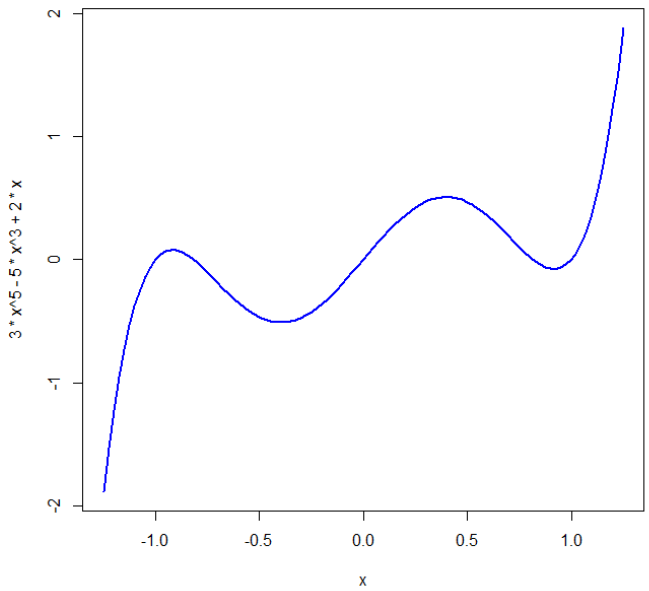


?barplot

Curve

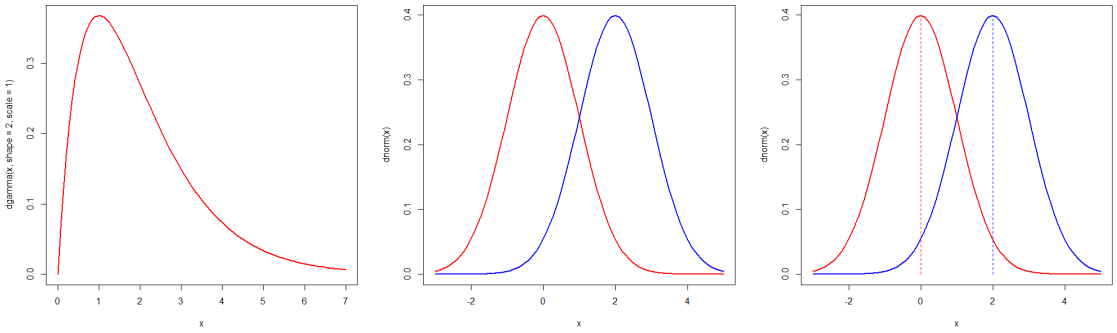
- The function **curve()** draws a curve corresponding to a given function
- If the function is written within curve() it needs to be a function of x
- If you want to use a multiple argument function, use x for the argument you wish to plot over

```
# Plot a 5th order polynomial
curve(3*x^5-5*x^3+2*x, from=-1.25, to=1.25, lwd=2, col="blue")
```




```
# Plot the gamma density
  curve(dgamma(x, shape=2, scale=1), from=0, to=7, lwd=2, col="red")
# Plot multiple curves, notice that the first curve determines the x-axis
curve(dnorm, from=-3, to=5, lwd=2, col="red")
curve(dnorm(x, mean=2), lwd=2, col="blue", add=TRUE)
```

```
# Add vertical lines at the means
lines(c(0, 0), c(0, dnorm(0)), lty=2, col="red")
lines(c(2, 2), c(0, dnorm(2, mean=2)), lty=2, col="blue")
```



Saving Graphs

- Graphs can be saved using several different formats, such as PDFs, JPEGs, and BMPs, by using pdf(), jpeg() and bmp(), respectively
- Graphs are saved to the current working directory

  Save graphics by choosing File -> Save as
- # Create a single pdf of figures, with one graph on each page

  pdf("SavingExample.pdf", width=7, height=5) # Start graphics device

  pdf("C://SavingExample.pdf", width=7, height=5)

```
x <- rnorm(100)
hist(x, main="Histogram of X")
plot(x, main="Scatterplot of X")
dev.off()                          # Stop graphics device
```

 # Create multiple pdfs of figures, with one pdf per figure

```
pdf(width=7, height=5, onefile=FALSE)
x <- rnorm(100)
hist(x, main="Histogram of X")
plot(x, main="Scatterplot of X")
dev.off()                          # Stop graphics device
```

6. Packages

- **Packages** are collections of **R** functions, data, and compiled code in a well-defined format. The directory where packages are stored is called the **library**
- The base distribution comes with some high priority add on packages, for example, boot, nlme, stats, grid, foreign, MASS, spatial etc

- The packages included as default in base distribution implement standard statistical functionality, for example, linear models, classical tests etc
- Packages not included in the base distribution can be downloaded and installed directly from R prompt
- Once installed, they have to be loaded into the session to be used
- Currently, the CRAN package repository has **4348 packages**
- **library()** **# To see all installed packages**
- **help("INSTALL")** or **help("install.packages")** in R for information on how to install packages from this repository

Adding Packages

- Choose **Install Packages** from the **Packages** menu
- Select a **CRAN Mirror**
- Select a package (e.g. car)
- Then use the **library(*package*)** function to load it for use (e.g. library(car))

7. Handling Data

Creating data frames

The command data.frame can be used to organize data of different kinds and to extract subsets of said data. Assume that we have data about three persons and that we store it as follows:

```
length <- c(180,175,190)
weight <- c(75,82,88);
name <- c("Anil","Ankit","Sunil")
friends <- data.frame(name,length,weight)
```

friends is now a data frame containing the data for the three persons. Data can easily be extracted:

```
> my.names <- friends$name
> length1 <- friends$length[1]
```

8. Reading Data

Reading data from files

There are a few principal functions reading data into R

- read.table, read.csv, for reading tabular data
- readLines, for reading lines of a text file
- source, for reading in R code files (inverse of dump)
- load, for reading in saved workspaces

read.csv(file, header = TRUE, sep = ",", quote="\"", dec=".", fill = TRUE, comment.char="", ...)

Specify the package

library (MASS)

Set working directory

setwd("G:/Course")

Reading ASCII Format

mydata=read.table("G:/Course/yielddata.txt")

dim(mydata)

summary(mydata)

mydata=read.table("G:/Course /yielddata.txt",header=T)

dim(mydata)

summary(mydata)

names(mydata)

mydata=read.table(file="yielddata.txt",header=T)

dim(mydata)

summary(mydata)

names(mydata)

```
 [1] "Dist"          "Yield"         "MARG_HH_F"     "HH_SIZE"
 [5] "NetArea"      "Cropedarea"    "Netirrig"      "GrossIrrigated"
 [9] "Rainfall"      "Fert"
```

mydata1=data.frame(mydata)

mydata1$Yield

mydata1$Fert

Extract district with yield less than median yield

mydata1$Yield[mydata1$Yield<median(mydata1$Yield)]

Extract data with yield less than median yield

mydata2=mydata1[mydata1$Yield<median(mydata1$Yield)]

mydata2=mydata1[mydata1$Yield<median(mydata1$Yield),]

dim(mydata2)

Read Data (Execl)

Call the require library

Load package XLConnect

The XLConnect package is part of the Comprehensive R Archive Network (CRAN). It can be easily installed by using the install.packages() command in your R session

install.packages ("XLConnect")

To load the package, use the library() or require() command in your R session

loadWorkbook() - loading/creating an Excel workbook

The loadWorkbook() function loads a Microsoft Excel workbook, so that it can then be further manipulated. Setting the create argument to TRUE will ensure the file will be created, if it does not exist yet. Both .xls and .xlsx file formats can be used.

```
loadWorkbook (filename , create = TRUE )
library(XLConnect)
library (MASS)
mydata2=loadWorkbook(file="yielddata.xls", create = TRUE)
readWorksheet(mydata3,sheet="yielddata",,header=T)
```

READING DATA IN OTHER FORMAT

library (foreign)

Read SPSS Dataset

```
MySpssdata=read.spss(file="yielddata.sav", use.value.labels=True, to.data.frame=True)
```

Read STAT Dataset

```
MyStatdata=read.dta(file="yielddata.dta")
```

Writing Data From Files

```
write.table(Result, ""MyResults.txt ")
write(Results,"MyResults2.txt")
write(Results,"MyResults2.txt",ncolumns=2)
```

How to save R workshop

```
save.image("myworshop.RData")
```

9. Analysis of a Data Set

We will study a data set from the early 70's, with data about different cars (Cars data set). Load the data set by writing

```
> data(mtcars)
```

You can read more about the data by looking at the help file:

```
> ?mtcars
```

mtcars package:datasets R Documentation

Motor Trend Car Road Tests

Description:

The data was extracted from the 1974 _Motor Trend_ US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Usage:

mtcars

Format:

A data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (lb/1000)

[, 7] qsec 1/4 mile time

[, 8] vs V/S

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

Source:

Henderson and Velleman (1981), Building multiple regression models interactively. _Biometrics_, *37*, 391-411.

Examples:

```
pairs(mtcars, main = "mtcars data")
coplot(mpg ~ disp | as.factor(cyl), data = mtcars, panel = panel.smooth, rows = 1)
```

Exercise. Answer the following questions using the help file:

1. How many cars are included in the data set?
2. Which years are the models from?
3. What does the mpg value desribe?

To see the entire data set, simply write

```
> mtcars
```

Exercise. To get familiar with the data set, answer the following non-statistical questions.

1. Are there any cars that weigh more than 5000 (lb/1000)?
2. How many cylinder has the motor of the Volvo 142E?
3. Are there any cars with 5 forward gears? Do they have automatic or manual transmisson?

Descriptive Statistics

Data can be summarized using simple measures such as mean, median, standard deviation, maximum and minimum and so on. A summary of a few such measures for the mtcars data set is obtained by writing

> summary(mtcars)

Measures can also be studied one at a time:

> mean(mtcars$hp); median(mtcars$hp); quantile(mtcars$wt); **max(mtcars$mpg)**

>sd(mtcars$mpg) # standard deviation

>var(mtcars$mpg) # variance

>sd(mtcars$mpg)^2 # sd*sd=var?

The command attach is very useful when dealing with data frames. By writing attach(mtcars) the references to the variables in mtcars can be shortened; instead of the long references above we can write:

> mean(hp); median(hp); quantile(wt); max(mpg)

> par(mfrow=c(1,2)); hist(mtcars$mpg); hist(mtcars$wt)

> boxplot(mtcars$mpg); x11(); boxplot(mtcars$wt)

The x11 command opens a new window which the next figure will be plotted in.

> plot(mtcars$wt,mtcars$mpg)

The correlation (which measures linear dependence) can be calculated using the command cor (use to help file to see how). What is the correlation in this case? Does it agree with the slope?

> cor(mtcars$wt,mtcars$mpg)

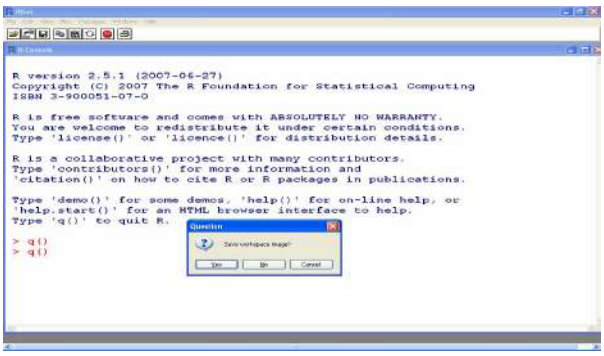
Linear regression

>lm(mtcars$wt~mtcars$mpg)

Try to see help (lm)

10. Quitting R

R can be closed with the command q( ). After issuing the quit command, R asks whether to save the workspace or not:



It is usually a good idea to save the workspace, since this creates a special file that can be directly read into R, and one can commence working with the same datasets and results already generated without a need to start from the scratch again. Saved workspace is in a file called .RData, and all the commands given during the same R session are saved in a file called .Rhistory. To load the workspace into R again, one can simply double-click on the file .Rdata, and R should open automatically with all the data and results loaded. Note however that libraries are not loaded automatically, and these should be loaded (if needed) before commencing the work.

Strengths And Weaknesses Of R

Strengths

- free and open source, supported by a strong user community
- highly extensible and flexible
- implementation of modern statistical methods
- moderately flexible graphics with intelligent defaults

Weaknesses

- slow or impossible with large data sets
- non-standard programming paradigms

References

- R Development Core Team (2012). R: A language and environment for statistical computing.
- R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

DATA VISUALIZATION USING R

Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110012

1. Introduction:

Data visualization is the graphical representation of data that turns raw data into clear and meaningful visuals. These visuals like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Important principles for effective data visualization include keeping it simple, accurate, relevant, consistent, and interactive. Key benefits of data visualization include:

- Visual representations of data are often easier to comprehend than raw numbers.
- Identifying trends, outliers, and correlations is easier with visual tools.
- Well-designed visualizations help convey complex data to others in an easily digestible format.

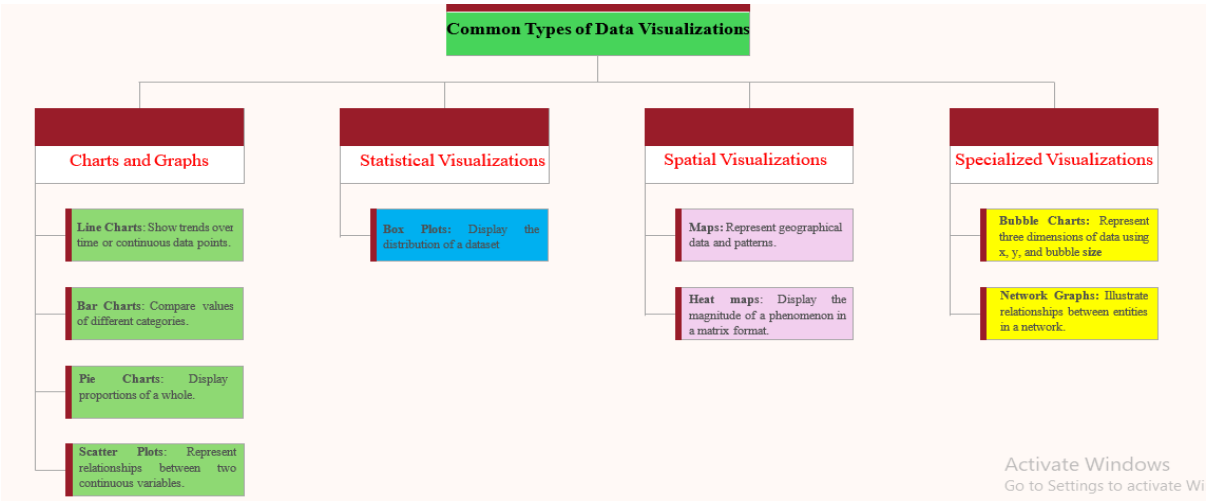
2. Getting Started with R and RStudio

Before diving into data visualization, ensure that you have **R** and **RStudio** installed on your computer. R is a language for statistical computing and graphics. RStudio is an Integrated Development Environment (IDE) for R. These can be downloaded from CRAN and RStudio from RStudio's website. Once installed, launch RStudio, where you can interact with R and visualize data effectively.

3. Basic Plotting in R

R provides a rich set of plotting functions in the **base** package, allowing users to quickly generate a variety of basic plots.

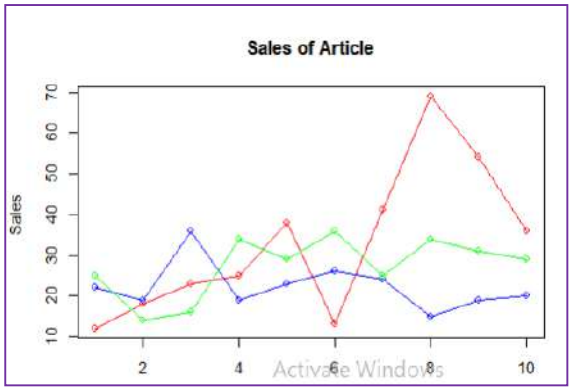
Common Types of Data Visualizations:



3.1 Charts and Graphs:

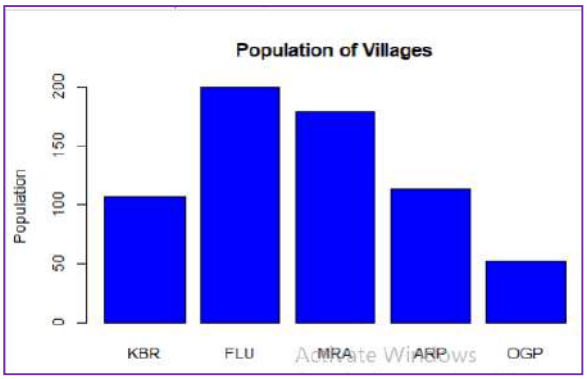
1. **Line Charts:** A line chart visually displays data trends over time using connected data points. It is widely used in various fields for analysing and representing data patterns.

```
# Create the data for the chart
Dia <- c(12,18, 23, 25, 38, 13, 41,69,54,36)
ht <- c(22, 19, 36, 19, 23,26,24,15,19,20)
yd <- c(25, 14, 16, 34, 29,36,25,34,31,29)

# Plot the bar chart.
plot(Dia, type = "o", col = "red",
     xlab = "Month", ylab = "Sales ",
     main = "Sales of Article")

lines(ht, type = "o", col = "blue")
lines(yd, type = "o", col = "green")
```



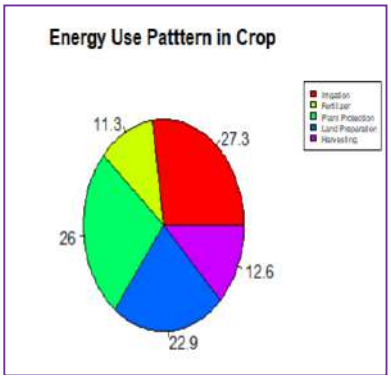
2. **Bar Charts:** A bar chart is a visual representation of data where individual bars represent different categories, and the length of each bar corresponds to the value it represents. It is commonly used to compare and show the relationships between different data sets.

```
# Create the data for the chart
Population <- c(107, 200, 180, 113, 52)
Village <- c("KBR", "FLU", "MRA", "ARP", "OGP")

# Plot the bar chart
barplot(Population, names.arg = Village, xlab ="Village",
        ylab ="Population", col ="blue",
        main ="Population of Villages", horiz = FALSE)
```



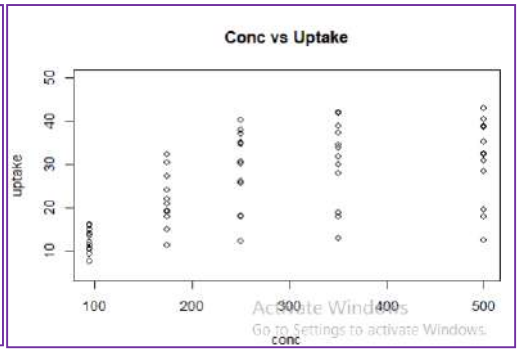
3. **Pie Charts:** A pie chart is a circle divided into sectors to show the proportion of different categories in a dataset. Each sector's size corresponds to the percentage it represents, making it effective for visualizing relative proportions.

```
# Create data for the graph.
data <- c(63, 26, 60, 53, 29)
labels <- c("Irrigation", "Fertilizer", "Plant Protection", "Land Preparation", "Harvesting")

piepercent<- round(100 * data / sum(data), 1)

# Plot the chart.
pie(data, labels = piepercent,
    main = "Energy Use Patttern in Crop", col = rainbow(length(data)))
legend("topright", labels,
       cex = 0.5, fill = rainbow(length(data)))
```



4. **Scatter Plots:** A scatter plot is a graph showing points in a coordinate system, with each point representing a pair of values for two variables. Scatter plots are crucial in statistical analysis for identifying associations and understanding data distribution.

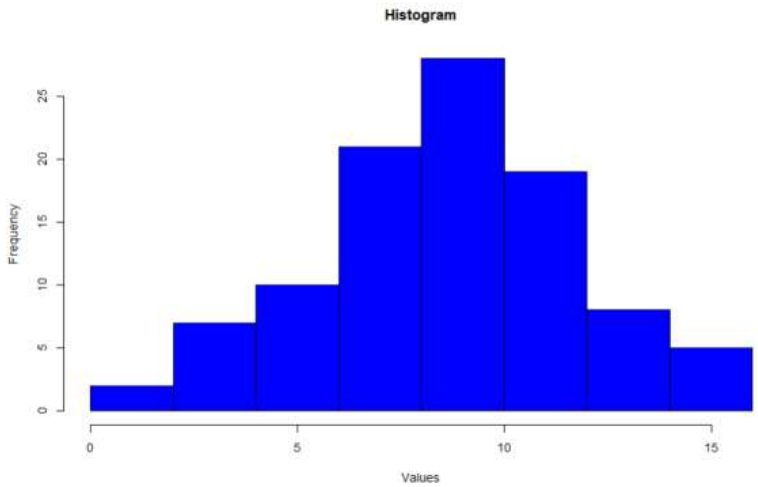
```
# Get the input values.
datam <- CO2[, c('conc', 'uptake')]

# Plot the chart for cars with
plot(x = datam$conc, y = datam$uptake,
     xlab = "conc",
     ylab = "uptake",
     xlim = c(95, 500),
     ylim = c(5, 50),
     main = "Conc vs Uptake"
)
```



3.2 Statistical Visualizations:

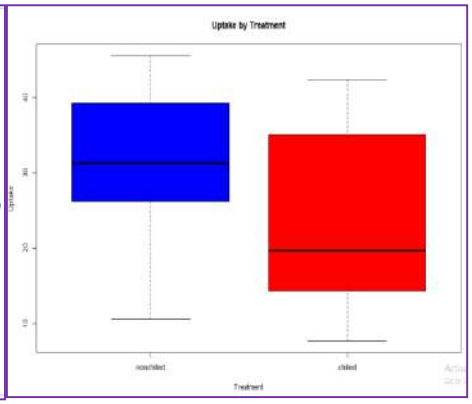
1. **Histogram:** It visually displays the distribution of data by illustrating how values are distributed across different ranges or bins. It provides insights into the central tendency and spread of the data, making it a valuable tool for understanding patterns and trends within a dataset.

```
hist(a, col = "blue", xlab='Values', ylab='Frequency', main='Histogram')
```



2. **Box Plots:** A box plot graphically represents the five-number summary, offering a concise overview of the data's central tendency and spread. The five-number summary includes the minimum, maximum, median (50th percentile), lower quartile (Q1, 25th percentile), and upper quartile (Q3, 75th percentile). In a box plot, a central box spans the interquartile range (from Q1 to Q3), with a line inside marking the median. Lines extend from the box to the smallest and largest observations. Box plots can be oriented either horizontally or vertically. Additionally, a box plot may identify potential outliers. They serve as effective tools for conveying information about the location and variation within datasets, particularly for highlighting changes between different data groups.

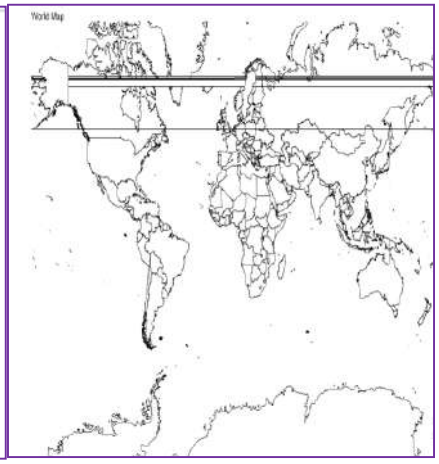
```r
# Load the dataset
data(CO2)

# Set up plot colors
my_colors <- c("blue","red")

# Create the box plot with customized aesthetics
boxplot(uptake ~ Treatment, data = CO2,
        main = "Uptake by Treatment", xlab = "Treatment", ylab = "Uptake",
        col = my_colors, border = "black")

# Add a legend
legend("topright", legend = unique(CO2$Treatment),
       fill = my_colors, border = "black", title = "Treatment")
```



3.3 Spatial Visualizations:

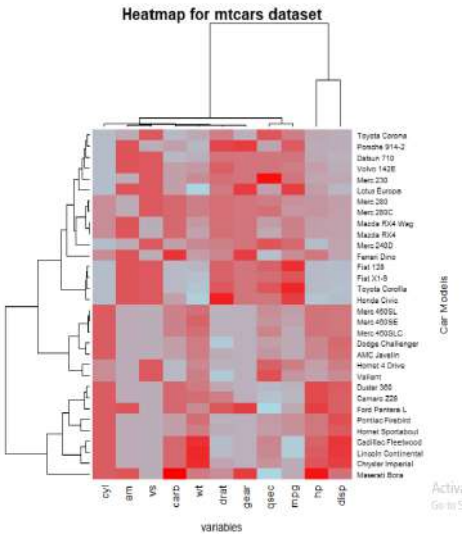
1. **Maps:** Represent geographical data and patterns.

```r
library(ggplot2)
library(maps)
world_map <- map_data("world")
ggplot(world_map, aes(long, lat, group = group)) +
  geom_polygon(fill = "white", color = "black") +
  coord_map("mercator") +
  ggtitle("World Map") +
  theme_void()
```



2. **Heatmaps:** Display the magnitude of a phenomenon in a matrix format.

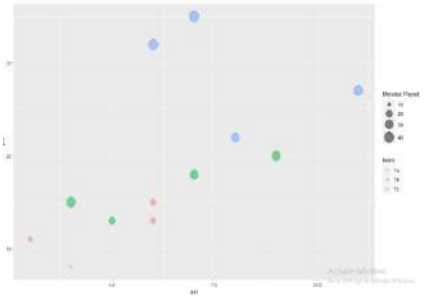
```r
# Heatplot from Base R
# using default mtcars dataset from the R
x  <- as.matrix(mtcars)

# custom colors
new_colors <- colorRampPalette(c("lightblue", "red"))

# plotting the heatmap
plt <- heatmap(x,
               # assigning new colors
               col = new_colors(100),

               # adding title
               main = "Heatmap for mtcars dataset",
               margins = c(5,10),
               # adding x-axis and y-axis labels
               xlab = "variables",
               ylab = "Car Models",
               scale = "column"
)
```



3.4 Specialized Visualizations:

Bubble Charts: Represent three dimensions of data using x, y, and bubble size.

```
#create data frame
fd <- data.frame(team=c('TA', 'TA', 'TA', 'TA','TB', 'TB', 'TB', 'TB','TC', 'TC', 'TC', 'TC'),
                 pts=c(8, 11, 13, 15, 13, 15, 18, 20, 22, 27, 32,35),
                 ast=c(4, 3, 6, 6, 5, 4, 7, 9, 8, 11, 6, 7),
                 min=c(9, 12, 15, 18, 20, 36, 30, 35, 31, 40, 43, 49))

#view data frame
fd
library(ggplot2)

#create bubble chart and color circles based on value of team variable
ggplot(fd, aes(x=ast, y=pts, size=min, color=team)) +
  geom_point(alpha=0.5)  +
  scale_size(range=c(2, 10), name='Minutes Played')
```



4. Conclusion

Data visualization is an essential skill for effective data analysis, as it allows complex data to be communicated in an intuitive and accessible way. R, a powerful programming language for statistical computing, offers an extensive range of visualization tools for various needs. It includes basic charting options like bar graphs, histograms, and line plots, which are ideal for simple data exploration. These tools allow users to easily examine trends, distributions, and relationships in data. With R's straightforward approach, users can quickly generate clear and meaningful visuals, making it a great choice for beginners and those seeking simplicity. This ease of use, combined with its versatility, ensures that data insights are presented in a clear and effective manner, helping analysts and decision-makers interpret information with confidence. R's basic visualization tools lay the foundation for any data story, supporting a wide range of analytical purposes.

ANALYSIS OF SURVEY DATA USING R SOFTWARE

Raju Kumar and Deepak Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

A sample survey is a method for collecting data from or about the members of a population so that inferences about the entire population can be obtained from a subset, or sample, of the population members. In other words, it is a way of collecting information from a random sample of observations drawn from a population of interest using a probability-based sample design. Certain strategies are frequently employed in sample surveys to improve precision and control survey data collection expenses. These methods introduce a complexity to the analysis, which must be taken into account in order to create unbiased estimates and their associated precision levels. This paper gives a quick overview of how these design complications affect sampling variance and then outlines how to use the survey function in software R to analyze sample survey data.

2. Complex Sample Designs

Statistical methods are involved in carrying out a study include planning, designing, collecting data, analyzing, drawing meaning interpretation and reporting of the research findings. Statistical methods for estimating population parameters and their associated variances are based on assumptions about the characteristics and underlying distribution of the observations. Statistical methods in most general-purpose statistical software tacitly assume that the data meet certain assumptions. Among these assumptions are that the observations were selected independently and that each observation had the same probability of being selected. Data collected through surveys often have sampling schemes that deviate from these assumptions. For logistical reasons, samples are often clustered geographically to reduce costs of administering the survey, and it is not unusual to sample households, then subsample families and/or persons within selected households. In these situations, sample members are not selected independently, nor are their responses likely to be independently distributed.

In addition, a common survey sampling practice is to oversample certain population subgroups to ensure sufficient representation in the final sample to support separate analyses. This is particularly common for certain policy-relevant subgroups, such as ethnic and racial minorities, the poor, the elderly, and the disabled. In this situation, sample members do not have equal probabilities of selection. Adjustments to sampling weights (the inverse of the probability of selection) to account for nonresponse, as well as other weighting adjustments (such as post stratification to known population totals), further exacerbate the disparity in the weights among sample members.

In brief, the complications in a complex survey sample result from following:

- **Stratification-** Dividing the population into relatively homogenous groups (strata) and sampling a predetermined number from each stratum will increase precision for a given sample size.

- **Clustering-** Dividing the population into groups and sampling from a random subset of these groups (e.g. geographical locations) will decrease precision for a given sample size but often increase precision for a given cost.
- **Unequal sampling-** Sampling small subpopulations more heavily will tend to increase precision relative to a simple random sample of the same size.
- **Finite population-** Sampling all of a population or stratum results in an estimate with no variability, and sampling a substantial fraction of a stratum results in decreased variability in comparison to a sample from an infinite population. I have described these in terms of their effect on the design of the survey.
- **Weighting -**When units are sampled with unequal probability it is necessary to give them correspondingly unequal weights in the analysis. The inverse-probability weighting has generally the same effect on point estimates as the more familiar inverse-variance weighting, but very different effects on standard errors.

Most standard statistical procedures in software packages commonly used for data analysis do not allow the analyst to take most of these properties of survey data into account unless specialized survey procedures are used. That is standard methods of statistical analysis assume that survey data arise from a *simple random sample* of the target population. Little attention is given to characteristics often associated with survey data, including missing data, unequal probabilities of observation, stratified multistage sample designs, and measurement errors. Failure to do so can have an important impact on the results of all types of analysis, ranging from simple descriptive statistics to estimates of parameters of multivariate models.

3. Impact of Complex Sample Design on Sampling Variance

Because of these deviations from standard assumptions about sampling, such survey sample designs are often referred to as complex. While stratification in the sampling process can decrease the sampling variance, clustering and unequal selection probabilities generally increase the sampling variance associated with resulting estimates. Not accounting for the impact of the complex sample design can lead to an underestimate of the sampling variance associated with an estimate. So while standard software packages can generally produce an unbiased weighted survey estimate, it is quite possible to have an underestimate of the precision of such an estimate when using one of these packages to analyze survey data.

That is, analyzing a stratified sample as if it were a simple random sample will *overestimate* the standard errors, analyzing a cluster sample as if it were a simple random sample will usually *underestimate* the standard errors, as will analyzing an unequal probability sample as if it were a simple random sample.

The magnitude of this effect on the variance is commonly measured by what is known as the design effect. The design effect is the sampling variance of an estimate, accounting for the complex sample design, divided by the sampling variance of the same estimate, assuming a sample of equal size had been selected as a simple random sample. A design effect of unity indicates that the design had no impact on the variance of the estimate. A design effect greater than one indicates that the design has increased the variance, and a design effect less than one indicates that the design actually decreased the variance of the estimate. The design effect can be used to determine the effective sample size, simply by dividing the nominal sample size by the design effect. The effective sample size gives the

number of observations that would yield an equivalent level of precision from an independent and identically distributed (iid) sample.

4. Software Packages R for Survey data analysis

Several packages are available to the public designed specifically for use with sample survey data. However, this lecture will discuss only Software R for analyzing complex surveys. The survey functions for R were contributed by Thomas Lumley, Department of Biostatistics, University of Washington, USA.

Types of designs that can be accommodated

- Designs incorporating stratification, clustering, and possibly multistage sampling, allowing unequal sampling probabilities or weights.
- Simple two-phase designs
- Multiply-imputed data

Types of estimates and statistical analyses that can be done in R

- Mean, Totals, Quantiles, Variance, Tables, Ratios,
- Generalised linear models (e.g. linear regression, logistic regression etc.)
- Proportional hazards models
- Proportional odds and other cumulative link models
- Survival curves
- Post-stratification, raking, and calibration
- Tests of association in two-way tables

Restrictions on number of variables or observations: Only those due to limitations of available memory or disk capacity.

Variance estimation methods: Taylor series linearization and replication weighting.

Platforms on which the software can be run

- Intel computers with Windows 2000 or better
- Mac OS X 10.3 or later
- Linux
- Most Unix systems.

Pricing and terms: Free download. R is updated about twice per year and the survey package is updated as needed.

5. Implementation of survey package in R

First install survey package. The command **svydesign** in **library (survey)** is used for survey data analysis in R, described as below.

svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)

where different arguments of function **svydesign()** are

ids	Formula or data frame specifying cluster ids from largest level to smallest level, ~0 or ~1 is a formula for no clusters.
probs	Formula or data frame specifying cluster sampling probabilities
strata	Formula or vector specifying strata, use NULL for no strata
variables	Formula or data frame specifying the variables measured in the survey. If NULL, the data argument is used.
fpc	Finite population correction
weights	Formula or vector specifying sampling weights as an alternative to prob
data	Data frame to look up variables in the formula arguments
nest	If TRUE, relabel cluster ids to enforce nesting within strata
check.strata	If TRUE, check that clusters are nested in strata

The **svydesign** object combines a data frame and all the survey design information needed to analyse it. These objects are used by the survey modelling and summary functions. The **id** argument is always required, the strata, fpc, weights and probs arguments are optional. If these variables are specified they must not have any missing values.

By default, svydesign assumes that all PSUs, even those in different strata, have a unique value of the id variable. This allows some data errors to be detected. If your PSUs reuse the same identifiers across strata then set nest=TRUE.

The finite population correction (fpc) is used to reduce the variance when a substantial fraction of the total population of interest has been sampled. It may not be appropriate if the target of inference is the process generating the data rather than the statistics of a particular finite population.

The finite population correction can be specified either as the total population size in each stratum or as the fraction of the total population that has been sampled. In either case the relevant population size is the sampling units. That is, sampling 100 units from a population stratum of size 500 can be specified as 500 or as 100/500=0.2.

If population sizes are specified but not sampling probabilities or weights, the sampling probabilities will be computed from the population sizes assuming simple random sampling within strata.

For multistage sampling the id argument should specify a formula with the cluster identifiers at each stage. If subsequent stages are stratified strata should also be specified as a formula with stratum identifiers at each stage. The population size for each level of sampling should also be specified in fpc. If fpc is not specified then sampling is assumed to be with replacement at the top level and only the first stage of cluster is used in computing variances. If fpc is specified but for fewer stages than id, sampling is assumed to be complete for subsequent stages. The variance calculations for multistage sampling assume simple or stratified random sampling within clusters at each stage except possibly the last.

If the strata with one only PSU are not self-representing (or they are, but svydesign cannot tell based on fpc) then the handling of these strata for variance computation is determined by options ("survey.lonely.psu").

Example -Read the api data - Academic Performance Index (api) is computed for all California schools. The full population data in **apipop** are a data frame with 6194 observations on the 37 variables. Read **apipop** data available in survey package

```
data(api)          #This load the api population data apipop
dim(apipop)        # Shows the dimension of the data set
```

The details of 37 variables are

1.	cds	Unique identifier
2.	stype	Elementary/Middle/High School
3.	name	School name (15 characters)
4.	sname	School name (40 characters)
5.	snum	School number
6.	dname	District name
7.	dnum	District number
8.	cname	County name
9.	cnum	County number
10.	flag	reason for missing data
11.	pcttest	percentage of students tested
12.	api00	API in 2000
13.	api99	API in 1999
14.	target	target for change in API
15.	growth	Change in API
16.	sch.wide	Met school-wide growth target?
17.	comp.imp	Met Comparable Improvement target
18.	both	Met both targets
19.	awards	Eligible for awards program
20.	meals	Percentage of students eligible for subsidized meals
21.	ell	`English Language Learners' (percent)
22.	yr.rnd	Year-round school
23.	mobility	percent of students for whom this is the first year at the school
24.	acs.k3	average class size years K-3
25.	acs.46	average class size years 4-6
26.	acs.core	Number of core academic courses
27.	pct.resp	percent where parental education level is known
28.	not.hsg	percent parents not high-school graduates
29.	hsg	percent parents who are high-school graduates
30.	some.col	percent parents with some college

31.	col.grad	percent parents with college degree
32.	grad.sch	percent parents with postgraduate education
33.	avg.ed	average parental education level
34.	full	percent fully qualified teachers
35.	emer	percent teachers with emergency qualifications
36.	enroll	number of students enrolled
37.	api.stu	number of students tested.

Type **summary(apipop)** and see what you get?

The other data sets contain additional variables pw for sampling weights and fpc to compute finite population corrections to variance. **apipop** is the entire population, **apiclus1** is a cluster sample of school districts, **apistrat** is a sample stratified by stype, and **apiclus2** is a two-stage cluster sample of schools within districts. The sampling weights in **apiclus1** are incorrect (the weight should be 757/15) but are as obtained from UCLA. Data were obtained from the survey sampling help pages of UCLA Academic Technology Services, at

http://www.ats.ucla.edu/stat/stata/Library/svy_survey.htm.

The API program and original data files are at http://api.cde.ca.gov/

api00 is API in 2000

mean (apipop$api00)

[1] 664.7126

enroll is number of students enrolled

sum (apipop$enroll, na.rm=TRUE)

[1] 3811472

Here na.rm=TRUE means –logical, Should missing values be removed?

Specifying a complex survey design – use function svydesign ()

[i] Stratified sample

Here we use data set apistrat, see dim(apistrat), c(apistrat[1,]), attach(apistrat) commands etc.

dstrat<- svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)

summary(dstrat)

Stratified Independent Sampling design

svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat, fpc = ~fpc)

Probabilities:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02262	0.02262	0.03587	0.04014	0.05339	0.06623

```
Stratum Sizes:
                E   H   M
obs           100  50  50
design.PSU    100  50  50
actual.PSU    100  50  50
Population stratum sizes (PSUs):
   E    M    H
4421 1018  755
Data variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"
 [7] "dnum"     "cname"    "cnum"     "flag"     "pcttest"  "api00"
[13] "api99"    "target"   "growth"   "sch.wide" "comp.imp" "both"
[19] "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col"
[31] "col.grad" "grad.sch" "avg.ed"   "full"     "emer"     "enroll"
[37] "api.stu"  "pw"       "fpc"
```

Some functions used to compute means, variances, ratios and totals for data from complex surveys are as follows.

svymean () and **svytotal** () functions are use to extract mean and total estimate along with their standard error, specified as below.

svymean(x, design, na.rm=FALSE,deff=FALSE,...)

svytotal(x, design, na.rm=FALSE,deff=FALSE,...)

Arguments

x	A formula, vector or matrix
design	survey.design or svyrep.design object
na.rm	Should cases with missing values be dropped?
rho	parameter for Fay's variance estimator in a BRR design
return.replicates	Return the replicate means?
deff	Return the design effect
object	The result of one of the other survey summary functions
quietly	Don't warn when there is no design effect computed
estimate.only	Don't compute standard errors (useful when svyvar is used to estimate the design effect)
names	vector of character strings

Also see

Svyvar (x, design, na.rm=FALSE,...)

svyratio (x, design, na.rm=FALSE,...)

svyquantile (x, design, na.rm=FALSE,...)

svymean(~api00, dstrat)

	mean	SE
api00	662.29	9.4089

svymean(~api00, dstrat, deff=TRUE)

	mean	SE	DEff
api00	662.29	9.4089	1.2045

svytotal(~enroll, dstrat, na.rm=TRUE)

	total	SE
enroll	3687178	114642

```
#stratified sample, Now try these code for your self
dstrat<-svydesign(id=~1, strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)
summary(dstrat)
svymean(~api00, dstrat)
svyquantile(~api00, dstrat, c(.25,.5,.75))
svyvar(~api00, dstrat)
svytotal(~enroll, dstrat)
svyratio(~api.stu, ~enroll, dstrat)
# coefficients of variation
cv(svytotal(~enroll,dstrat))
```

[ii] One-stage cluster sample

```
dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
summary(dclus1)
svymean(~api00, dclus1, deff=TRUE)
svymean(~factor(stype),dclus1)
svymean(~interaction(stype, comp.imp), dclus1)
svyquantile(~api00, dclus1, c(.25,.5,.75))
svyvar(~api00, dclus1)
svytotal(~enroll, dclus1, deff=TRUE)
svyratio(~api.stu, ~enroll, dclus1)
```

summary(dclus1)

```
1 - level Cluster Sampling design
With (15) clusters.
svydesign(id = ~dnum, weights = ~pw, data = apiclus1, fpc = ~fpc)
Probabilities:
```

```
   Min.      1st Qu.      Median      Mean      3rd Qu.    Max.
0.02954     0.02954      0.02954     0.02954   0.02954   0.02954
Population size (PSUs): 757
Data variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"
 [7] "dnum"     "cname"    "cnum"     "flag"     "pcttest"  "api00"
[13] "api99"    "target"   "growth"   "sch.wide" "comp.imp" "both"
[19] "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col"
[31] "col.grad" "grad.sch" "avg.ed"   "full"     "emer"     "enroll"
[37] "api.stu"  "fpc"      "pw"
```

svymean(~api00, dclus1)

```
              mean     SE
api00       644.17   23.542
```

svytotal(~enroll, dclus1, na.rm=TRUE)

```
              total     SE
enroll      3404940    932235
```

[iii] Two-stage cluster sample

dclus2<-svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)

summary(dclus2)

```
2 - level Cluster Sampling design
With (40, 126) clusters.
svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
Probabilities:
   Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
0.003669    0.037740    0.052840    0.042390    0.052840    0.052840
Population size (PSUs): 757
Data variables:  [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"
 [7] "dnum"     "cname"    "cnum"     "flag"     "pcttest"  "api00"
[13] "api99"    "target"   "growth"   "sch.wide" "comp.imp" "both"
[19] "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col"
[31] "col.grad" "grad.sch" "avg.ed"   "full"     "emer"     "enroll"
[37] "api.stu"  "pw"       "fpc1"     "fpc2"
```

svymean(~api00, dclus2)

	mean	SE
api00	670.81	30.099

svytotal(~enroll, dclus2, na.rm=TRUE)

	total	SE
enroll	2639273	799638

[iv] Two-stage `with replacement'

dclus2wr<-svydesign(id=~dnum+snum, weights=~pw, data=apiclus2)

summary(dclus2wr)

```
2 - level Cluster Sampling design (with replacement)
With (40, 126) clusters.
svydesign(id = ~dnum + snum, weights = ~pw, data = apiclus2)
Probabilities:
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.003669 0.037740 0.052840 0.042390 0.052840 0.052840
Data variables:
 [1] "cds"      "stype"    "name"     "sname"    "snum"     "dname"
 [7] "dnum"     "cname"    "cnum"     "flag"     "pcttest"  "api00"
[13] "api99"    "target"   "growth"   "sch.wide" "comp.imp" "both"
[19] "awards"   "meals"    "ell"      "yr.rnd"   "mobility" "acs.k3"
[25] "acs.46"   "acs.core" "pct.resp" "not.hsg"  "hsg"      "some.col"
[31] "col.grad" "grad.sch" "avg.ed"   "full"     "emer"     "enroll"
[37] "api.stu"  "pw"       "fpc1"     "fpc2"
```

svymean(~api00, dclus2wr)

	mean	SE
api00	670.81	30.712

svytotal(~enroll, dclus2wr, na.rm=TRUE)

	total	SE
enroll	2639273	820261

Reference

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology.

DEVELOPMENT OF R PACKAGE

Pankaj Das

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

R is a powerful statistical programming language widely used in data science, machine learning, and research. One of its key strengths lies in its extensibility, allowing users to develop their own **R packages** to share code, functions, and datasets efficiently. Developing an R package is essential for creating reusable and maintainable code, contributing to the open-source community, and enhancing reproducibility in research. This article provides a step-by-step guide to developing an R package, covering package structure, documentation, testing, and publishing on **CRAN (Comprehensive R Archive Network)** or **GitHub**.

Why Develop an R Package?

Developing an R package offers several benefits:

1. **Code Reusability** – Functions can be easily shared and reused in different projects.
2. **Collaboration** – Team members can work with standardized functions.
3. **Documentation** – Well-structured packages enhance usability and understanding.
4. **Contribution to the Community** – Packages can be published for global use.

Steps to Develop an R Package

1. Setting Up the Package

To create an R package, start by installing the devtools and usethis packages:

```
#r console
install.packages("devtools")
install.packages("usethis")
library(devtools)
library(usethis)
```

Now, create the package structure using:

```
#r console
create_package("path/to/package_name")
```

This command generates a folder structure with necessary files.

2. Understanding Package Structure

An R package consists of:

- **DESCRIPTION** – Metadata about the package (title, author, dependencies).
- **NAMESPACE** – Specifies which functions are exported.

• **R/** – Contains all R scripts with functions.
• **man/** – Stores documentation for functions.
• **tests/** – Includes unit tests for checking function correctness.
• **vignettes/** – Provides long-form documentation and use cases.

3. Writing Functions

Develop functions inside the R/ directory. Example:

```
#r console
# Save this in R/myfunction.R
my_function <- function(x, y) {
  return(x + y)
}
```

4. Documenting Functions

Use **roxygen2** for documentation. Add comments like:

```
#r console
#' Add Two Numbers
#'
#' This function takes two numbers and returns their sum.
#'
#' @param x First number.
#' @param y Second number.
#' @return Sum of x and y.
#' @examples
#' my_function(3, 5)
#' @export
my_function <- function(x, y) {
  return(x + y)
}
```

Run document() to generate documentation:

```
#r console
devtools::document()
```

5. Testing the Package

Testing ensures reliability. Use testthat to create test cases:

```
#r console
usethis::use_testthat()
```

Write tests inside tests/testthat/:

```
#r console
test_that("my_function works correctly", {
  expect_equal(my_function(2, 3), 5)
  expect_equal(my_function(-1, 1), 0)
})
```

Run tests using:

```
#r console
devtools::test()
```

6. Checking and Building the Package

Before releasing, check the package:

```
#r console
devtools::check()
```

To build the package:

r

devtools::build()

7. Publishing the Package

Publishing on GitHub

If you want to share your package on GitHub, use:

```
#r console
usethis::use_git()
usethis::use_github()
```

Users can install your package via:

```
#r console
devtools::install_github("username/package_name")
```

Publishing on CRAN

To publish on CRAN:

1. Check your package with devtools::check()

2. Submit using usethis::use_cran_submission()
3. Follow CRAN guidelines and respond to reviewer feedback.

8. Conclusion

Developing an R package is a structured process that enhances reproducibility, usability, and collaboration in research and data science. By following best practices in documentation, testing, and distribution, you can create robust and valuable R packages for personal or community use.

PYTHON – AN OVERVIEW

Md Ashraful Haque

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

1. Introduction:

Python is the one of the most popular programming languages now-a-days. It is a high-level, interpreted, interactive, object-oriented programming language. Python language was created by Guido van Rossum in 1991 at the National Research Institute for Mathematics and Computer Science in the Netherlands. Python programming language is mainly used for-

- Data handling and visualization
- Analysis of variety of data such as numerical, textual, image, videos, audio etc.
- Performing complex mathematical computations
- Server-side scripting for developing web applications.
- Standalone software development etc.

Why Python?

Python is very easy learn language. It can work in any system irrespective of the operating system. The syntax of python language is very simple and allows programmers to write programs in very few lines. Python runs on an interpreter system, which means that the code is being executed as soon as it is written. And last but not least, that python has a very large and mature community for the developers. There are lots of blogs, tutorials, documents, guide videos available online for python developers.

Python Installation:

Most of the latest computer systems have python already installed. To check if you have python installed on a Windows PC, search in the start bar for Python or run the following on the Command Line (cmd.exe):

C:\your\python\installation\folder>python --version

If not, then one can download the latest version of python (latest version is 3.9.2) from https://www.python.org/downloads/ for the particular operating system and follow the guidelines while installation.

Getting Started with Python:

Any python script or file is saved with .py file extension. Let's write the first python program that prints 'Hello, Everyone!!!'. So, first open a text editor and write the following code in it:

e.g.

```
print("Hello, Everyone!!!")
```

Now save it as 'first.py'. Now open command prompt, go to the python installation folder and type the following command:

C:\your\python\installation\path>python /your/program/path/first.py

The output should read:

Hello, Everyone!!!

Python from Command Line:

In case of python, it is possible to run the code as a command line itself using the command prompt.

Type the following on the Windows, Mac or Linux command line:

C:\your\python\installation\path>python

From there one can write any python code, including our first example from earlier in the:

C:\your\python\installation\path>python

Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>

Which will write "Hello, Everyone!!!" in the command line:

C:\your\python\installation\path>python

Python 3.6.4 (v3.6.4:d48eceb, Dec 19 2017, 06:04:45) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print("Hello, Everyone!!!")

Hello, Everyone!!!

Whenever you are done in the python command line, you can simply type the following to quit the python command line interface:

exit()

Python Syntax:

The major syntactical rules of python programs has been provided below-

Execution of code

a. python can be executed directly from command line.

>>> print("Hello, Everyone!!!")

Hello, Everyone!!!

b. Python can also be executed using a file with '.py' extension

C:\your\python\installation\path>python /your/program/path/first.py

Indentation

The indentation refers to the spaces at the beginning of a program line. Indentation is very important and stricter in python. Python uses indentation as a block of code.

e.g.

```
if 5 > 2:
  print("Five is greater than two!")
```

Comments

In python, comments can be included in the code by using '#' symbol. Comments can be used in the beginning, middle, or in the end pf the code. Comments can be multiline. For multiline comments one can use triple quotes (""").

Variables in Python:

In python, the variables are simple storage structures for storing data values. There is no requirement of *type* declaration for the variables in python. The *type* of any variable can be acquired by *type()* function.

e.g.

```
x = 5
y = "python"
print(type(x))
print(type(y))
```

In python variables names -

- are case sensitive
- Must starts with a letter or underscore
- Can be alphanumeric

Python variables can store different types of data.

Text Type:	str
Numeric Types:	int, float, complex
Sequence Types:	list, tuple, range
Mapping Type:	dict
Set Types:	set, frozenset
Boolean Type:	bool
Binary Types:	bytes, bytearray, memoryview

Operators in Python:

Python divides the operators in the following groups:

Arithmetic operators	+, -, /, *, %, **, //
Assignment operators	=, +=, -=, *=, /=
Comparison operators	==, !=, >, <, >=, <=
Logical operators	And, or, not
Identity operators	is, is not,
Membership operators	in, in not
Bitwise operators	&, |, ^, ~, >>, <<

Data structures in python:

Data Structures are the way of organizing, storing, manipulating, and accessing data in better way. The data structures enable us to can be access and update data in a more efficient manner depending upon the situation. Data Structures are fundamentals of any programming language around which a program is built. There are mainly four types of built-in data structures in python. Python helps to learn the fundamental of these data structures in a simpler way as compared to other programming languages. These data structures are-

- List
- Tuple
- Set
- Dictionary

1. List Data Structures:

List are used to store more than one data in single variable. In python lists are flexible i.e. it can store multiple data type in a single list.

The characteristics of Lists Data Structures in python are:

- Items are indexed (starting from 0)
- Items are ordered
- Items are changeable
- Lists allow duplicate values of items

Creation of List: Lists are created by placing the comma separated items inside the square brackets.

```
## creation of lists
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3]
list3 = [True, False, False]
list4 = [1, 2, 3, "GFG", 2.3]
list5 = [1,2,3,4,4,]
```

Accessing Items from List: Items of the lists can be access by mentioning the index or indices inside the square brackets.

```
## accessing items
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3, 6, 9, 2, 1, 10]
x = list1[0]
print(x)
y = list2[1:4]
print(y)
```

```
## new list
new_list = [1, 2, 3, 'example', 3.132, 10, 30]
#access all elements
print(new_list)
#access index 3 element
print(new_list[3])
#access elements from 0 to 1 and exclude 2
print(new_list[0:2])
#access elements in reverse
print(new_list[::-1])
```

Updating the list: Items in the list at particular position can be updated by mentioning the values in the left-hand side of the assignment operator.

```
list2 = [1, 5, 7, 9, 3, 6, 9, 2, 1, 10]
list2[2] = 34
print(list2)
```

Remove items: Items in the list at particular position can be deleted by del statement.

```
list2 = [1, 5, -12, 9, 3, 6, 9, 2, 1, 10]
del list2[2] print(list2)
```

Some common functions operate on list data structures:

```
## append(): adds an items or a list of items in at the end of a list
list1.append(list2)
## insert(): adds an items at a particular location of a list
list1.insert(1,'mango')
## remove(): deletes an item by its value from a list
list1.remove('banana')
## clear(): deletes all the elements from the
list list1.clear()
## index(): finds the index of the given element in the list
list1.index('mango')
## finds the count of the given element present in the list
list1.count('mango')
#sorted(): temporarily sorts the elements of the list
sorted(list1)
#sort(): permanantly sorts the elements of the list
list1.sort(reverse=True)
```

Some basic operations on list data structures:

```
## Get number of items in a list
n = len(list1)
## Concatenate two lists together
list_new = list1 + list2
## check membership of an item in a list
100 in list2 # (gives true or false)
```

2. Tuple Data Structure:

Tuples are sequence of immutable objects in python. Tuples can store more than one datatype in a single instance of tuple.

The characteristics of Tuple Data Structures in python are:

- Items are indexed (starting from 0)
- Items are ordered
- Items are non-changeable
- Tuples allow duplicate values for the items

Creation of tuples: Tuples are created by placing the comma separated items inside the round brackets or parenthesis.

```
tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3)
tuple3 = (True, False, False)
```

Accessing items from Tuple: Items of the tuple can be access by mentioning the index or indices inside the square brackets.

```
## accessing items
tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3, 6, 9, 2, 1, 10)
x = tuple1[0]
print(x)
y = tuple2[1:4]
print(y)
```

Updating the tuple: Items in the tuples can't be changes once the tuple is created.

```
tuple2 = (1, 5, 7, 9, 3, 6, 9, 2, 1, 10)
tuple2[2] = 34 ## will raise an error print(tuple2)
```

Remove items: Items in the tuple can't be deleted as tuples are immutable. However, del statement can be used to delete whole tuple instead.

```
tup = ('physics', 'chemistry', 1997, 2000)
```

```
print(tup)
del tup
print(tup) ## will raise an error
```

Some basic operations on tuple data structures:

```
## Get number of items in a tuple
n = len(tuple1)
tuple_new = tuple1 + tuple2
## check membership of an item in a tuple
100 in tuple2 # (gives true or false)
```

3. Set Data Structure:

Mathematically, a set is a collection of items in any order. The sets in python are typically used for mathematical operations like union, intersection, difference and complement etc.

The characteristics of Set Data Structures in python are:

- Items are unindexed
- Items are unordered
- Items are non-changeable.
- Sets doesn't allow duplicate values

Creation of Sets: Sets are created by placing the comma separated items inside curly brackets.

```
set1 = {"apple", "banana", "cherry"}
set2 = {1, 5, 7, 9, 3}
set3 = {True, False, False}
```

Accessing items in Sets: Items in the sets can't be access by mentioning the index number. For accessing the items in the Sets one can use any loop structure.

```
Days=set(["Mon","Tue","Wed","Thu","Fri","Sat","Sun"]
for d in Days:
        print(d)
```

Adding and deleting items: In Sets, a new item can be added using *add()* function and an existing item can be deleted by *discard()* function.

```
Days=set(["Mon","Tue","Wed","Thu","Fri","Sat"])
Days.add("Sun")
print(Days)
Days.discard("Mon")
print(Days)
```

Different set operations:

Union of Sets: The union operation on two sets produces a new set containing all the distinct elements from both the sets. In the below example the element "Wed" is present in both the sets. Here, pipe (|) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA|DaysB
print(AllDays)
```

Intersection of Sets: The intersection operation on two sets produces a new set containing only the common elements from both the sets. Here, ampersand (&) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA & DaysB
print(AllDays)
```

Difference of Sets: The difference operation on two sets produces a new set containing only the elements from the rst set and none from the second set. Here, minus (-) operator is used.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Wed","Thu","Fri","Sat","Sun"])
AllDays = DaysA - DaysB
print(AllDays)
```

Compare Sets: We can check if a given set is a subset or superset of another set.

```
DaysA = set(["Mon","Tue","Wed"])
DaysB = set(["Mon","Tue","Wed","Thu","Fri","Sat","Sun"])
SubsetRes = DaysA <= DaysB
SupersetRes = DaysB >= DaysA
print(SubsetRes)
print(SupersetRes)
```

4. Dictionary Data Structure:

Dictionaries are the type of data structure that are used to store data in **key:value** pair. In Dictionary, each key is separated from its value by a colon (:), the items are separated by commas, and the whole thing is enclosed in curly braces.

The characteristics of Dictionaries Data Structures in python are:

- Items are ordered
- Items are changeable
- Dictionary doesn't allow duplicate values

Creation of dictionaries: Dictionaries are created by placing the comma separated **key:values** pairs inside curly brackets.

```
dictionary1 = {"brand": "Ford",
               "model": "Mustang",
               "year": 1964}
x = dictionary1 ["model"]
print(x)
```

Accessing items: Items ca be access by mentioning the key name inside the square bracket.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
print ("dict['Name']: ", dict['Name'])
print ("dict['Age']: ", dict['Age'])
```

Updating Dictionary: One can update a dictionary by adding a new entry or a key-value pair, modifying an existing entry, or deleting an existing entry.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
dict['Age'] = 8;
# update existing entry
dict['School'] = "DPS School"
# Add new entry
print ("dict['Age']: ", dict['Age'])
print ("dict['School']: ", dict['School'])
```

Delete Dictionary Elements: One can either remove individual dictionary elements or clear the entire contents of a dictionary.

```
dict = {'Name': 'Zara', 'Age': 7, 'Class': 'First'}
del dict['Name'] # remove entry with key 'Name'
dict.clear() # remove all entries in dict
del dict  # delete entire dictionary
print ("dict['Age']: ", dict['Age'])
print ("dict['School']: ", dict['School'])
```

Control Structures in Python:

There is mainly one control structure that is *if...else* . The *if...else* structures are used to implement the logical conditions of the program and allow the program to branch based on the evaluation of an expression.

General syntax of *if...else* :

```
if expression :
   statement 1
   statement 2
   ...
```

```
   statement n else:
   statement 1
   statement 2
   ...
statement always executed
```

N.B. Indentation in the control and loop structures are very crucial in case of python programming language.

```
## examples of if..else
## if statement value = 5
threshold= 4
print("value is", value, "threshold is ",threshold)
if value > threshold :
        print(value, "is bigger than ", threshold)
## if..else statement a = 330 b = 200 if b > a:
        print("b is greater than a") else:   print("error")
```

Nested control structures: The *if..else* structures can be used in nested manner by using *elif* statement.

```
## nested if.. statements
### if ... elif ... else ...
a = 5
b = 4
print("a = ", a, "and b = ", b)
if a > b :
   print(a, " is greater than ", b)
elif a == b :
   print(a, " equals ", b)
else :
   print(a, " is less than ", b)
```

Loop Structures in Python:

In python generally two types of loop structures are used: while loop and for loop.

1. *while* loop:

With the 'while' loop, a set of statements can be executed repeatedly long as a condition is true . For the loop to terminate, there has to be some termination criteria mentioned in the code which will potentially change the condition and stop the iteration.

```
   ## Simple example
```

```
i=1
while i < 6:
    print(i)
    i = i + 1
## sum of n numbers using a while loop
n = 10
cur_sum = 0
i = 1
while  i <= n :
   cur_sum = cur_sum + i
   i = i + 1
print("The sum of the numbers from 1 to", n, "is ", cur_sum)
```

Points to note:

- Here, the conditional clause (i <= n) in the *while* statement can be anything which would return a boolean value of either *True* or *False* upon execution.
- Initially *i* has been set to 1 (before the start of the loop) and therefore the condition is *True*.
- The clause can be made more complex by using parentheses, *and* and *or* operators amongst others
- The statements after the *while* clause are only executed if the condition evaluates as *True*.
- Within the statements after the *while* clause there should be something which potentially will make the condition evaluate as *False* next time around. If not the loop will never end.
- In this case the last statement in the loop changes the value of *i* which is part of the condition clause, so hopefully the loop will end.

2. *for* loop:

A *for* loop is used for iterating over a sequence (that is either a list, a tuple, a dictionary, a set, or a string) for executing a set of statements. The difference between *while* and *for* loop is that in *for* loop we know that at the outset how often the statements in the loop will be executed, we don't have to rely on a variable being changed within the looping statements as in *while* loop.

General syntax of *for* loop:

```
for variable_name in some_sequence :
   statement1
   statement2
   ...
   statementn
## simple example
```

```
for i in [1,2,3] :
   print(i)
print("\nExample 1\n")
fruits = ["apple", "banana", "cherry"]
for x in fruits:
 print(x)

print("\nExample 2\n")
for x in "banana":
 print(x)

print("\nExample 3\n")
for name in ["Tom", 42, 3.142] :
   print(name)

print("\nExample 4\n")
for i in range(10) :
   print(i)

print("\nExample 5\n")
longString = "The quick brown fox jumped over the lazy sleeping dog"
for word in longString.split() :
   print(word)
```

Python Functions:

A function is a block of code that contains a set of statements and runs only when it is called explicitly. One can pass data, known as parameters, into a function. A function can return data as a result.

e.g.

```
def my_function(str):
  print(str + "! Welcome to the class.")
my_function("Bob")
```

Packages in Python:

Package or module is a python object with arbitrarily named attributes that one can bind and reference. Packages allows us to logically locate the python code. Simply a package or module is file containing a set of python codes. Packages are also referred as library

Packages or modules or libraries can be imported by using the *'import'* keyword.

e.g.

```
import os
import sys
```

PIP is a package manager available in python. PIP is used to install, upgrade, or uninstall a packages in python environment.

C:\your\python\installation\path>pip install numpy

Some important packages or modules in Python:

NumPy:

NumPy is python library or packages used for working with arrays. NumPy was created by Travis Oliphant in 2005 and it is open source.

In python, the concepts of arrays is served by the List data structure but it is too slow in processing. NumPy provides a 50x faster access speed for the array objects in python than the List. NumPy has a lots of applications in the domain of -

- Arrays
- Matrices
- Linear Algebra
- Fourier Transformation

Creating Arrays

The object of NumPy that deals with the arrays is known as '*ndarray'*. One can create a *'ndarray'* object by using *array()* function. One can pass any type of array-like object in the *array()* function.

e.g.

```
import numpy as np
array_var = np.array([1, 2, 3, 4, 5])
```

Array can be of 0, 1, 2 or 3 dimensions.

e.g.

```
import numpy as np
array0 = np.array(42) #0 dimension
array1 = np.array([1, 2, 3, 4, 5, 6, 7, 8]) # 1 dimension
array2 = np.array([[1, 2, 3], [4, 5, 6]]) # 2 dimension
array3 = np.array([[[1, 2, 3], [4, 5, 6]], [[1, 2, 3], [4, 5, 6]]]) #3 dimension
```

Accessing Array elements

Array elements can be accessed by its index number

```
print(array1 [2]) #accessing the 3rd item from the array 'array1'
```

Slicing an Array

Slicing in python means taking elements from one given index to another given index.

print(array1 [1:3]) #slicing from 2nd item to the 4th element

print(array1 [2:]) #slicing from 3rd item to the last element

print(array1 [:6]) #slicing from beginning to the 5th element

Properties and functions:

dtype- returns the type of values stored in the array object

shape- gives the number of elements in each dimension of the array object

reshape– allows to change the shape of the array either by adding adding/removing dimensions or changing the number of elements in each dimension

concatenate()- joins two or more arrays axis wise.

array_split()– splitting an array into two or more parts

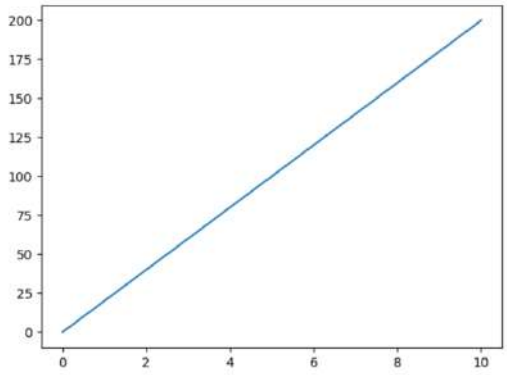
Matplotlib:

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter. Matplotlib is open source and we can use it freely.

Most of the Matplotlib utilities lies under the pyplot submodule, and are usually imported under the plt alias.

e.g. Draw a line in a diagram from position (0,0) to position (10, 200):

```
import matplotlib.pyplot as plt
import numpy as np
xpoints = np.array([0, 10])
ypoints = np.array([0, 200])
plt.plot(xpoints, ypoints)
plt.show()
```



Properties and functions:

marker- keyword argument to emphasize each point in the plot

linestyle/ls- keyword argument to change the style of the plotted line

xlabel()- functions for setting a label for x-axis

ylabel()- function for setting a label for y-axis

title() - function for giving the title for the plot

grid() -function to add grid lines to the plot

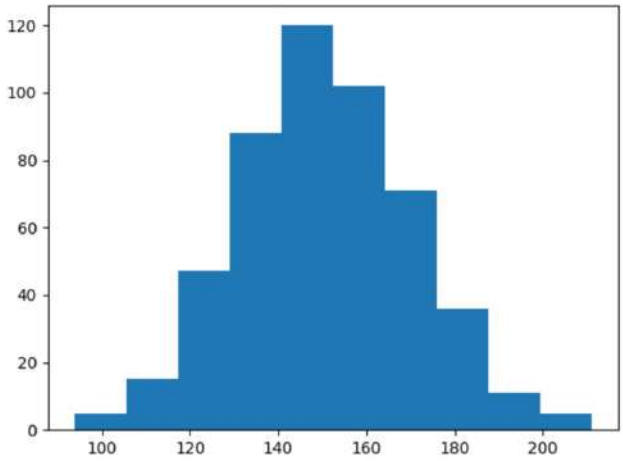
scatter()-function to draw a scatter plot

bar()- function to draw bar graphs

hist()- function to create histograms

e.g.

```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(150, 20 , 250)
plt.hist(x)
plt.show()
```



Pandas:

Pandas is a one of the most popular python package providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word 'Panel Data' – an Econometrics from Multidimensional data. Pandas is well suited for many different kinds of data:

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations

There are mainly two data structures of pandas which handle the majority of typical use cases in finance, statistics, social science and Engineering are Series (1-dimensional) and DataFrame (2-dimensional).

DataFrame

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Features of DataFrame:

- Potentially columns are of different types
- Size – Mutable
- Labelled axes (rows and columns)
- Can Perform Arithmetic operations on rows and columns

e.g.1

```
import pandas as pd
data = [1,2,3,4,5]
df = pd.DataFrame(data)
print df
```

```
   0
0  1
1  2
2  3
3  4
4  5
```

e.g. 2

```
import pandas as pd
data = [['Alex',10],['Bob',12],['Clarke',13]]
df = pd.DataFrame(data,columns=['Name','Age'],dtype=float)
print df
```

```
    Name   Age
0   Alex   10.0
1   Bob    12.0
2   Clarke 13.0
```

Importing data files using pandas

Pandas provides the means for datafiles to be imported to the python environment. External files in any format (.csv, .xls, .txt, .pdf, etc.) can be imported using pandas.

e.g. 1: .csv file can be imported by read_csv() function

```
data = pd.read_csv('/content/sample_data/california_housing_test.csv')
```

e.g. 2: .xls file can be imported by read_excel() function

```
data = pd.read_excel('/content/sample_data/shishamharvesteddata.xls')
```

Measure of central tendency

Mean, Median and Mode of the dataset can be calculated using mean(), median() and mode() functions available in Pandas

e.g.:

```
## mean
```

```
data[].mean()
## median
data[].median()
## mode
data[].mode()
```

Description statistics

Description statistics can be calculated by describe()function available in Pandas

e.g.:

```
data[['dbhcm','Branchkg','Stemkg']].describe()
```

output:

	dbhcm	Branchkg	Stemkg
count	42.000000	42.000000	42.000000
mean	18.927701	27.347262	91.985714
std	4.520851	14.871299	36.560946
min	10.828025	7.630000	20.640000
25%	16.037675	16.015000	66.947500
50%	19.135000	24.550000	96.705000
75%	21.702500	35.378750	114.047500
max	29.681529	70.235000	171.460000

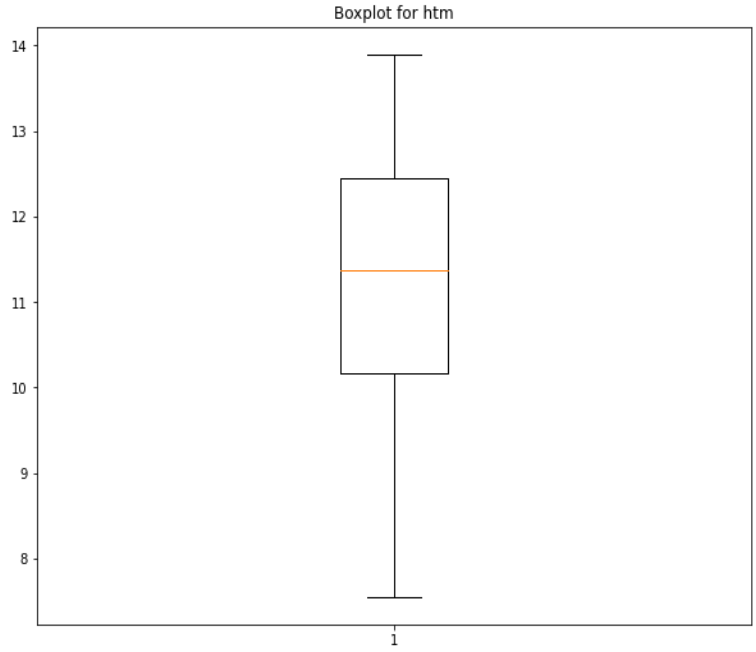
Boxplot

The boxplots can be drawn with the help of pyplot.boxplot function available with matplotlib.

e.g.:

```
## Boxplot
from matplotlib import pyplot as plt
fig = plt.figure(figsize=(10,8))
plt.boxplot(data['htm'])
plt.title('Boxplot for htm')
plt.show()
```

Output:



References:

1. https://colab.research.google.com/github/tensorflow/examples/blob/master/courses/udacity_intro_to_tensorflow_for_deep_learning/l01c01_introduction_to_colab_and_python.ipynb#scrollTo=F8YVA_634OFk
2. https://docs.python.org/3/tutorial/
3. https://numpy.org/
4. https://pandas.pydata.org/
5. https://www.guru99.com/python-tutorials.html
6. https://www.programiz.com/python-programming
7. https://www.tutorialspoint.com/python/index.htm
8. https://www.w3schools.com/python/default.asp

SPSS – AN OVERVIEW

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

SPSS is a widely used software package for statistical analysis in social science. The original SPSS manual (Nie *et al.*, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Originally it is an acronym of *Statistical Package for the Social Science* but now it stands for *Statistical Product and Service Solutions*. The current versions (2015) are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with rather simple instructions. This package of programs is available for both personal as well as mainframe computers. SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation. SPSS integrates complex data and file management, statistical analysis and reporting functions. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

Some versions of SPSS released in recent years are

- SPSS Statistics 17.0.1 - December 2008
- PASW Statistics 17.0.3 - September 2009
- PASW Statistics 18.0, 18.0.1, 18.0.2, 18.0.3
- IBM SPSS Statistics 19.0 - August 2010
- IBM SPSS Statistics 19.0.1, 20.0, 20.0.1, 21.0

Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services). Purpose of this chapter is to introduce the basic features of the SPSS and also to provide some basic statistical analysis using this software.

2. Key features of SPSS

Some of the key features of SPSS are

- It is easy to learn and use with its pull-down menu features
- It includes a full range of data management system and editing tools
- It offers comprehensive range of plotting, reporting and presentation features.
- It provides in-depth statistical analysis capabilities

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary stored in the data file) are features of the base software. There are varieties of statistics included in the base software. Some of the important statistics are:

Descriptive statistics: Cross tabulation, Frequencies, Descriptives, Explore, Descriptive Ratio Statistics etc.

Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests etc.

Prediction for numerical outcomes: Linear regression, Multiple Regression

Prediction for identifying groups: Factor analysis, Cluster analysis (two-step, K-means, hierarchical), Discriminant analysis etc.

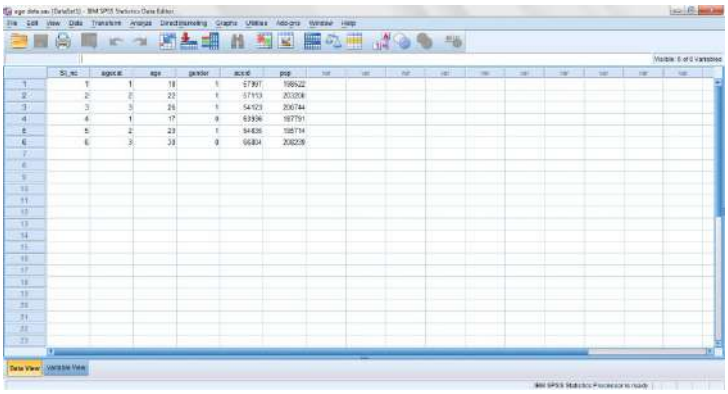
3. Basic features of SPSS

SPSS makes statistical analysis manageable for the naive user and convenient for the experts. There are a number of different types of windows in SPSS. The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file.

Data Editor: This graphical user interface displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when an SPSS session is started. This editor has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window.

- ✓ **Data view**: Displays the actual data values or defined value labels. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. One can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases;
- ✓ **Variable view**: Displays variable definition information contained or metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. One can modify variable properties in the Variable view for example, add and delete variables, change the order of variables etc.

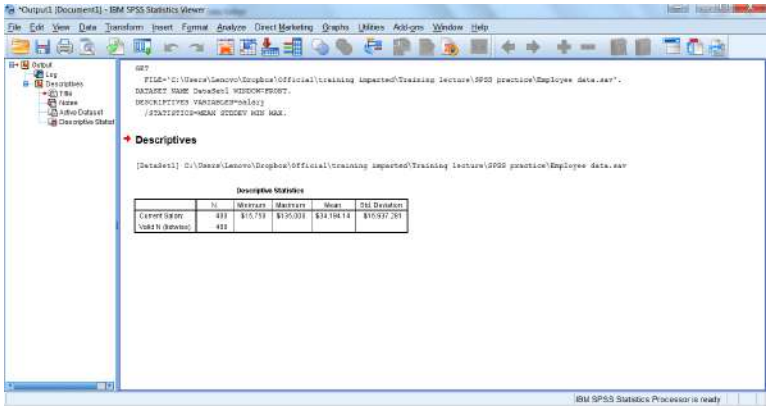
Extension of the saved data file will be ".sav".



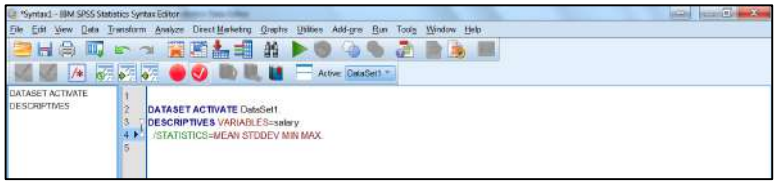
Viewer: All results, tables, and charts performed by different statistical analysis are displayed in the Viewer. Extension of the saved output file will be ".spv". One can use the Viewer to browse results, show or hide selected tables and charts, change the display order of results by moving selected items or move items between the Viewer and other applications. The output presented in Viewer can be edited and saved for later use. A

Viewer window opens automatically the first time a procedure is run that generates output. The Viewer is divided into two panes:

- ✓ The left pane contains an outline view of the contents. One can click an item in the outline to go directly to the corresponding table or chart.
- ✓ The right pane contains statistical tables, charts, and text output.



Syntax Editor: The pull-down menu interface generates command syntax: this can be displayed in the output. These command syntax can also be pasted into a syntax file in a syntax window using the "paste" button present in each menu. One can then edit the command syntax to utilize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions. Extension of the saved syntax file will be ".sps". Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses.



Pivot Table Editor: The results from most statistical procedures are displayed in pivot tables. These pivot tables outputs can be modified in many ways with pivot table editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results. Changing the layout of the table does not affect the results. Instead, it's a way to display information in a different or more desirable manner.

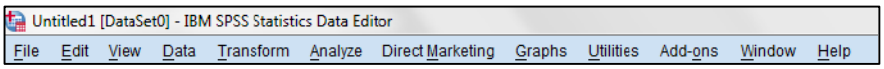
Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

Chart Editor: High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes, switch the horizontal and vertical axes, rotate 3-D scatterplots, and even change the chart type.

Script Window: It provides the opportunity to write full-blown programs, in a BASIC-like language. It is a text editor for syntax composition. Extension of the saved script file will be ".sbs"

Many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Many of the tasks that are to be performed with SPSS start with **menu** selections. Each window has its own

menu bar with menu selections appropriate for that window type. The menu options available in SPSS are



Most menu selections open dialog boxes. These dialog boxes can be used to select variables and various options for analysis. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in sub-dialog boxes. All these above mentioned options have further sub-options. The three dots after an option term (...) on a drop-down menu, such as **Define Variable...** option in Data option, signifies that a dialog box will appear when this option is chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facing arrowhead after an option term indicates that a further submenu will appear to the right of the drop-down menu. An option with neither of these signs means that there are no further dropdown menus to select. There are five standard command push buttons in most dialog boxes.

OK: It runs the procedure. After the variables and additional specifications are selected, click OK to run the procedure.

Paste: It generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset: It deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel: It cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help: It contains information about the current dialog box.

Basic Steps in Data Analysis using SPSS

- **Get data into SPSS**. There are several ways to get data in the SPSS. One can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter data directly in the Data Editor.
- **Select a procedure**. Select an appropriate procedure from the menus in order to perform appropriate analysis on the data file and calculate statistics or create charts.
- **Select the variables for the analysis**. The variables in the data file are displayed in a dialog box for the procedure.
- **Run the procedure**. Results are displayed in the Viewer.

4. Entering and Editing Data

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view.

To be able to retrieve a file, the file must be saved with a proper name. The default extension name for saving files is **sav**. To save this file on hard disk, we carry out the following sequence:

> **File →Save As...** [opens **Save Data As** dialog box]→box under **File Name:** delete the asterisk and type file name →**OK**

The output file can also be printed and saved. The extension name for output file is .**spo**.

To retrieve this file, use the following procedure:

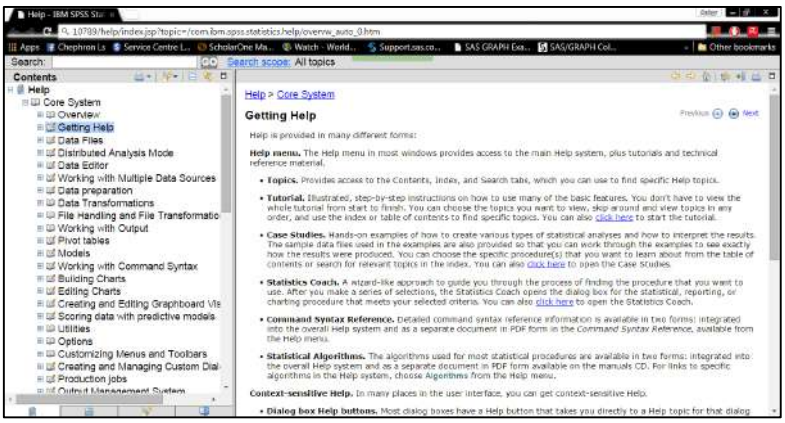
> **File→Open→Data...**[opens the **Open Data File** dialog box] →choose drive from options listed →type name under **File Name: → OK**

5. Statistical Procedures

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyse it. The **Analyse** option has the following sub options:

Reports, Descriptive Statistics, Tables, Compare means, General Linear model, Mixed Models, Correlate, Regression, Loglinear, Neural Networks, Classify, Dimension Reduction, Scale, Non parametric tests, Forecasting, Time Series, Survival, Multiple response, Missing value analysis, Multiple imputation, Complex samples, Quality control, ROC curve.

Help topics available at IBM SPSS Statistics is so enriched that it helps naive users to manage their desired statistical analysis efficiently.



Some of the important statistical analysis options are described in detail as follows

5.1. Descriptive Statistics:

This submenu provides techniques for summarising data with statistics, charts, and reports. This is most useful for providing useful descriptive statistics for different types of dataset. There various sub-sub menus under this submenu.

Frequencies option helps in generating information about the relative frequency of the occurrence of each category of a variable. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptives option carry out statistical analysis that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Explore produces and displays summary statistics for all cases or separately for groups of cases. Several other additional features like boxplots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained using this submenu.

Crosstabs is used to carry out cross-tabulation in order to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests.

P-P plots provides the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Q-Q plots provide the quantiles of a variable's distribution against the quantiles of the normal distribution.

5.2. Compare Means:

This submenu provides techniques for testing differences among two or more means for both independent and related samples.

Means computes summary statistics for a variable when the cases are subdivided into groups based on their values.

One-sample t test procedure tests whether the mean of a single variable differs from a specified constant.

Independent sample t test is used to test if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the *One-way ANOVA* option could be used.

Paired sample t test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly non-normal you should consider one of the nonparametric tests.

One-way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through *Post Hoc Multiple Comparison option* which consist of the options like *Least-significant difference, Duncan's multiple range test, Scheffe* etc. The data obtained using completely randomised design can be analysed through this option.

5.3. General Linear Model:

This submenu provides techniques for testing univariate and multivariate Analysis of Variance models, including repeated measures.

Univariate sub-option could be used to analyse the experimental designs like Completely randomised design, Randomised block design, Latin square design, Designs for factorial

experiments etc. The covariance analysis can also be performed and alternate methods for partitioning sums of squares can be selected.

Multivariate analyses analysis of variance and analysis of covariance designs when there are two or more correlated dependent variables. Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, one can test whether verbal and mathematical test scores are related to instructional method used, sex of the subject, and the interaction of method and sex. This procedure should be used only if there are several dependent variables which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject. Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

5.4. Correlate:

This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate calculates matrices of Pearson product-moment correlations, and of Kendall and Spearman nonparametric correlations, with significance levels and optional univariate statistics. The correlation coefficient is used to quantify the strength of the linear relationship between two variables. The *Pearson correlation coefficient* should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly non-normal.

Partial calculates *partial correlation coefficients* that describe the relationship between two variables, while adjusting for the effects of one or more additional variables. If the value of a dependent variable from a set of independent variables is to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted.

Distances calculates statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, etc.

5.5. Regression:

This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two stage least squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic

regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Curve Estimation produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. One can also save predicted values, residuals, and prediction intervals as new variables.

Logistic estimates regression models in which the dependent variable is dichotomous. If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables. In recent versions there are two options **Binary Logistic** as well as **Multinomial Logistic.**

Probit performs probit analysis which is used to measure the relationship between a response proportion and the strength of a stimulus. For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomous buy/not buy, alive/dead and several groups of subjects are exposed to different levels of some stimulus.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained. The nonlinear regression procedure can be used if one knows the equation whose parameters are to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively. If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option. If the variance of the dependent variable is not constant for all of the values of the independent variable, weights which are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution. The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option. For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

5.6. Classify:

This submenu provides cluster and discriminant analysis.

Two Step Cluster performs Two Step Cluster Analysis procedure which is an exploratory data analysis tool designed to reveal natural clustering within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques. The Log-

likelihood and Euclidean Distance Measures are used as the similarity measure between two clusters.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters. The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics. If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables. If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

Nearest Neighbor performs Nearest Neighbor Analysis for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

5.7. Dimension Reduction:

This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Correspondence Analysis analyses correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Distances computes many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider the characteristics of the data.

15.8. Nonparametric Tests:

This submenu provides nonparametric tests for one sample, or for two and more paired or independent samples.

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The

expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten (p=0.1), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs is used to test whether the two values of a dichotomous variable occur in a random sequence. The runs test is appropriate only when the order of cases in the data file is meaningful.

1-Sample K-S is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson**. Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.

2-Independent Samples is used to compare the distribution of a variable between two non-related groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz runs tests sorts the data values from smallest to largest and then performs a runs test on the group's numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median tests counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Wilcoxon and Sign tests are nonparametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test. *McNemar's test* is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. *The Friedman test* is a nonparametric alternative to a single-factor repeated measures analysis of variance. You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale. *Cochran's Q test* can be used to test whether several dichotomous variables have the same mean.

5.9. Forecasting:

This submenu provides create models, seasonal decomposition, spectral analysis, autocorrelations, cross-correlations etc.

Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Spectral analysis calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series) as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

15.10. Complex Samples:

This submenu provides procedures for Sampling from Complex Designs. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

18. Graphs

The Chart Builder available in Graph menu allows to build charts from predefined gallery charts or from the individual parts (for example, axes and bars). Build a chart by dragging and dropping the gallery charts or basic elements onto the canvas, which is the large area to the right of the Variables list in the Chart Builder dialog box.

Legacy Dialogs submenu provides following graph options

Bar generates a simple, clustered, or stacked bar chart of the data.

3-D Bar Charts generates bar graph in 3-dimensional axis.

Line generates a simple or multiple line chart of the data.

Area generates a simple or stacked area chart of the data.

Pie generates a simple pie chart or a composite bar chart from the data.

High-Low plots pairs or triples of values, for example high, low, and closing prices.

Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.

Error Bar Charts plot the confidence intervals, standard errors, or standard deviations of individual variables.

Scatter/dot generates a simple or overlay scatterplot, a scatterplot matrix, or a 3-D scatterplot from the data.

Histogram generates a histogram showing the distribution of an individual variable.

19. Exercises

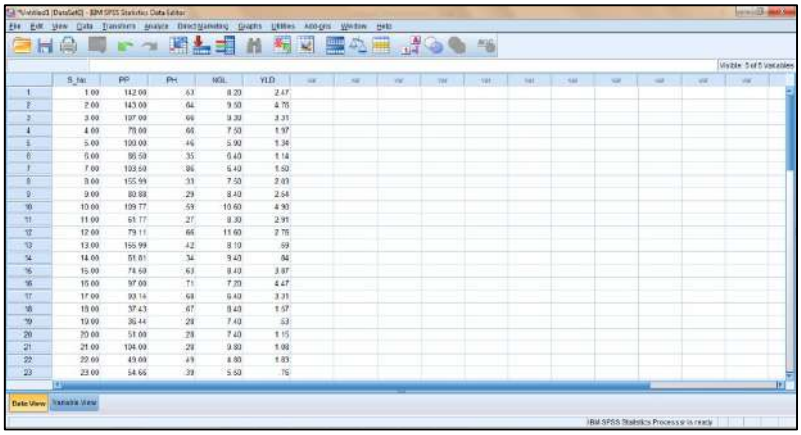
Exercise 1. The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570
5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

Source: Design Resources Server. ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).

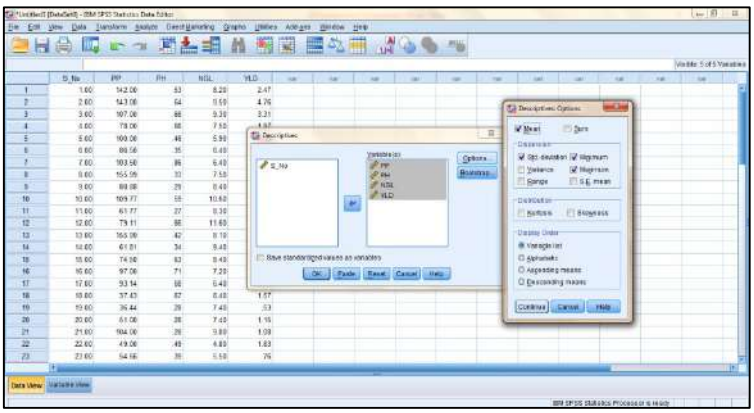
1. Find mean, standard deviation, minimum and maximum values of all the characters.
2. Find correlation coefficient between each pair of the variables.
3. Give a scatter plot of the variable PP with dependent variable yield.
4. Fit a multiple linear regression equation where yield is dependent variable whereas all other characters as independent variables.

At first enter the entire data in the data editor as given below,

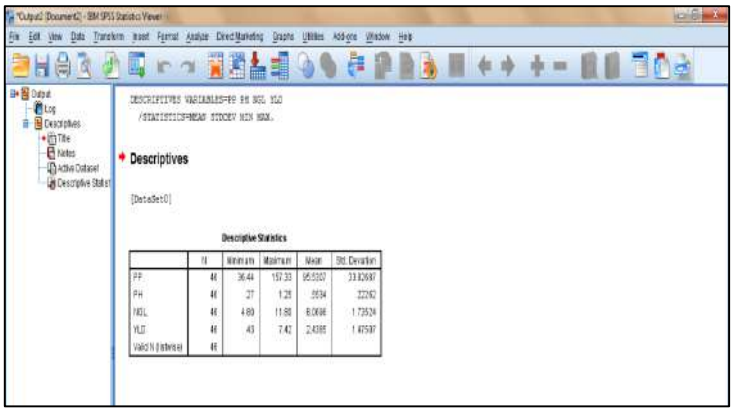


There are several ways to answer the Q no. 1in SPSS. Commands following first way is as follows,

Analyze → Descriptive Statistics → Descriptives…→ Put PP, PH, NGL, YLD in the variables list→ Choose appropriate options from Options tab→Press Continue → Press Ok

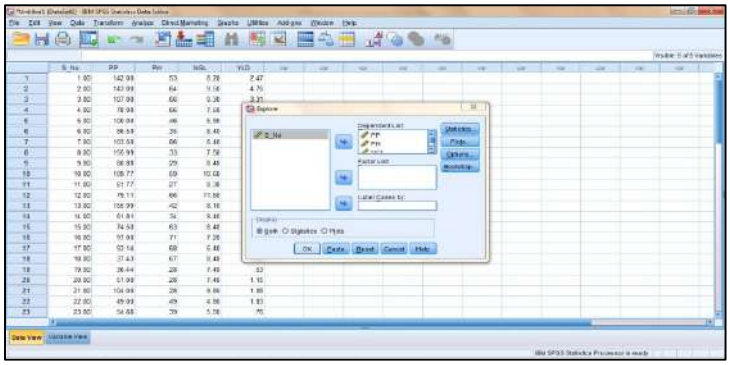


Output:

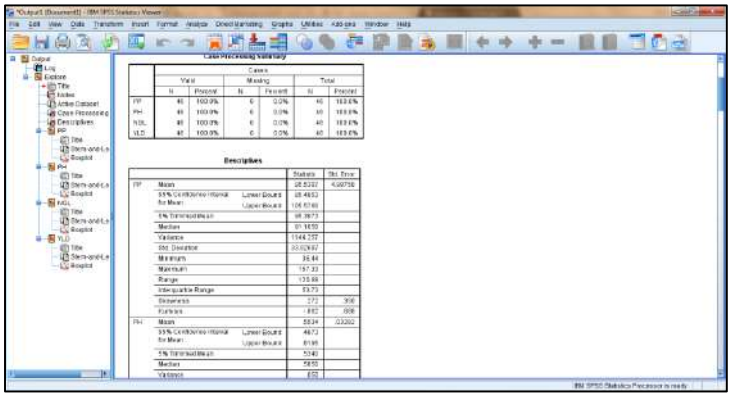


Another way:

Analyze → Descriptive Statistics → Explore…→ Put PP, PH, NGL, YLD in the Dependent list→ Choose both Statistics and plot→ Press Ok

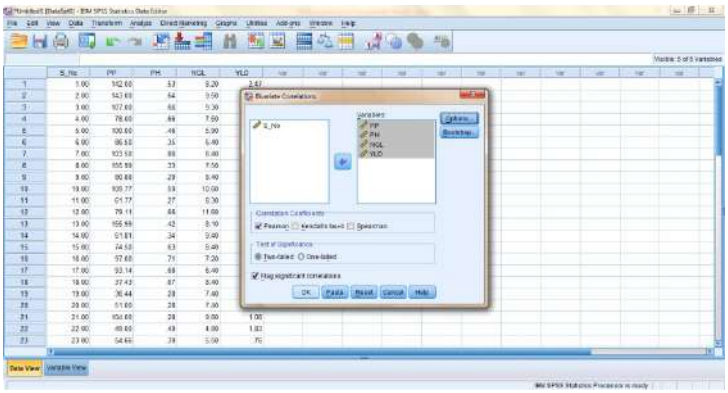


Output:

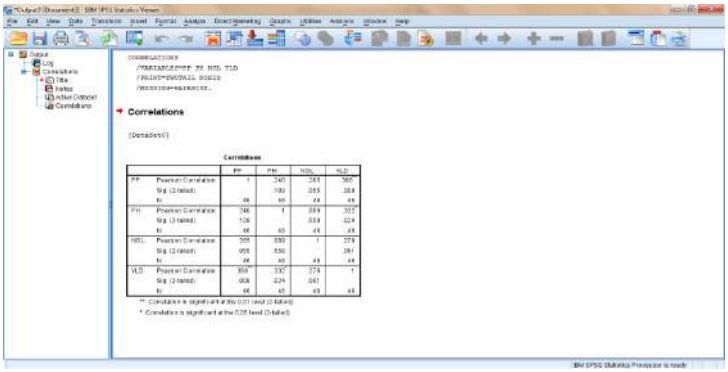


To answer Q no. 2 follow the following steps

Analyze → Correlate → Bivariate→ Put PP, PH, NGL, YLD in the Valiables list→ Choose Pearson's correlation coefficient→ Press Ok

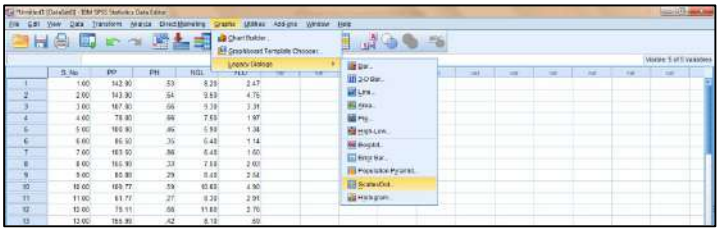


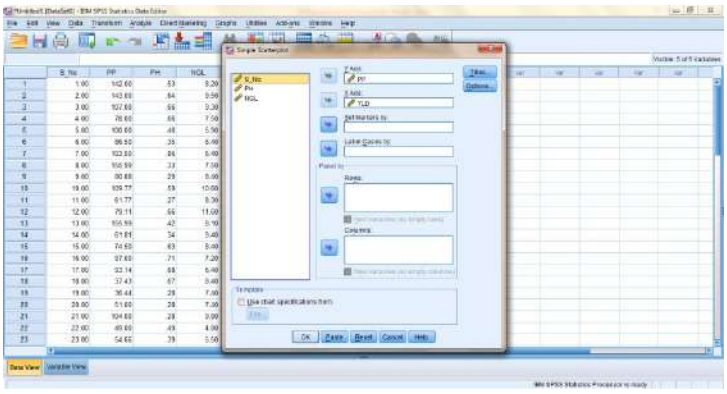
Output:



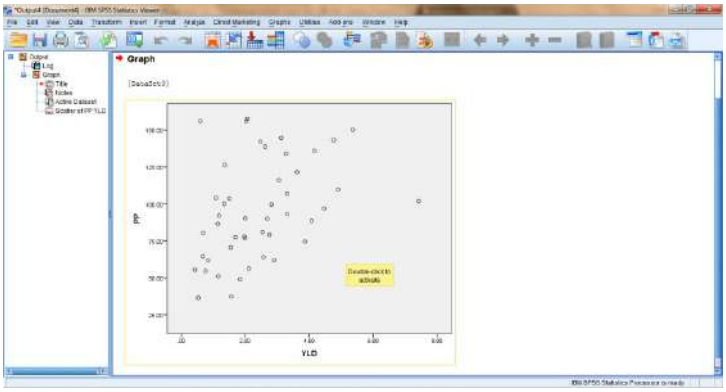
To give the scatter plot of the variable PP with dependent variable yield use following steps:

Graphs → Legacy dialogs→ Scatterplot→ Put PP at Y axis and YLD at X axis→ Ok





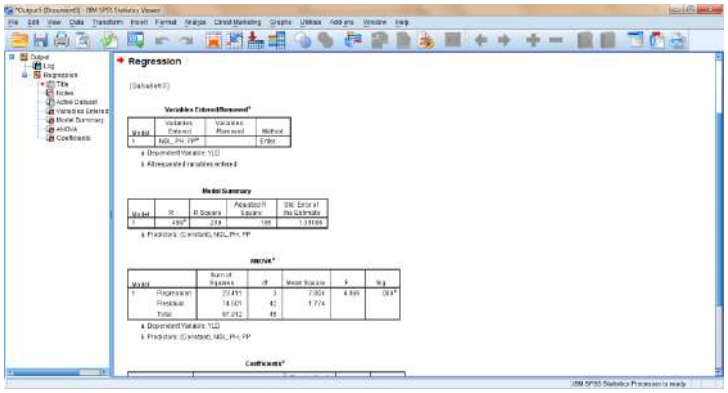
Output:



To fit a multiple linear regression equation taking yield as dependent variable and all other characters as independent variables perform following steps

Analyze → Regression → Linear → Put Yld in Dependent variable and PP, PH, NGL in independent variable list → Press Ok

Output:



Exercise 2. Practical exercise using SPSS for Survey Data

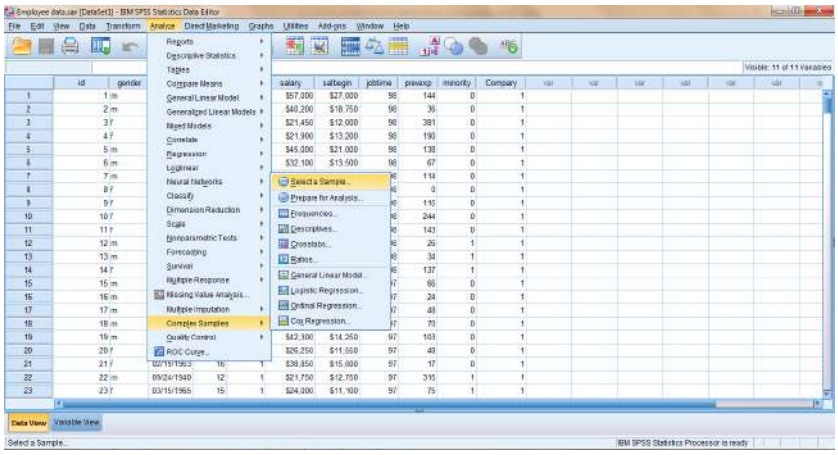
In this section, a practical exercise is provided which has analyzed using a popular statistical software, SPSS. For illustration purpose, we are going to use "Employee data" from the Sample folder of SPSS available at C:\Program Files (x86)\IBM\SPSS\Statistics\20\Samples\English. In addition a new variable "Company" has been added to the dataset which is having values from 1,2,…,10. Thus, there are 400 data-points clustered into 10 clusters each of size 40 usu. This dataset has been considered as population used for further illustration (available at https://www.dropbox.com/s/rxxccpuk3iieepa/Employee%20data.sav?dl=0).

The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind. The Complex Samples option allows to select a sample according to a complex design and incorporate the design specifications into the data analysis.

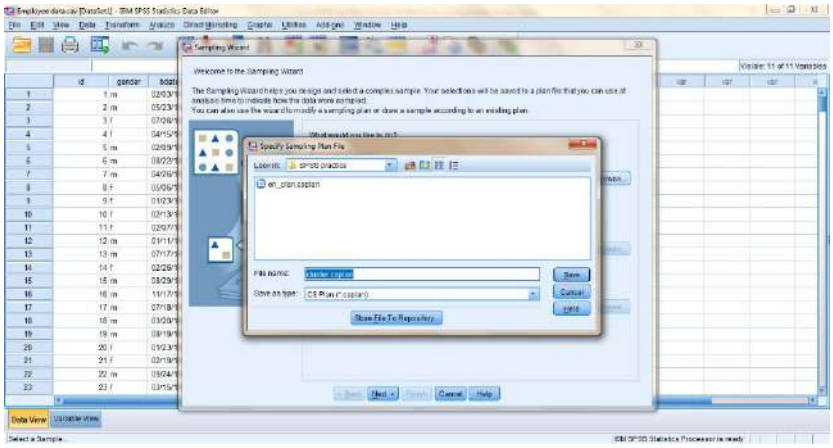
Creating a New Sample Plan

1. From the menus choose

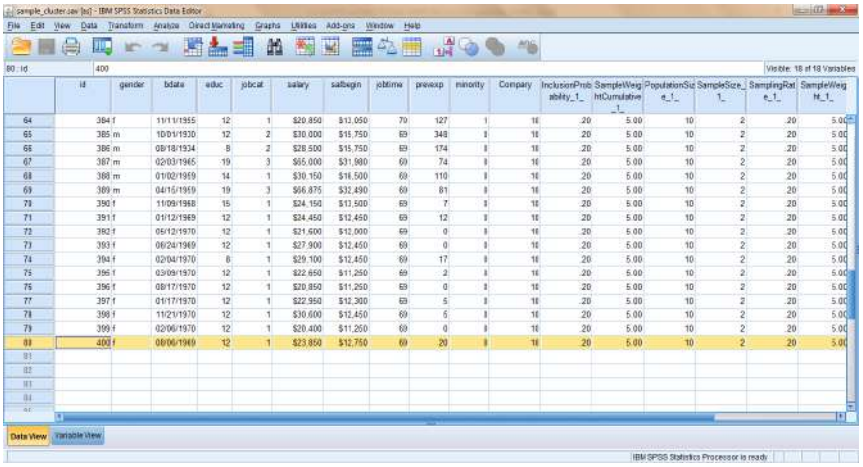
   Analyze → Complex Samples → Select a Sample....



2. Select *Design a sample* and choose a plan filename to save the sample plan.



3. Click *Next* to continue through the Wizard.
4. Optionally, in the Design Variables step, one can define strata, clusters, and input sample weights. Select the variable "Company" as cluster. Then, click Next.
5. Optionally, in the Sampling Method step, one can choose a method for selecting items.
    - If one select PPS Brewer or PPS Murthy, one can click Finish to draw the sample. Otherwise, click Next.
6. In the Sample Size step, specify the number or proportion of units to sample.
7. Optionally, in further steps one can:
    - Choose output variables to save.
    - Add a second or third stage to the design.
    - Set various selection options, including which stages to draw samples from, the random number seed, and whether to treat user-missing values as valid values of design variables.
    - Choose where to save output data.
8. Now click Finish to draw the sample.



Developed Sample Plan can be used for furthermore random sample selection as follows

Analyze → Complex Samples → Draw the Sample….

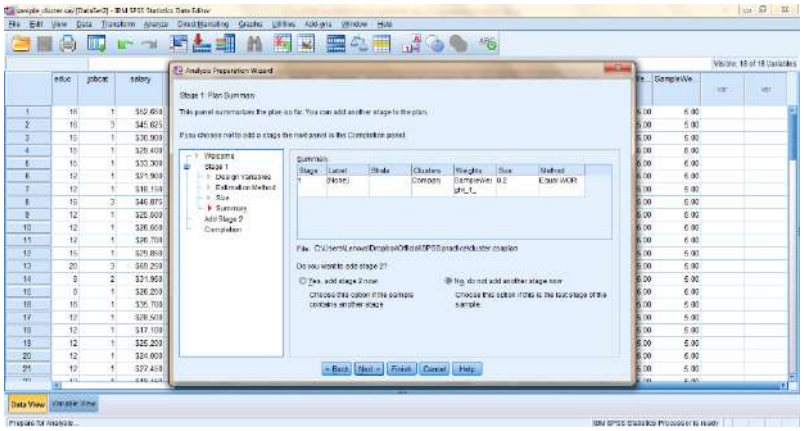
After selection of Sample next step is to prepare the sample for analysis. The Analysis Preparation Wizard guides through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, one should have a sample drawn according to a complex design.

Creating a New Analysis Plan

1. From the menus choose:

   Analyze → Complex Samples → Prepare for Analysis...

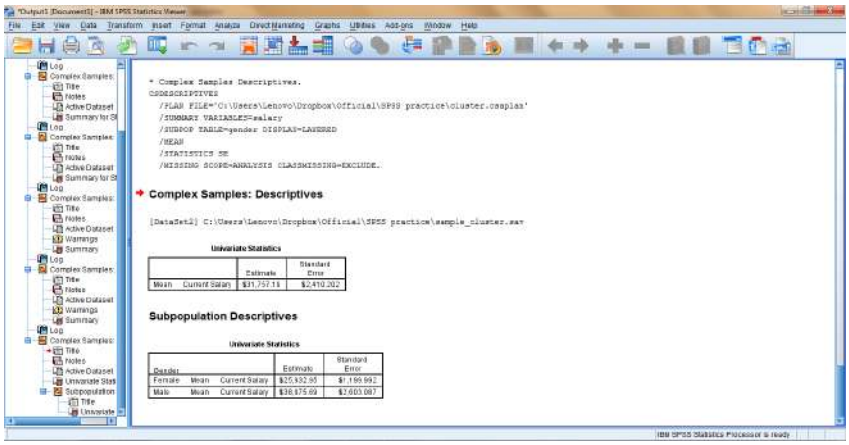
2. Select Create a plan file, and choose a plan filename to save the analysis plan.
3. Click Next to continue through the Wizard.
4. Specify the variable containing sample weights in the Design Variables step, optionally defining strata and clusters.
5. Optionally, in further steps one can:
   a. Select the method for estimating standard errors in the Estimation Method step.
   b. Specify the number of units sampled or the inclusion probability per unit in the Size step.
   c. Add a second or third stage to the design.
6. Now click Finish to save the plan.



Now using this Analysis Plan file one generates several types of outputs available in the *Complex Samples* option like

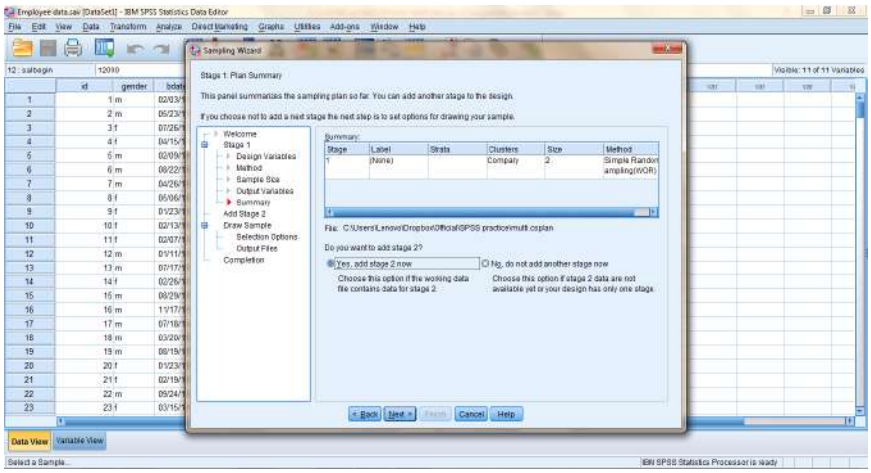
- Frequencies
- Descriptive
- Crosstabs
- Ratios
- General Linear Model
- Logistic Regression
- Ordinal Regression
- Cox Regression

Results from the Descriptive options using the "Current Salary" is given by

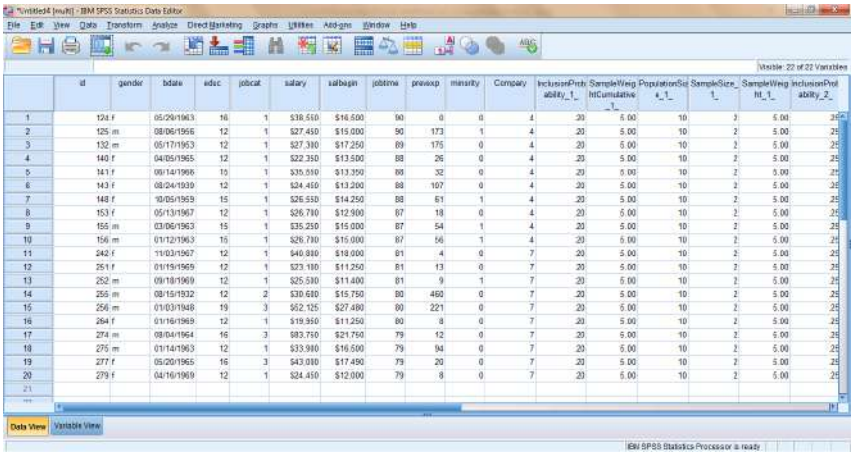


For selection of samples by Multistage sampling design one can edit the existing Sample Plan for cluster sampling or prepare new sampling plan according to Multistage sampling.

At the seventh step of the earliar shown "***Creating a New Sample Plan***", one should select "*Yes, add stage 2 now*" when the question "*Do you want to add Stage 2*" pops up in the sampling wizard as shown below:



Then define "*sample size*" for the stage 2 and path where to save the output file. An output file is given below when, first, 2 clusters are selected by SRSWOR and, the, within each selected cluster 10 units are selected by SRSWOR.



For analysis as per two stage sampling design, New Analysis Plan shall be created and further analysis of the sample shall be carried out.

REFERENCES:

1. Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).
2. Morgan, G.A., Leech, N.L., Gloeckner G.W. and Barrett, K. C. (2012).*IBM SPSS for Introductory Statistics: Use and Interpretation*. Fifth Edition, Routledge.
3. Nie, N. H., Bent, D. H. and Hull, C. H.(1970). *SPSS: Statistical Package for the Social Sciences*. New York: McGraw-Hill.

ANALYSIS OF SURVEY DATA USING SPSS

Deepak Singh and Raju kumar

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. INTRODUCTION

SPSS is a widely used software package for statistical analysis in social science. SPSS is capable of handling large amounts of data and can perform all of the analyses covered in the text and much more. The current versions (2015) are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics Software*) *Statistics*. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with rather simple instructions. SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation. SPSS integrates complex data and file management, statistical analysis and reporting functions. Purpose of this chapter is to introduce the basic features of the SPSS for its application in survey data analysis.

When surveying a population, choosing a simple random sample may not be the best approach. A probability sample that uses strategies like stratification, clustering, and multistage sampling has many advantages over simple random sample under certain conditions like to increase precision, decrease cost, ensuring sub-populations are included etc. Under these situations, it is recommended to use techniques dedicated to producing correct estimates for complex sample data.

IBM SPSS Complex Samples can compute statistics and standard errors from complex sample designs by incorporating the designs into survey analysis. It offers planning tools such as stratified, clustered or multistage sampling. From the planning stage and sampling through the analysis stage, SPSS Complex Samples allows one to select a sample according to a complex design and incorporate the design specifications into the data analysis making it easy to obtain accurate and reliable results. SPSS Complex Samples considers up to three states when analyzing data from a multistage design therefore multistage analysis up to three stages is possible through it.

2. STATISTICAL PROCEDURE FOR SURVEY DATA ANALYSIS IN SPSS

Complex Samples submenu under the Analyze menu provides procedures for Sampling from Complex Designs and incorporate the design specifications into the data analysis, thus ensuring the results are valid. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

3. PROPERTIES OF COMPLEX SAMPLES

A complex sample can differ from a simple random sample in many ways. In a simple random sample, individual sampling units are selected at random with equal probability and without replacement (WOR) directly from the entire population. By contrast, a given complex sample can have some or all of the following features:

3.1 **STRATIFICATION**

Stratified sampling involves selecting samples independently within non-overlapping subgroups of the population, or strata. For example, strata may be socioeconomic groups, job categories, age groups, or ethnic groups. With stratification, one can ensure adequate sample sizes for subgroups of interest, improve the precision of overall estimates, and use different sampling methods from stratum to stratum.

3.2 **CLUSTERING**

Cluster sampling involves the selection of groups of sampling units, or clusters. For example, clusters may be schools, hospitals, or geographical areas, and sampling units may be students, patients, or citizens. Clustering is common in multistage designs and area (geographic) samples.

3.3 **MULTIPLE STAGES**

In multistage sampling, one selects a first-stage sample based on clusters. Then it creates a second-stage sample by drawing subsamples from the selected clusters. If the second-stage sample is based on sub-clusters, one can then add a third stage to the sample. For example, in the first stage of a survey, a sample of cities could be drawn. Then, from the selected cities, households could be sampled.

Finally, from the selected households, individuals could be polled. The Sampling and Analysis Preparation wizards allow you to specify three stages in a design.

3.4 **NON RANDOM SAMPLING**

When selection at random is difficult to obtain, units can be sampled systematically (at a fixed interval) or sequentially.

3.5 **UNEQUAL SELECTION PROBABILITIES**

When sampling clusters that contain unequal numbers of units, one can use probability-proportional-to-size (PPS) sampling to make a cluster's selection probability equal to the proportion of units it contains. PPS sampling can also use more general weighting schemes to select units.

3.6 **UNRESTRICTED SAMPLING**

Unrestricted sampling selects units with replacement (WR). Thus, an individual unit can be selected for the sample more than once.

3.7 **SAMPLING WEIGHTS**

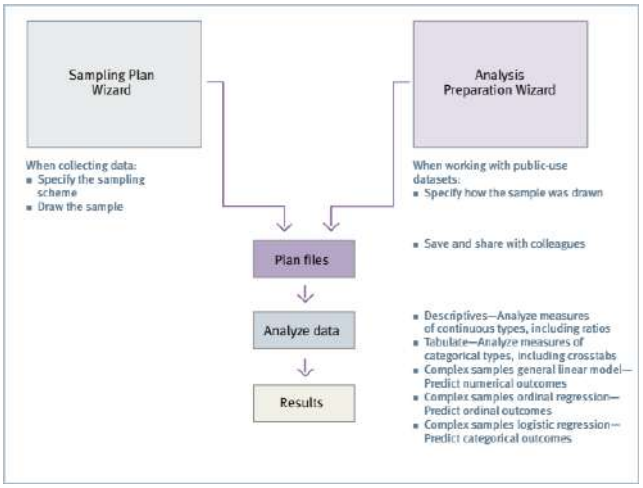
Sampling weights are automatically computed while drawing a complex sample and ideally correspond to the "frequency" that each sampling unit represents in the target population. Therefore, the sum of the weights over the sample should estimate the population size. Complex Samples analysis procedures require sampling weights in order to properly

analyze a complex sample. Note that these weights should be used entirely within the Complex Samples option and should not be used with other analytical procedures via the Weight Cases procedure, which treats weights as case replications.

4. USAGE OF COMPLEX SAMPLES PROCEDURES

The usage of Complex Samples procedures depends on the particular needs. The primary types of users are those who: Plan and carry out surveys according to complex designs, possibly analyzing the sample later.

The first step for SPSS Complex Samples is to use the wizards. If you are creating your own samples, use the Sampling Wizard to define the sampling scheme but if using datasets that have been sampled, such as those provided by the CDC, DHS surveys etc. use the Analysis Preparation Wizard to specify how the samples were defined and how to estimate standard errors. Once you create a sample or specify standard errors, you can create plans, analyze your data, and produce results.



SPSS complex samples helps to obtain correct estimates such as Population totals, means, ratios, standard errors, produce correct confidence intervals and hypothesis tests and predict outcomes.

4.1 **Complex Samples Plan (CSPLAN)**

Before using the Complex Samples analysis procedures, one may need to use the Analysis Preparation Wizard Regardless of which type of user one may be, one need to supply design information to Complex Samples procedures. This information is stored in a **plan file** for easy reuse. CSPLAN does not actually extract the sample or analyze data.

PLAN FILES

A plan file contains complex sample specifications. There are two types of plan files:

Sampling Plan To sample cases, sample design created by CSPLAN is used as input to the CSSELECT (discussed next) procedure.

Analysis Plan This plan file contains information needed by Complex Samples analysis procedures to properly compute variance estimates for a complex sample. The plan includes the sample structure, estimation methods for each stage, and references to required

variables, such as sample weights. The Analysis Preparation Wizard allows you to create and edit analysis plans. To analyze sample data, use an analysis design created by CSPLAN as input to the CSDESCRIPTIVES, CSTABULATE, CSGLM, CSLOGISTIC, or CSORDINAL procedures.

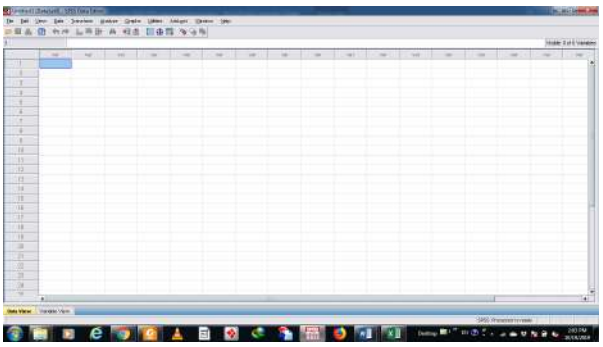
There are several advantages to saving your specifications in a plan file, including:

A surveyor can specify the first stage of a multistage sampling plan and draw first-stage units now, collect information on sampling units for the second stage, and then modify the sampling plan to include the second stage.

An analyst who doesn't have access to the sampling plan file can specify an analysis plan and refer to that plan from each Complex Samples analysis procedure. A designer of large-scale public use samples can publish the sampling plan file, which simplifies the instructions for analysts and avoids the need for each analyst to specify his or her own analysis plans.

Steps for Drawing the Sample and Analysis of Sampled Data

- START – IBM SPSS for windows



- Prepare a file from which data to be sampled in **SPSS Data Editor**
  or browse your data file by using following procedure:
  File - Open - Data

4.1.1 **SAMPLING WIZARD FOR COMPLEX DESIGN**

The sampling wizard is used to create, modify or executing the sampling plan file. We should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind for carrying this feature in SPSS.

Creating a New Sample Plan

1. From the menus choose:
   Analyze > **Complex Samples** > **Select a Sample...**
2. Select **Design a sample** and choose a plan filename to save the sample plan (demo_cs1)

3. Click Next to continue through the Wizard.
4. Optionally, in the Design Variables step, you can define strata, clusters, and input sample weights. After you define these, click Next.

5. Optionally, in the Sampling Method step, you can choose a method by which items can be selected like Simple random sampling with or without replacement, probability proportional to size etc. If you select **PPS Brewer** or **PPS Murthy**, you can click **Finish** to draw the sample. Otherwise, click **Next** and then:
6. In the Sample Size step, specify the number or proportion of units to sample. You can now click Finish to draw the sample.

4.1.2 **Sampling Wizard: Design Variables**

This step allows you to select stratification and clustering variables. In addition, if the current sample design is part of a larger sample design, you may have sample weights from a previous stage of the larger design. You can specify a numeric variable containing these weights in the first stage of the current design. Sample weights are computed automatically for subsequent stages of the current design.

This step has four options viz;

Stratify By; for stratification,

Clusters; for clustering variables,

Input sample weight; when sample weights of each unit are available

Stage label; for specifying an optional string label for each stage. This is used in the output to help identify stage wise information.

4.1.3 **Sampling Wizard: Sampling Method**

Some sampling types allow one to choose whether to sample with replacement (WR) or without replacement (WOR). See the type descriptions for more information. Note that some probability-proportional-to-size (PPS) types are available only when clusters have been defined and that all PPS types are available only in the first stage of a design. Moreover, WR methods are available only in the last stage of a design.

- **Simple Random Sampling** Units are selected with equal probability. They can be **selected** with or without replacement.
- **Simple Systematic** Units are selected at a fixed interval throughout the sampling frame (or strata, if they have been specified) and extracted without replacement. A randomly selected unit within the first interval is chosen as the starting point.
- **Simple Sequential** Units are selected sequentially with equal probability and without replacement.
- **PPS** This is a first-stage method that selects units at random with probability proportional to size. Any units can be selected with replacement; only clusters can be sampled without replacement.
- **PPS Systematic** This is a first-stage method that systematically selects units with probability proportional to size. They are selected without replacement.
- **PPS Sequential** This is a first-stage method that sequentially selects units with probability proportional to cluster size and without replacement.
- **PPS Brewer** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Murthy** This is a first-stage method that selects two clusters from each stratum with probability proportional to cluster size and without replacement. A cluster variable must be specified to use this method.
- **PPS Sampford** This is a first-stage method that selects more than two clusters from each stratum with probability proportional to cluster size and without replacement. It is an extension of Brewer's method. A cluster variable must be specified to use this method.
- **Use WR estimation for analysis**. By default, an estimation method is specified in the plan file that is consistent with the selected sampling method. This allows one to use with-replacement estimation even if the sampling method implies WOR estimation. This option is available only in stage 1.
- **Measure of size (mos):** If a PPS method is selected, one must specify a measure of size that defines the size of each unit. These sizes can be explicitly defined in a variable or they can be computed from the data. Optionally, one can set lower and upper bounds on the MOS, overriding any values found in the MOS variable or computed from the data. These options are available only in stage 1.

4.1.4 **Sampling Wizard: Sample Size**

This step allows you to specify the number or proportion of units to sample within the current stage. The sample size can be fixed or it can vary across strata. For specifying sample size, clusters chosen in previous stages can be used to define strata.

- **Units.** You can specify an exact sample size or a proportion of units to sample.
- **Value.** A single value is applied to all strata. If **Counts** is selected as the unit metric, you should enter a positive integer. If **Proportions** is selected, you should enter a

non-negative value. Unless sampling with replacement, proportion values should also be no greater than 1.

- **Unequal values for strata.** Allows you to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows you to select a numeric variable that contains size values for strata.

4.1.5 **Sampling Wizard: Output Variables**

This step allows you to choose variables to save when the sample is drawn.

Population size. The estimated number of units in the population for a given stage. The root name for the saved variable is *Population Size*.

Sample proportion. The sampling rate at a given stage. The root name for the saved variable is *Sampling Rate_*.

Sample size. The number of units drawn at a given stage. The root name for the saved variable is *Sample Size_*.

Sample weight. The inverse of the inclusion probabilities. The root name for the saved variable is *Sample Weight_*.

Some stage wise variables are generated automatically. These include:
Inclusion probabilities. The proportion of units drawn at a given stage. The root name for the saved variable is *Inclusion Probability_*.

Cumulative weight. The cumulative sample weight over stages before and including the current one. The root name for the saved variable is *Sample Weight Cumulative_*.

Index. Identifies units selected multiple times within a given stage. The root name for the saved variable is *Index_*.

4.1.6 **Sampling Wizard: Plan Summary**

This is the last step within each stage, providing a summary of the sample design specifications through the current stage. From here, one can either proceed to the next stage (creating it, if necessary) or set options for drawing the sample.

4.2 **Complex Samples Selection (CSSELECT)**

This step selects complex, probability-based samples from a population. One can also control other sampling options, such as the random seed and missing-value handling. It chooses units according to a sample design created through the CSPLAN procedure.Write sampled units to an external file using an option to keep/drop specified variables. It has two sub-parts i.e. DRAW SAMPLE SELECTION OPTIONS and DRAW SAMPLE OUTPUT FILES, which are discussed below:

4.2.1 **Sampling Wizard: Draw Sample Selection Options**

Draw sample. In addition to choosing whether to draw a sample, one can also choose to execute part of the sampling design. Stages must be drawn in order that is, stage 2 cannot be drawn unless stage 1 is also drawn. When editing or executing a plan, one cannot resample locked stages.

Seed. This allows one to choose a seed value for random number generation.

Include user-missing values. This determines whether user-missing values are valid. If so, user-missing values are treated as a separate category.

Data already sorted. If your sample frame is pre-sorted by the values of the stratification variables, this option allows one to speed the selection process.

4.2.2 **Sampling Wizard: Draw Sample Output Files**

This step allows one to choose where to direct sampled cases, weight variables, joint probabilities, and case selection rules.

Sample data. These options let one determine where sample output is written. It can be added to the active dataset, written to a new dataset, or saved to an external IBM® SPSS® Statistics data file. Datasets are available during the current session but are not available in subsequent sessions unless one explicitly save them as data files.

Joint probabilities. These options let one determine where joint probabilities are written. They are saved to an external SPSS Statistics data file. Joint probabilities are produced if the PPS WOR, PPS Brewer, PPS Sampford, or PPS Murthy method is selected and WR estimation is not specified.

Case selection rules. If one are constructing oner sample one stage at a time, one may want to save the case selection rules to a text file. They are useful for constructing the subframe for subsequent stages.

4.2.3 **Sampling Wizard: Finish**

This is the final step. One can save the plan file and draw the sample now or paste your selections into a syntax window.

4.3 **Preparing a Complex Sample for Analysis: The Analysis Preparation Wizard**

After selection of Sample next step is to prepare the sample for analysis. The Analysis Preparation Wizard guides one through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, one should have a sample drawn according to a complex design. Creating a new plan is most useful when one do not have access to the sampling plan file used to draw the sample. If one do have access to the sampling plan file used to draw the sample, one can use the default analysis plan contained in the sampling plan file or override the default analysis specifications and save your changes to a new file.

Complex Samples analysis procedures require analysis specifications from an analysis or sample plan file in order to provide valid results.

Plan. Specify the path of an analysis or sample plan file.

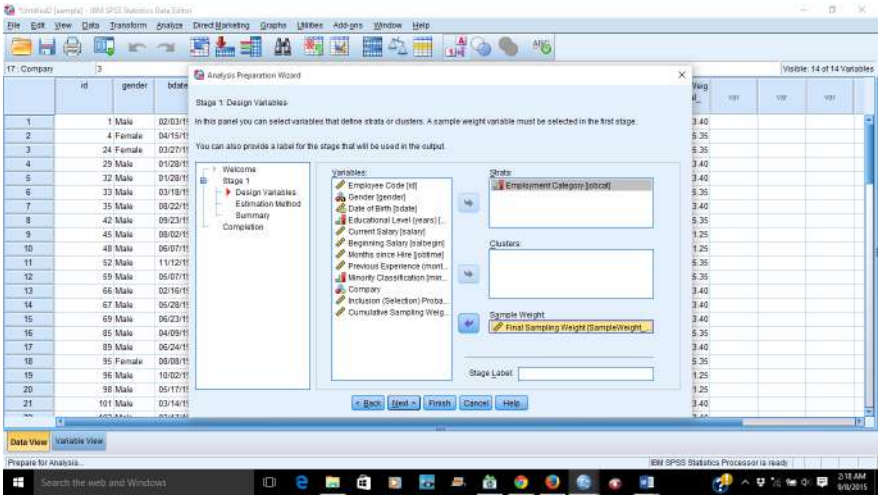
Joint Probabilities. In order to use Unequal WOR estimation for clusters drawn using a PPS WOR method, one need to specify a separate file or an open dataset containing the joint probabilities. This file or dataset is created by the Sampling Wizard during sampling.

Creating a New Analysis Plan

1. From the menus choose:

   Analyze → Complex Samples → Prepare for Analysis...

2. Select Create a plan file, and choose a plan filename to save the analysis plan.
3. Click Next to continue through the Wizard.
4. Specify the variable containing sample weights in the Design Variables step. Select the variable "Employee category" as strata.

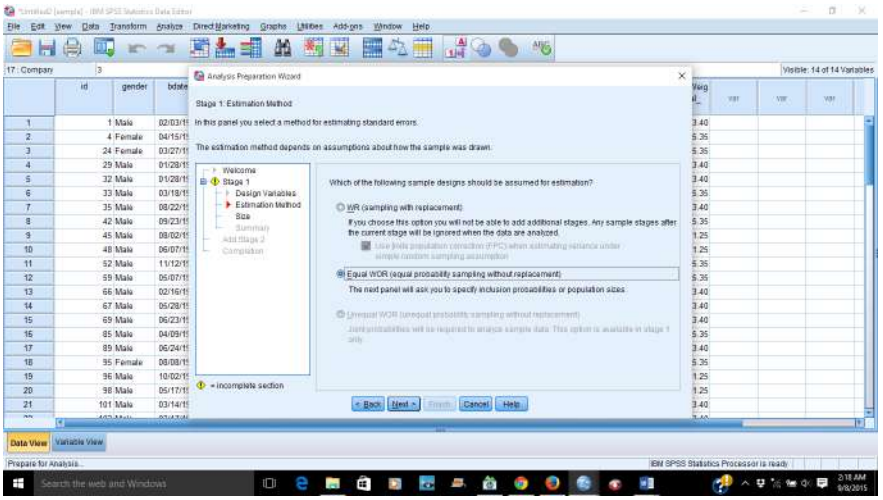


5. Optionally, in further steps one can:

a. Select the method for estimating standard errors in the Estimation Method step.

b. Specify the number of units sampled or the inclusion probability per unit in the Size step.

c. Add a second or third stage to the design.



6. Now click Finish to save the plan.

Analysis Preparation Wizard: Design Variables

This step allows one to identify the stratification and clustering variables and define sample weights. One can also provide a label for the stage.

Strata. The cross-classification of stratification variables defines distinct subpopulations, or strata. Oner total sample represents the combination of independent samples from each stratum.

Clusters. Cluster variables define groups of observational units, or clusters. Samples drawn in multiple stages select clusters in the earlier stages and then subsample units from the selected clusters. When analyzing a data file obtained by sampling clusters with replacement, one should include the duplication index as a cluster variable.

Sample Weight. One must provide sample weights in the first stage. Sample weights are computed automatically for subsequent stages of the current design.

Stage Label. One can specify an optional string label for each stage. This is used in the output to help identify stagewise information.

4.4 **Analysis Preparation Wizard: Estimation Method**

This step allows one to specify an estimation method for the stage.

WR (sampling with replacement). WR estimation does not include a correction for sampling from a finite population (FPC) when estimating the variance under the complex sampling design. One can choose to include or exclude the FPC when estimating the variance under simple random sampling (SRS).

Equal WOR (equal probability sampling without replacement). Equal WOR estimation includes the finite population correction and assumes that units are sampled with equal probability. Equal WOR can be specified in any stage of a design.

Unequal WOR (unequal probability sampling without replacement). In addition to using the finite population correction, Unequal WOR accounts for sampling units (usually clusters) selected with unequal probability. This estimation method is available only in the first stage.

4.5 **Analysis Preparation Wizard: Size**

This step is used to specify inclusion probabilities or population sizes for the current stage. Sizes can be fixed or can vary across strata. For the purpose of specifying sizes, clusters specified in previous stages can be used to define strata. Note that this step is necessary only when Equal WOR is chosen as the Estimation Method.

- **Units.** One can specify exact population sizes or the probabilities with which units were sampled.
- **Value.** A single value is applied to all strata. If ***Population Sizes*** is selected as the unit metric, one should enter a non-negative integer. If ***Inclusion Probabilities*** is selected, one should enter a value between 0 and 1, inclusive.
- **Unequal values for strata.** Allows one to enter size values on a per-stratum basis via the Define Unequal Sizes dialog box.
- **Read values from variable.** Allows one to select a numeric variable that contains size values for strata.

4.6 **Analysis Preparation Wizard: Plan Summary**

This is the last step within each stage, providing a summary of the analysis design specifications through the current stage. From here, one can either proceed to the next stage (creating it if necessary) or save the analysis specifications.

Analysis Preparation Wizard: Finish

This is the final step. One can save the plan file now or paste your selections to a syntax window.

When making changes to stages in the existing plan file, one can save the edited plan to a new file or overwrite the existing file. When adding stages without making changes to existing stages, the Wizard automatically overwrites the existing plan file. If one want to save the plan to a new file, choose to ***Paste the syntax generated by the Wizard into a syntax window*** and change the filename in the syntax commands.

4.7 **Analysis Preparation Wizard: Plan Summary**

This step allows one to review the analysis plan and remove stages from the plan.

Remove Stages. One can remove stages 2 and 3 from a multistage design. Since a plan must have at least one stage, one can edit but not remove stage 1 from the design.

5. Analysis Outputs

Now using this Analysis Plan file one generates several types of outputs available in the *Complex Samples* option like

Frequencies	General Linear Model
Descriptives	Logistic Regression
Crosstabs	Ordinal Regression
Ratios	Cox Regression

5.1 **Complex Samples Frequencies (CSFREQUENCIES)**

The Complex Samples Frequencies procedure produces frequency tables for selected variables and displays univariate statistics. Optionally, you can request statistics by subgroups, defined by one or more categorical variables. Variables for which frequency tables are produced should be categorical. Subpopulation variables can be string or numeric but should be categorical.
Complex Samples Frequencies-

1. From the menus choose:
   Analyze > Complex Samples > Frequencies
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Click Continue.
4. In the Complex Samples Frequencies dialog box, select at least one frequency variable.

5.2 Complex Samples Descriptives (CSDESCRIPTIVES)

CSDESCRIPTIVES estimates means, sums, and ratios, and computes their standard errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSDESCRIPTIVES performs analyses for subpopulations.

Descriptives-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Descriptive analysis
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples Descriptives Wizard – Measures - Sub Population – ok
5. Output- **SPSS Viewer**

5.3 Complex Samples Tabulate (CSTABULATE)

CSTABULATE displays one-way frequency tables or two-way cross tabulations and associated standard errors, design effects, confidence intervals, and hypothesis tests for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSTABULATE creates tables for subpopulations.

Crosstabs-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Crosstabs analysis Wizard
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples Crosstabs Wizard – Rows –Columns- Sub Population – ok
5. Output- **SPSS Viewer**

5.4 Complex Samples Ratios

The Complex Samples Ratios procedure displays univariate summary statistics for ratios of variables. Optionally, one can request statistics by subgroups, defined by one or more categorical variables.

Statistics. The procedure produces ratio estimates, *t* tests, standard errors, confidence intervals, coefficients of variation, unweighted counts, population sizes, design effects, and square roots of design effects.

Data. Numerators and denominators should be positive-valued scale variables. Subpopulation variables can be string or numeric but should be categorical.

Assumptions. The cases in the data file represent a sample from a complex design that should be analyzed according to the specifications in the file selected in the Complex Samples Plan dialog box.

Complex Samples Ratios-

1. From the menus choose:
2. Analyze > Complex Samples > Ratios...
3. Select a plan file. Optionally, select a custom joint probabilities file.
4. Click ***Continue***.
5. Select at least one numerator variable and denominator variable, here take "income" and "Age" respectively.
6. Optionally, one can specify variables to define subgroups for which statistics are produced, here take "ed"(ed is for sducation).

5.5 Complex Samples General Linear Model (CSGLM)

This procedure enables you to build linear regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA) models for samples drawn using complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSGLM performs analyses for subpopulations.

General Linear Model-

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for General Linear Model
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples General Linear Model Wizard – Dependent variable –Factors- Covariates – Subpopuation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.6 Complex Samples Ordinal (CSORDINAL)

CSORDINAL performs regression analysis on a binary or ordinal polychromous dependent variable using the selected cumulative link function for samples drawn by complex sampling methods. The procedure estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, as well as WR and WOR sampling procedures. Optionally, CSORDINAL performs analyses for a subpopulation.

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Plan for Complex Samples for Ordinal regression
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples for Ordinal regression– Dependent variable –Factors- Covariates –link function- Subpopulation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.7 Complex Samples Logistic Regression (CSLOGISTIC)

This procedure performs binary logistic regression analysis, as well as multinomial logistic regression (MLR) analysis, for samples drawn by complex sampling methods. CSLOGISTIC estimates variances by taking into account the sample design used to select the sample, including equal probability and PPS methods, and WR and WOR sampling procedures. Optionally, CSLOGISTIC performs analyses for subpopulations.

1. From the menus choose:
   Analyze > Complex Samples > Complex Samples Logistic Regression
2. Select a plan file by: File – Browse - Plan file name (demo_cs.csplan)
3. Continue
4. Complex Samples for Ordinal regression– Dependent variable –Factors- Covariates –link function- Subpopulation variable (Category, if category wise analysis is required)-ok
5. Output- **SPSS Viewer**

5.8 Complex Samples Cox Regression

The Complex Samples Cox Regression procedure performs survival analysis for samples drawn by complex sampling methods. Optionally, one can request analyses for a subpopulation.

Examples. A government law enforcement agency is concerned about recidivism rates in their area of jurisdiction. One of the measures of recidivism is the time until second arrest for offenders. The agency would like to model time to re-arrest using Cox Regression but are worried the proportional hazards assumption is invalid across age categories.

To Obtain Complex Samples Cox Regression

This feature requires the Complex Samples option.

From the menus choose:

Analyze > ***Complex Samples*** > ***Cox Regression***...

- Select a plan file. Optionally, select a custom joint probabilities file.
- Click ***Continue***.
- Specify the survival time by selecting the entry and exit times from the study.
- Select an event status variable.
- Click ***Define Event*** and define at least one event value.

SAS – AN OVERVIEW

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

SAS is a collection of modules that are used to process and analyze data. It began in the late ’60s and early ’70s as a statistical package (the name *SAS* originally stood for Statistical Analysis System). However, unlike many competing statistical packages, SAS is also an extremely powerful, general-purpose programming language. SAS is a predominant software in many industries. In recent years, it has been enhanced to provide state-of-the-art statistical tools for analysis. The only way to really learn a programming language is to write lots of programs, make some errors, correct the errors, and then make some more. If one already has access to SAS at work or school, he/she is ready to go. SAS Learning Edition 4.1 is useful for those who are learning SAS by themselves and do not have a copy of SAS to play with. This is a relatively inexpensive, fully functional version of SAS.

2. Getting Data into SAS

SAS can read data from almost any source. Common sources of data are raw text files, Microsoft Office Excel spreadsheets, Access databases, and most of the common database systems such as DB2 and Oracle. Most of this book uses either text files or Excel spreadsheets as data sources.

3. Components of SAS Programs

SAS programs often contain DATA steps and PROC steps. DATA steps are parts of the program where you can read or write the data, manipulate the data, and perform calculations. PROC (short for procedure) steps are parts of your program where you ask SAS to run one or more of its procedures to produce reports, summarize the data, generate graphs, and much more. DATA steps begin with the word DATA and PROC steps begin with the word PROC. Most DATA and PROC steps end with a RUN statement. SAS processes each DATA or PROC step completely and then goes on to the next step.

SAS also contains *global* statements that affect the entire SAS environment and remain in effect from one DATA or PROC step to another. In the program above, the OPTIONS and TITLE statements are examples of global statements. It is important to keep in mind that the actions of global statements remain in effect until they are changed by another global statement or until you end your SAS session.

All SAS programs, whether part of DATA or PROC steps, are made up of statements. Here is the rule: all SAS statements end with semicolons. This is an important rule because if you leave out a semicolon where one is needed, the program may not run correctly, resulting in hard-to-interpret error messages. Let’s discuss some of the basic rules of SAS statements. First, they can begin in any column and can span several lines, if necessary. Because a semicolon determines the end of a SAS statement, you can place more than one statement on a single line (although this is not recommended as a matter of style).

SAS is not case sensitive. Well, this is almost true. Of course references to external files must match the rules of your particular operating system. So, if you are running SAS under UNIX or Linux, file names will be case-sensitive. As you will see later, you get to name the variables in a SAS data set. The variable names in Program 1 are Name, Code, Days,

Number, Price, and CostPerSeed. Although SAS doesn't care whether you write these names in uppercase, lowercase, or mixed case, it does "remember" the case of each variable the *first* time it encounters that variable and uses that form of the variable name when producing printed reports.

4. SAS Names

SAS names follow a simple naming rule. All SAS variable names and data set names can be no longer than 32 characters and must begin with a letter or the underscore (_) character. The remaining characters in the name may be letters, digits, or the underscore character. Characters such as dashes and spaces are not allowed. Here are some valid and invalid SAS names.

Valid SAS Names

Parts, LastName, First_Name, Ques5, Cost_per_Pound, DATE, time, X12Y34Z56

Invalid SAS Names

8_is_enough - Begins with a number,

Price per Pound - Contains blanks,

Month-total - Contains an invalid character (-),

Num% - Contains an invalid character (%),

5. SAS Data Sets and SAS Data Types

When SAS reads data from anywhere (for example, raw data, spreadsheets), it stores the data in its own special form called a SAS data set. Only SAS can read and write SAS data sets. If you opened a SAS data set with another program (Microsoft Word, for example), it would not be a pretty sight. Even if SAS is reading data from Oracle tables or DB2, it is actually converting the data into SAS data set format in the background. The good news is that you don't ever have to worry about how SAS is storing its data or the structure of a SAS data set. However, it is important to understand that SAS data sets contain two parts: a descriptor portion and a data portion. Not only does SAS store the actual data values for you, it stores information about these values (things like storage lengths, labels, and formats). SAS has two types of variables: *character* and *numeric*. This makes it much simpler to use and understand than some other programs that have many more data types (for example, integer, long integer, and logical).

6. The SAS Display Manager and SAS Enterprise Guide

Because SAS runs on many different platforms (mainframes, microcomputers running various Microsoft operating systems, UNIX, and Linux), the way you write and run programs will vary. You might use a general-purpose text editor on a mainframe to write a SAS program, submit it, and send the output back to a terminal or to a file. On PCs, you might use the SAS Display Manager, where you write your program in the *Enhanced Editor (Editor* window*)*, see any error messages and comments about your program and the data in the *Log* window, and view your output in the *Output* window. In addition to the Enhanced Editor, an older program, simply called the *Program Editor*, is available for Windows and UNIX users. As an alternative to the Display Manager, you may enter the SAS environment using *SAS Enterprise Guide*, which is a front-end to SAS that allows you to use a menu-driven system to write SAS programs and produce reports. There are many excellent books published by SAS that offer detailed instructions on how to run SAS programs on each specific platform and the appropriate access method into SAS.

7. A Sample SAS Program

Let's start out with a simple SAS program that reads data from a text file and produces some basic reports to give you an overview of the structure of SAS programs. For this example, we have a text file with data on vegetable seeds. Each line of the file contains the following pieces of information (separated by spaces):

- Vegetable name
- Product code
- Days to germination
- Number of seeds
- Price

In SAS terminology, each piece of information is called a *variable*. (Other database systems, and sometimes SAS, use the term *column*.) A few sample lines from the file are shown here:

File **c:\my folder\crop.txt**

```
Crop_1 50104 55 30 195
Crop_1 51789 56 30 225
Crop_2 50179 68 150 395
Crop_2 50872 65 150 225
Crop_3 57224 75 200 295
Crop_3 62471 80 200 395
Crop_3 57828 66 200 295
Crop_4 52233 70 30 225
```

In this example, each line of data produces what SAS calls an *observation* (also referred to as a *row* in other systems). A complete SAS program to read this data file and produce a list of the data, a frequency count showing the number of entries for each crop, the average price per seed, and the average number of days until germination is shown here.

Program - A sample SAS program

```
*Comment 1: SAS Program to read veggie data file and to produce several reports;
    * Comment 2: Entering data using program editor;
    data crop;
        input Name $ Code Days Number Price;
            *$ for character variable;
        cards;
        Crop_1 50104 55 30 195
        Crop_1 51789 56 30 225
        Crop_2 50179 68 150 395
```

```
        Crop_2 50872 65 150 225
        Crop_3 57224 75 200 295
        Crop_3 62471 80 200 395
        Crop_3 57828 66 200 295
        Crop_4 52233 70 30 225
        ;
Run;

* Comment 3: To print the inserted data;
title "List of the Raw Data";
footnote "Overview of SAS";
proc print data= crop;
run;

* Comment 4: Alternative way for running data;
DATA crop;
        input Name $ Code Days Number Price @@;
        cards;
        Crop_1 50104 55 30 195 Crop_1 51789 56 30 225
        Crop_2 50179 68 150 395 Crop_2 50872 65 150 225
        Crop_3 57224 75 200 295 Crop_3 62471 80 200 395
        Crop_3 57828 66 200 295 Crop_4 52233 70 30 225
        ;
Run;

* Comment 5: To import from external sources - txt;
data crop;
        infile "c:\mywork\crop.txt";
        input Name $ Code Days Number Price;
run;

* Comment 6: To import from external sources - csv;
data crop;
        infile 'c:\mywork\crop.csv'  dlm=',' ;
        input Name $ Code Days Number Price;
```

```
run;

* Comment 7: Alternative way using IMPORT procedure*/
proc import datafile = 'c:\mywork\crop.csv'
        out = crop dbms=csv replace;
        getnames=no;
run;

* Comment 8: To import from external sources - xls */
proc import datafile = 'c:\mywork\crop.xls'
        out = crop dbms=excel replace;
        getnames=yes;
run;
* Comment 9: To modify the data;
data crop;
        set crop;
        CostPerSeed = Price / Number;*add new variable;
        *drop days;                    *delete variable;
        rename Number=Number_seeds;  *change variable name;
run;

* Comment 10: To sort the data;
proc sort data=crop;
        by Code;
run;

* Comment 11: To find the frequency counts;
title "Frequency Distribution of crop Names";
proc freq data= crop;
        tables Name;
run;

* Comment 12: To find means of the variables;
title "Average Cost of Seeds";
proc means data= crop;
```

```
        var Price CostPerSeed;
run;

* Comment 13: To find Scatter Plot;
proc plot data = crop;
        plot Days*Price = '*';
run;

* Comment 14: To fit linear regression;
proc reg data = crop;
        model Price = Days;
run;
```

References

Cody, R. (2018). *Learning SAS® by Example: A Programmer's Guide*. 2nd ed., Cary NC: SAS Institute Inc.

Delwiche, L. D., & Slaughter, S. J. (2002). *The little SAS book: A primer*. Cary, NC: SAS Institute.

ANALYSIS OF SURVEY DATA USING SAS

Ankur Biswas

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

A sampling method is a scientific and objective procedure of selecting units from the population and provides a sample that is expected to be representative of the population. A sampling method makes it possible to estimate the population parameters while reducing at the same time the size of survey operations. Some of the advantages of sample surveys as compared to complete enumeration are reduction in cost, greater speed, wider scope and higher accuracy. A function of the unit values of the sample is called an estimator. Various measures, like bias, mean square errors, variance etc. are used to assess the performance of the estimator. See Lohr (2010), Kalton (1983), Sukhatme *et al.* (1984), Cochran (1977), Murthy (1977), Raj (1968) and Kish (1965) for more information about statistical sampling and analysis of complex survey data.

The prime objective of a sample survey is to obtain inferences about the characteristic of a population. Population is defined as a group of units defined according to the objectives of the survey. The population may consist of all the households in a village / locality, all the fields under a particular crop. We may also consider a population of persons, families, fields, animals in a region, or a population of trees, birds in a forest depending upon the nature of data required. The information that we seek about the population is normally, the total number of units, aggregate values of various characteristics, averages of these characteristics per unit, proportions of units possessing specified attributes etc. The data can be collected in two different ways. The first one is complete enumeration which means collection of data on the survey characteristics from each unit of the population.

The main problem in sample surveys is the choice of a proper sampling strategy, which essentially comprise of a sampling method and the estimation procedure. In the choice of a sampling method there are some methods of selection while some others are control measures which help in grouping the population before the selection process. In the methods of selection, schemes such as simple random sampling, systematic sampling and varying probability sampling are generally used. Among the control measures are procedures such as stratified sampling, cluster sampling and multi-stage sampling etc. A combination of control measures along with the method of selection is called the sampling scheme.

2. Use of SAS Software for Survey Data Analysis

Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population. Due to variability among items, researchers apply scientific probability-based designs to select the sample. This reduces the risk of a distorted view of the population and enables statistically valid inferences to be made from the sample. To select probability-based random samples from a study population, you can use the SURVEYSELECT procedure, which provides a variety of methods for probability sampling. To analyze sample survey data, you can use the SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures, which incorporate the sample design into the analyses.

Many SAS/STAT procedures, such as the MEANS, FREQ, GLM, LOGISTIC, and PHREG procedures, can compute sample means, produce crosstabulation tables, and estimate regression relationships. However, in most of these procedures, statistical

inference is based on the assumption that the sample is drawn from an infinite population by simple random sampling. If the sample is in fact selected from a finite population by using a complex survey design, these procedures generally do not calculate the estimates and their variances according to the design actually used. Using analyses that are not appropriate for your sample design can lead to incorrect statistical inferences.

The SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures properly analyze complex survey data by taking into account the sample design. These procedures can be used for multistage or single-stage designs, with or without stratification, and with or without unequal weighting. The survey analysis procedures provide a choice of variance estimation methods, which include Taylor series linearization, balanced repeated replication (BRR), and the jackknife.

2.1 Proc SURVEYSELECT Procedure

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or can sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

To select a sample with PROC SURVEYSELECT, you input a SAS data set that contains the sampling frame, which is the list of units from which the sample is to be selected. The sampling units can be individual observations or groups of observations (clusters). You also specify the selection method, the desired sample size or sampling rate, and other selection parameters. PROC SURVEYSELECT selects the sample and produces an output data set that contains the selected units, their selection probabilities, and their sampling weights. When you select a sample in multiple stages, you invoke the procedure separately for each stage of selection, inputting the frame and selection parameters for each current stage.

PROC SURVEYSELECT provides methods for both equal probability sampling and probability proportional to size (PPS) sampling. In equal probability sampling, each unit in the sampling frame, or in a stratum, has the same probability of being selected for the sample. In PPS sampling, a unit's selection probability is proportional to its size measure.

PROC SURVEYSELECT provides the following equal probability sampling methods:

- simple random sampling (without replacement)
- unrestricted random sampling (with replacement)
- systematic random sampling
- sequential random sampling

This procedure also provides the following probability proportional to size (PPS) sampling methods:

- PPS sampling without replacement

- PPS sampling with replacement
- PPS systematic sampling
- PPS algorithms for selecting two units per stratum
- sequential PPS sampling with minimum replacement

The procedure uses fast, efficient algorithms for these sample selection methods. Thus, it performs well even for large input data sets or sampling frames. PROC SURVEYSELECT can perform stratified sampling by selecting samples independently within strata, which are non-overlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice toward meeting a variety of survey objectives. For example, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification toward improving the precision of the overall estimates. When you use a systematic or sequential selection method, PROC SURVEYSELECT can also sort by control variables within strata for the additional control of implicit stratification.

For stratified sampling, PROC SURVEYSELECT provides survey design methods to allocate the total sample size among the strata. Available allocation methods include proportional, Neyman, and optimal allocation. Optimal allocation maximizes the estimation precision within the available resources, taking into account stratum sizes, costs, and variances.

PROC SURVEYSELECT provides replicated sampling, where the total sample is composed of a set of replicates, and each replicate is selected in the same way. You can use replicated sampling to study variable non-sampling errors, such as variability in the results obtained by different interviewers. You can also use replication to estimate standard errors for combined sample estimates and to perform a variety of other resampling and simulation tasks.

Simple Random Sampling

The following PROC SURVEYSELECT statements select a probability sample of customers from the '*Customers'* data set by using simple random sampling:

```
title1 'Customer Satisfaction Survey';
title2 'Simple Random Sampling';
proc surveyselect data=Customers method=srs n=100
        out=SampleSRS;
run;
```

The PROC SURVEYSELECT statement invokes the procedure. The DATA= option names the SAS data set Customers as the input data set from which to select the sample. The METHOD=SRS option specifies simple random sampling as the sample selection method. In simple random sampling, each unit has an equal probability of selection, and sampling is without replacement. Without-replacement sampling means that a unit cannot be selected more than once. The N=100 option specifies a sample size of 100 customers. The OUT= option stores the sample in the SAS data set named SampleSRS.

Figure 1 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A sample of 100 customers is selected from the data set Customers by simple random sampling. With simple random sampling and no stratification in the sample design, the selection probability is the same for all units in the sample. In this sample, the selection probability for each customer equals 0.007423, which is the sample size (100) divided by the population size (13,471). The sampling weight equals 134.71 for each customer in the sample, where the weight is the inverse of the selection probability. If you specify the STATS option, PROC SURVEYSELECT includes the selection probabilities and sampling weights in the output data set. (This information is always included in the output data set for more complex designs.)

The random number seed is 39647. PROC SURVEYSELECT uses this number as the initial seed for random number generation. Because the SEED= option is not specified in the PROC SURVEYSELECT statement, the seed value is obtained by using the time of day from the computer's clock. You can specify SEED=39647 to reproduce this sample.

Figure 1. Sample Selection Summary

Customer Satisfaction Survey

Simple Random Sampling

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Input Data Set	CUSTOMERS
Random Number Seed	39647
Sample Size	100
Selection Probability	0.007423
Sampling Weight	134.71
Output Data Set	SAMPLESRS

The sample of 100 customers is stored in the SAS data set SampleSRS. PROC SURVEYSELECT does not display this output data set. The following PROC PRINT statements display the first 20 observations of SampleSRS:

```
title1 'Customer Satisfaction Survey';
title2 'Sample of 100 Customers, Selected by SRS';
title3 '(First 20 Observations)';
proc print data=SampleSRS(obs=20);
run;
```

Figure 2 displays the first 20 observations of the output data set SampleSRS, which contains the sample of customers. This data set includes all the variables from the DATA= input data set Customers. If you do not want to include all variables, you can use the ID statement to specify which variables to copy from the input data set to the output (sample) data set.

Figure 2. Customer Sample (First 20 Observations)

Customer Satisfaction Survey
Sample of 100 Customers, Selected by SRS
(First 20 Observations)

Obs	CustomerID	State	Type	Usage
1	036-89-0212	FL	New	74
2	045-53-3676	AL	New	411
3	050-99-2380	GA	Old	167
4	066-93-5368	AL	Old	1232
5	082-99-9234	FL	New	90
6	097-17-4766	FL	Old	131
7	110-73-1051	FL	Old	102
8	111-91-6424	GA	New	247
9	127-39-4594	GA	New	61
10	162-50-3866	FL	New	100
11	162-56-1370	FL	New	224
12	167-21-6808	SC	New	60
13	168-02-5189	AL	Old	7553
14	174-07-8711	FL	New	284
15	187-03-7510	SC	New	21
16	190-78-5019	GA	New	185
17	200-75-0054	GA	New	224
18	201-14-1003	GA	Old	3437
19	207-15-7701	GA	Old	24
20	211-14-1373	AL	Old	88

Stratified Sampling

In this section, stratification is added to the sample design for the customer satisfaction survey. The sampling frame, which is the list of all customers, is stratified by *State* and *Type*. This divides the sampling frame into non-overlapping subgroups formed from the

values of the State and Type variables. Samples are then selected independently within the strata.

PROC SURVEYSELECT requires that the input data set be sorted by the STRATA variables. The following PROC SORT statements sort the *Customers* data set by the stratification variables *State* and *Type*:

```
proc sort data=Customers;
  by State Type;
run;
```

The following PROC SURVEYSELECT statements select a probability sample of customers from the *Customers* data set according to the stratified sample design:

```
title1 'Customer Satisfaction Survey';
title2 'Stratified Sampling';
proc surveyselect data=Customers method=srs n=15
                  seed=1953 out=SampleStrata;
  strata State Type;
run;
```

The STRATA statement names the stratification variables State and Type. In the PROC SURVEYSELECT statement, the METHOD=SRS option specifies simple random sampling. The N=15 option specifies a sample size of 15 customers for each stratum. If you want to specify different sample sizes for different strata, you can use the N=SAS-data-set option to name a secondary data set that contains the stratum sample sizes. The SEED=1953 option specifies '1953' as the initial seed for random number generation. Figure 3 displays the output from PROC SURVEYSELECT, which summarizes the sample selection. A total of 120 customers are selected.

Figure 3. Sample Selection Summary

Customer Satisfaction Survey
Stratified Sampling

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variables	State
	Type
Input Data Set	CUSTOMERS
Random Number Seed	1953
Stratum Sample Size	15

Number of Strata	8
Total Sample Size	120
Output Data Set	SAMPLESTRATA

2.2 Proc SURVEYMEANS Procedure

The SURVEYMEANS procedure produces estimates of population means and totals from sample survey data. The procedure also computes estimates of proportions for categorical variables, estimates of quantiles for continuous variables, and ratio estimates of means and proportions. For all of these statistics, PROC SURVEYMEANS provides standard errors, confidence limits, and *t* tests.

PROC SURVEYMEANS provides domain analysis, which computes estimates for domains (subpopulations), in addition to analysis for the entire study population. Formation of subpopulations can be unrelated to the sample design, and so the domain sample sizes can actually be random variables. Domain analysis takes this variability into account by using the entire sample to estimate the variance of domain estimates. Domain analysis is also known as subgroup analysis, subpopulation analysis, and subdomain analysis.

Simple Random Sampling

This example illustrates how you can use PROC SURVEYMEANS to estimate population means and proportions from sample survey data. The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. Researchers want to know how much these students spend weekly for ice cream, on average, and what percentage of students spend at least $10 weekly for ice cream.

To answer these questions, 40 students were selected from the entire student population by using simple random sampling (SRS). Selection by simple random sampling means that all students have an equal chance of being selected and no student can be selected more than once. Each student selected for the sample was asked how much he or she spends for ice cream per week, on average. The SAS data set *IceCream* saves the responses of the 40 students:

```
data IceCream;
  input Grade Spending @@;
  if (Spending < 10) then Group='less';
  else Group='more';
  datalines;
7 7  7  7  8 12  9 10  7  1  7 10  7  3  8 20  8 19  7 2
7 2  9 15  8 16  7  6  7  6  7  6  9 15  8 17  8 14  9 8
9 8  9  7  7  3  7 12  7  4  9 14  8 18  9  9  7  2  7 1
7 4  7 11  9  8  8 10  8 13  7  2  9  6  9 11  7  2  7 9
;
```

The variable Grade contains a student's grade. The variable Spending contains a student's response regarding how much he spends per week for ice cream, in dollars. The variable Group is created to indicate whether a student spends at least $10 weekly for ice cream: Group='more' if a student spends at least $10, or Group='less' if a student spends less than $10.

You can use PROC SURVEYMEANS to produce estimates for the entire student population, based on this random sample of 40 students:

```
title1 'Analysis of Ice Cream Spending';
title2 'Simple Random Sample Design';
proc surveymeans data=IceCream total=4000;
  var Spending Group;
run;
```

The PROC SURVEYMEANS statement invokes the procedure. The TOTAL=4000 option specifies the total number of students in the study population, or school. The procedure uses this total to adjust variance estimates for the effects of sampling from a finite population. The VAR statement names the variables to analyze, Spending and Group. Figure 4 displays the results from this analysis. There are a total of 40 observations used in the analysis. The "Class Level Information" table lists the two levels of the variable Group. This variable is a character variable, and so PROC SURVEYMEANS provides a categorical analysis for it, estimating the relative frequency or proportion for each level. If you want a categorical analysis for a numeric variable, you can name that variable in the CLASS statement.

Figure 4. Analysis of Ice Cream Spending

Analysis of Ice Cream Spending
Simple Random Sample Design

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	40

Class Level Information		
Class Variable	Levels	Values
Group	2	less more

Statistics						
Variable	**Level**	**N**	**Mean**	**Std Error of Mean**	**95% CL for Mean**	
Spending		40	8.750000	0.845139	7.04054539	10.4594546
Group	**less**	23	0.575000	0.078761	0.41568994	0.7343101
	more	17	0.425000	0.078761	0.26568994	0.5843101

2.3 The SURVEYFREQ Procedure

This procedure produces one-way to *n*-way frequency and cross tabulation tables from sample survey data. These tables include estimates of population totals, population proportions (overall proportions, and also row and column proportions), and corresponding standard errors. Confidence limits, coefficients of variation, and design effects are also available. The procedure provides a variety of options to customize the table display.

2.4 PROC SURVEYREG PROCEDURE

The SURVEYREG procedure performs regression analysis for sample survey data. The procedure fits linear models and computes regression coefficients and their variance-covariance matrix. The procedure enables you to specify classification effects by using the same syntax as in the GLM procedure.

2.5 PROC SURVEYLOGISTIC PROCEDURE

The SURVEYLOGISTIC procedure provides logistic regression analysis for sample survey data. Logistic regression analysis investigates the relationship between discrete responses and a set of explanatory variables. PROC SURVEYLOGISTIC fits linear logistic regression models for discrete response survey data by the method of maximum likelihood and incorporates the sample design into the analysis. The SURVEYLOGISTIC procedure enables you to specify categorical classification variables (also known as CLASS variables) as explanatory variables in the model by using the same syntax for main effects and interactions as in the GLM and LOGISTIC procedures.

Table 1: Survey Sampling and Analysis Procedures in SAS/STAT Software

PROC SURVEYSELECT	
Selection Methods	Simple random sampling (without replacement)
	Unrestricted random sampling (with replacement)
	Systematic
	Sequential
	Probability proportional to size (PPS) sampling, with and without replacement
	PPS systematic
	PPS for two units per stratum
	PPS sequential with minimum replacement
Allocation Methods	Proportional

	Optimal
	Neyman
Sampling Tools	Cluster sampling
	Replicated sampling
	Serpentine sorting
PROC SURVEYMEANS	
Statistics	Estimates of population means and totals
	Estimates of population proportions
	Estimates of population quantiles
	Ratio estimates
	Standard errors
	Confidence limits
	Hypothesis tests
	Domain analysis
PROC SURVEYFREQ	
Tables	One-way frequency tables
	Two-way and multiway crosstabulation tables
	Estimates of population totals and proportions
	Standard errors
	Confidence limits
Analyses	Tests of goodness of fit
	Tests of independence
	Risks and risk differences
	Odds ratios and relative risks
Graphics	Weighted frequency and percent plots
	Odds ratio, relative risk, and risk difference plots
PROC SURVEYREG	
Analyses	Linear regression model fitting
	Regression coefficients
	Covariance matrices
	Confidence limits
	Hypothesis tests
	Estimable functions
	Contrasts
	Least squares means (LS-means) of effects
	Custom hypothesis tests among LS-means
	Regression with constructed effects
	Predicted values and residuals
	Domain analysis
PROC SURVEYLOGISTIC	
Analyses	Cumulative logit regression model fitting

Logit, probit, and complementary log-log link functions
Generalized logit regression model fitting
Regression coefficients
Covariance matrices
Confidence limits
Hypothesis tests
Odds ratios
Estimable functions
Contrasts
Least squares means (LS-means) of effects
Custom hypothesis tests among LS-means
Regression with constructed effects
Model diagnostics
Domain analysis

References

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition. John Wiley and Sons.

Raj, D. (1968). *Sampling Theory*. TATA McGRAW-HILL Publishing Co. Ltd.

Kalton, G. (1983), *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Beverly Hills, CA and London: Sage Publications.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, Second Edition, Pacific Grove, CA: Duxbury Press.

Murthy, M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Singh, Daroga and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. Wiley Eastern Limited.

Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Third Revised Edition, Iowa State University Press, USA.

MAPI (MOBILE ASSISTED PERSONAL INTERVIEW) - ICAR-IASRI APP FOR COLLECTION OF SURVEY DATA

Kaustav Aditya

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

The need for timely, reliable and comprehensive statistics on crop area and production assumes special significance in view of the vital role played by the Agricultural Sector in the Indian Economy. To collect data on crop area and production various large scale surveys are conducted by the Ministry of Agriculture and Farmers Welfare and National Sample Survey Office, Ministry of Statistics and Programme Implementation, every year which incur a huge cost. Further, in large scale surveys conducted for collection of the field level primary data, the data quality is declining as there was increase in occurrence of many type of non-sampling errors while collection, tabulation and cleaning of the data and due to that the expected outcome from the data is hard to obtain. To tackle these kinds of problems and reduce the cost of data collection, ICAR- Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi team under the project entitled "Pilot study for developing state level estimates of crop area and production on the basis of sample sizes recommended by professor Vaidyanathan committee report (funded by Department of Agriculture &Cooperation &Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India)", have developed an Android Mobile Based Application named **Mobile Assisted Personal Interview** (MAPI) for Data Collection which is compatible with all the Android smart phones available in the market. This software is built in android version 4.1(Jelly Bean) and inbuilt database support named Sqlite Database which support average models of android mobile phones and tablets. The **MAPI** software has been developed to collect the data for crop area, yield, production and other demographic and social information as per the requirement of the pilot study and already available online at sample survey resource server of ICAR-IASRI, New Delhi at http://sample.iasri.res.in/ssrs/android.html. For other customized surveys the developed software can be modified based on the request of the registered user. To develop MAPI android eclipse software is used which generate .apk (android package kit) file for installation as per the device requirement. Purpose of this application is to collect the data for crop area, yield, production and other demographic and social information of the area under consideration. The questionnaires are developed in IASRI as per the requirement of the project. For other customized surveys the developed software can be modified based on the request of the user. To access the entered data, MS Excel file can be generated directly from the application and can be accessed from the phone memory or can be uploaded to any kind of cloud storage (i.e. dropbox, google drive etc.) or can be emailed to the headquarters or authorities conducting the survey directly from the phone. For Authentication user has to register to the application before using it with valid email id. This feature maintains the confidentiality of the data. Further, to verify the actual data collection event there is a provision in the questionnaire to record the GPS location and snapshot of the field where the survey is done. Likewise, there are many features in this software which increase the quality of the data. The online version of the MAPI software is also available. In the online version there is a need for good internet connectivity as in this

version data gets uploaded directly to the server during the process of data collection. For using this version, we need to set up a server and install few prerequisite softwares that are readily available in the market and there is no need to depend on other organizations for setting up the server with some unknown softwares which may put the data in risk of cyber crimes. After setting up the software we have to prepare the questionnaire in the MAPI software and connect the MAPI software with the sever for deployment. Further many features are under development and soon the software will be upgraded with those features like roasters, calculators and GPS area measurement with Google Maps etc.

2. Steps to use MAPI software

Steps to access MAPI in android support device:-

1. Download the software from http://sample.iasri.res.in/ssrs/android.html
2. Install the setup file named "iasrischedule.apk" in memory of the smart phone or tablet.
3. After installation, MAPI app icon displayed on the main screen of the device, see Figure 1.
4. After user click on the icon, app will open the home page. See Figure 2 for home page of MAPI. The "Menu" tab in MAPI is provided for the user for navigation between several questionnaires at the time of survey.

Figure 1. MAPI app icon.

Figure 2. Home page of MAPI.

5. As shown in the home screen of the application, user can easily visit the institute's website or our project team with simple click. However to use the questionnaires, user has to signup first with the application as shown in Figure 3.

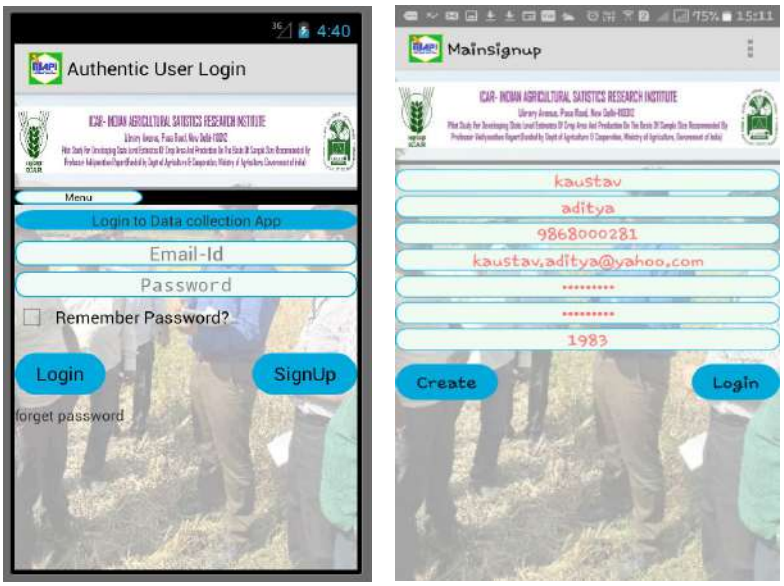


Figure 3. Login and signup page of MAPI

6. After filling one time registration form user need to click on create button, user can easily login to the application and access the Schedules/Questionnaires. The software is built with a remember password tab which helps the user in frequent login to the software without much constrain. The software is made to fit with the questionnaires prepared under the pilot survey planned at the institute for enumeration of crop area and yield under each selected village of each tehsil of each district of the selected 5 states of India.

The following questionnaires are prepared in the MAPI software.

a. Enumeration Schedule for Crop Area Enumeration (Figure 4)

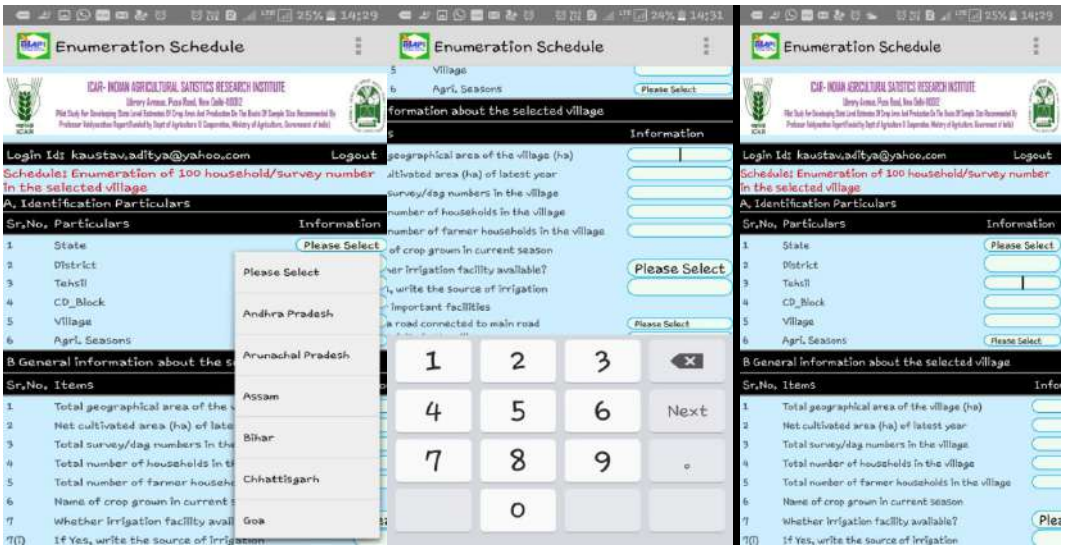


Figure 4. Enumeration Schedule

b. CCE Schedule-1 (For selection of plots for Crop Cutting Experiment (CCE))

Particular of plots selected under each survey number for crop cutting experiment is recorded in CCE schedule 1. This questionnaire consists of questions to record the GPS and pictorial information of the selected plot of CCE as shown in Figure 11. These questions were added for verification of the user or field investigators that he actually visited the plot and conducted the selection of plot of the CCE. The users have to complete the whole questionnaire before submitting the data using the "Submit" button in the questionnaire.

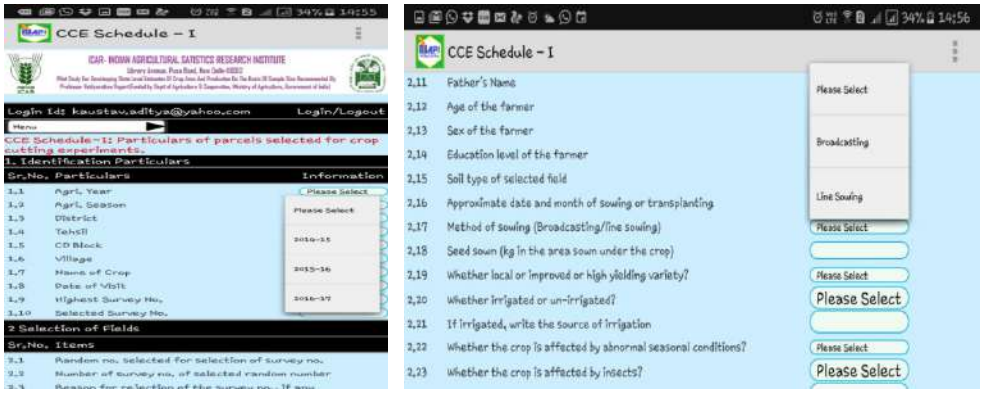


Figure 5. CCE Schedule I

Figure 6. **Q**uestionnaire for GPS location and Picture of the plot

c. CCE Schedule II (For recording the produce obtained from CCE plot)

This questionnaire or schedule is formulated to capture the information regarding the outcome of CCE from the farmer's field (Figure 7).

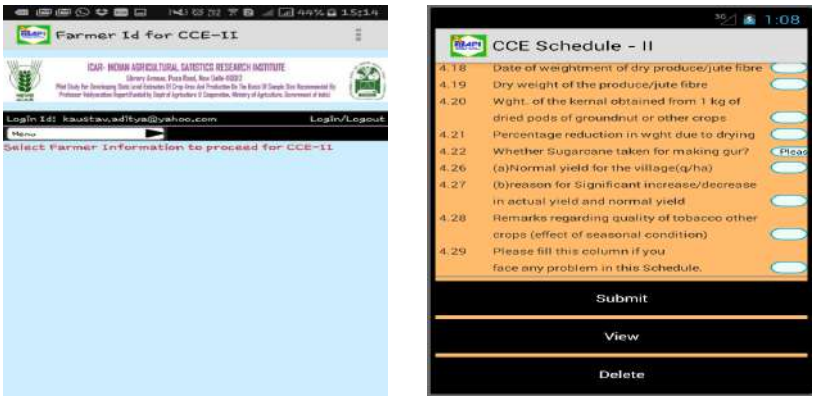


Figure 7. CCE Schedule II

3. Field Testing

Both offline and online version of the software is available for use but during field testing we used the offline version of the software as for the online version there is need of fast internet connection while rural India does not provide that support. We have implemented the MAPI software during Rabi 2015-16 two districts namely Bulandshahar and Pratapgarh of the state of Uttar Pradesh to validate the functionality and applicability of MAPI. In this survey, we have considered one tehsil/sub-division each from both the districts. The selected tehsil/sub-division from Bulandshahar and Pratapgarh districts are Bulandsahar and Kunda respectively. From the selected tehsils we have selected a total of 12 villages and 8 villages respectively for both Bulandsahar and Kunda. From each village we have to collect information about crop area and yield of the major food grain crops prevailing in that region. In the study area, the most predominant crop during the RABI season is Wheat while few pulses along with few oilseed crops mostly rapeseed and mustard were also cultivated. Almost 90% area is occupied by wheat crop in that state. In our proposed methodology of the project, we have to first select 100 survey numbers randomly from total survey numbers in each village in clusters of 5 survey numbers and then for each survey number we have to visit the farmer and collect information from him about crop area through enumeration schedule. This work is to be done within one month of sowing of the crop. Then 15 days before harvesting we have to select two survey numbers randomly from the selected 100 survey numbers and visit the selected survey number and collect information about the CCE plot demarcation for the crop persisting on that plot. If the survey number has no crop then we shift to the most adjacent survey number. Then on the day of harvesting another visit will be made to the farmer to record the weight of the produce using CCE schedule II. To verify that the developed software we have compared the MAPI data with the data collected through traditional PAPI survey in the selected district under the pilot project through the state officials of Department of Agriculture and Crop Insurance, Lukhnow, Govt. of Uttar Pradesh. For working with MAPI software we have hired two young professionals who were acquainted with the use of smart phones and tablets. Then they were trained with the process of data collection and methods of CCE and sent for the work of the data collection in both the selected tehsil mentioned above. It was found very efficient and less time consuming and more accurate than the ordinary paper based surveys.

References

1. Wikipedia, the Free Encyclopedia, 2016. Computer assisted personal interviewing (CAPI). Available at https://en.wikipedia.org/wiki/Computer-assisted_personal_interviewing.
2. Hanna, P., 2003. The Complete Reference JSP 2.0. Tata McGraw Hill Education Private Limited, New Delhi, India.

Overview of Open Source GIS Software-QGIS

Bharti

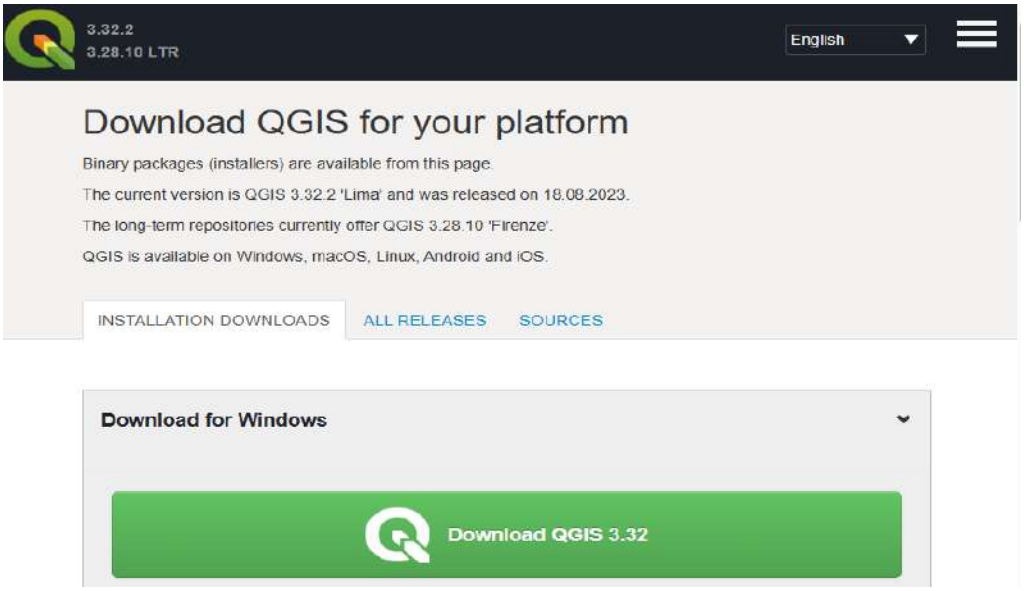
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

1. Introduction

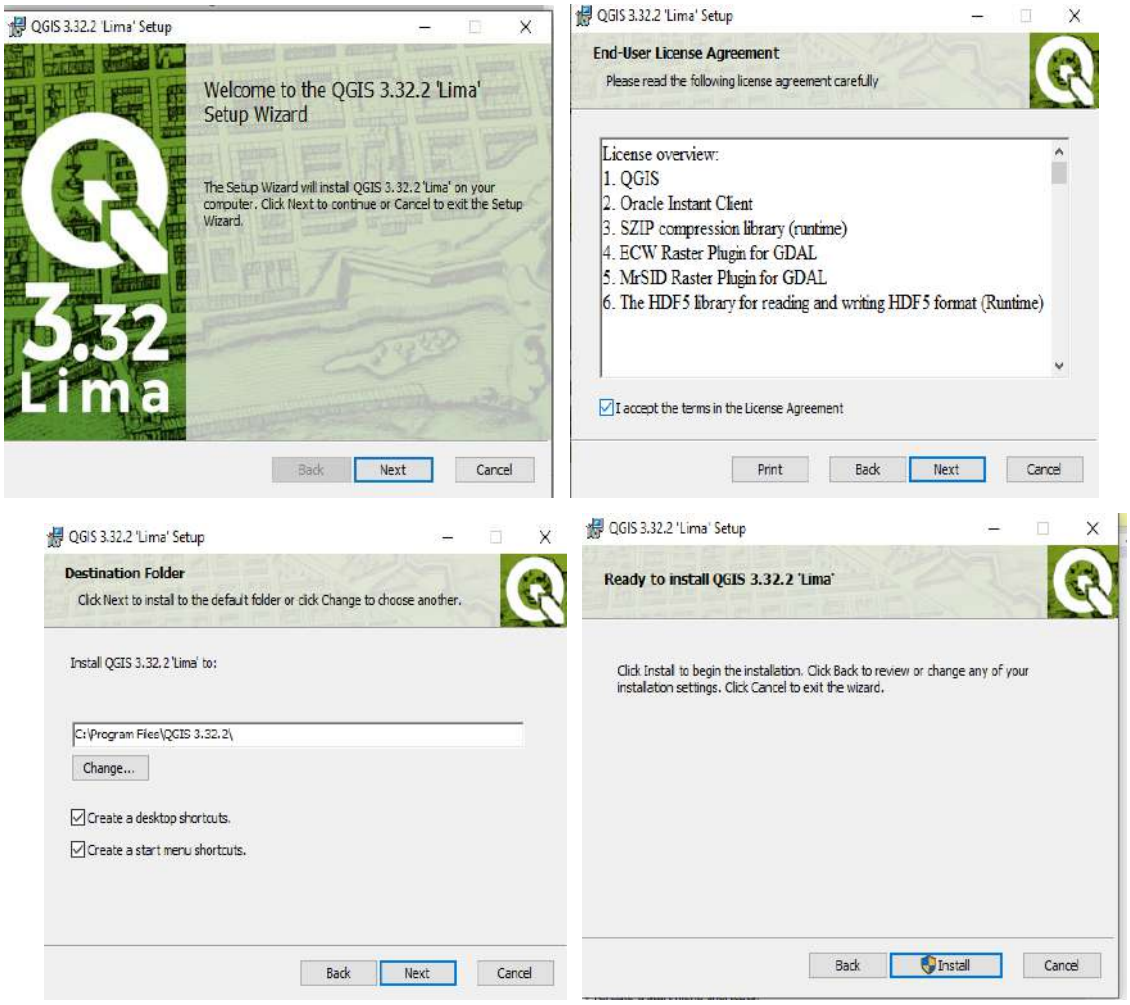
QGIS (Quantum GIS) is a free and open-source geospatial software used for analyzing and managing geographic data. It is a GIS (Geographic Information System) software designed to assist in working with both local and global data. QGIS supports vector and raster data, enabling graphical and spatial data analysis. QGIS offers various plugins that can enhance the software's features according to user needs. It supports multiple coordinate systems, allowing data from different locations to be accurately displayed. With QGIS, you can create interactive maps and manage metadata, ensuring data quality and consistency.

2. Downloading QGIS

- Visit the official QGIS website: https://qgis.org/



- Select "Download Now" or choose the supported version (Windows, macOS, Linux, etc.).
- On the download page, select the appropriate option based on your operating system and click the download button.
- Once the download is complete, open the downloaded file and start the installation process.



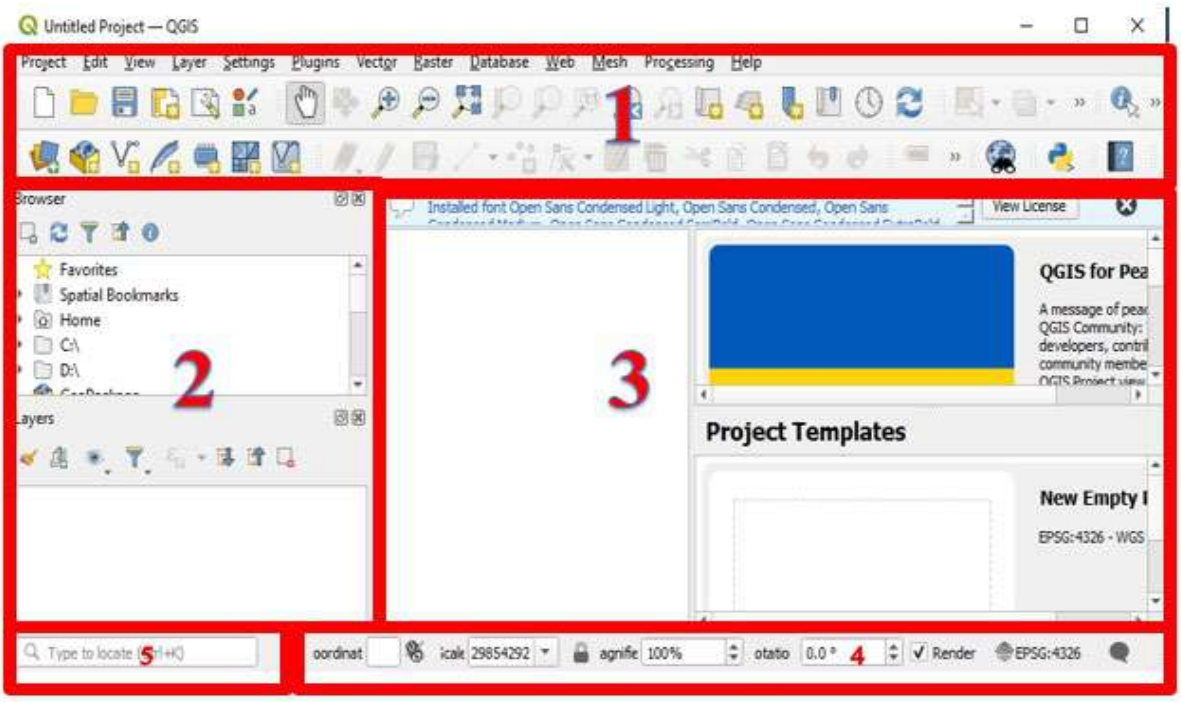
3. Different Versions of QGIS:

QGIS has released several versions, each with new and improved features and system updates. Some of the important QGIS versions include:

- QGIS 1.x Series
- QGIS 2.x Series
- QGIS 3.x Series
- QGIS 3.16 'Hannover'
- QGIS 3.18 'Zürich'

4. QGIS Interface:

- Menu Bar and Toolbars
- Layer Panel / Browser Panel
- Map Canvas
- Status Bar
- Locator Bar

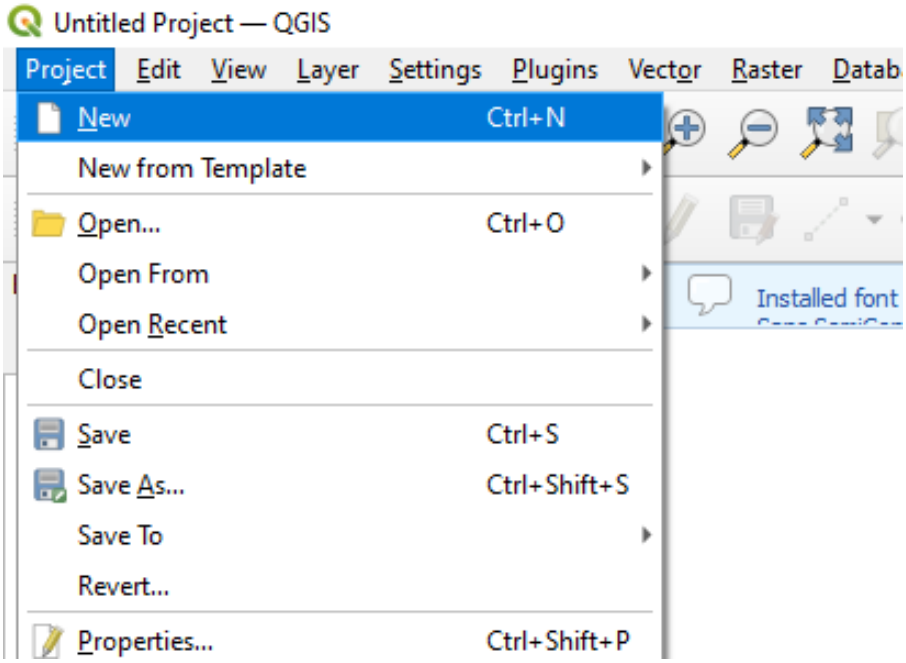


1. Menu Bar and Toolbars; 2. Layers List / Browser Panel; 3. Map canvas; 4. Status bar; 5. Locator bar

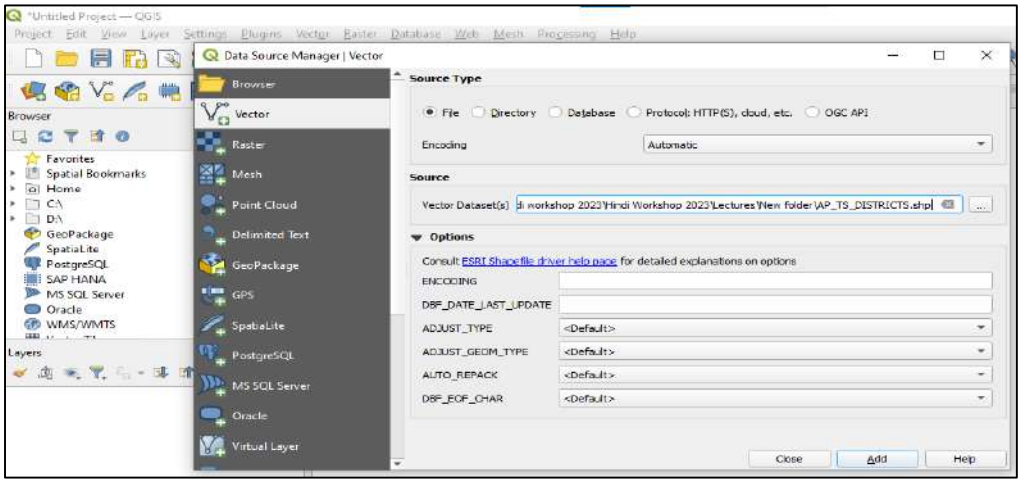
5. Data Import and Export in QGIS:

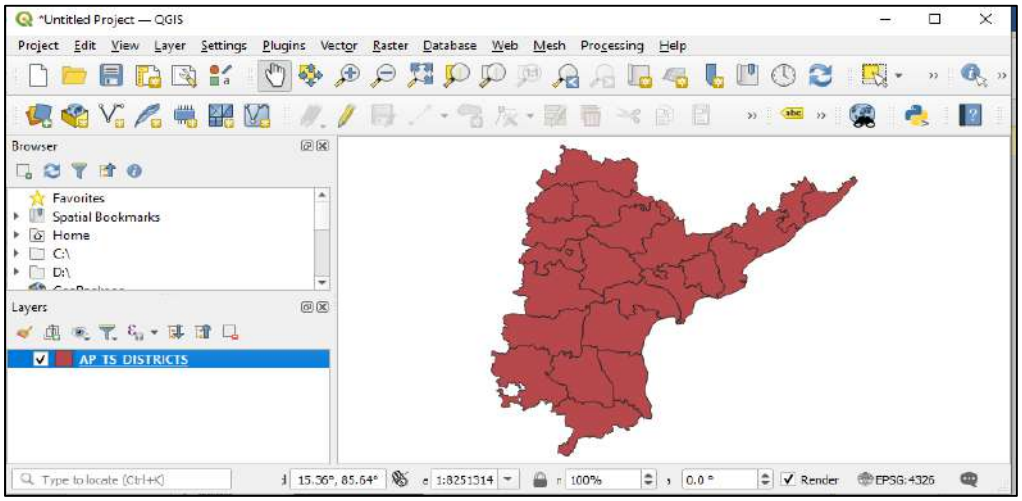
The steps to add the first layers in QGIS are:

- First, open the QGIS software.
- After opening the project, go to the "Project" menu and select "New".

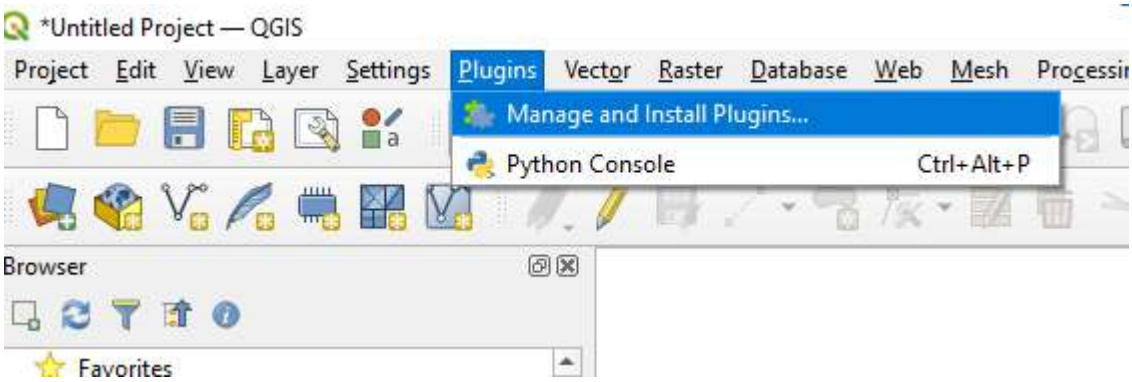


- After creating a new project, go to the "Layer" menu and select "Add Layer." Choose the type of data like vector, raster, Database, etc.



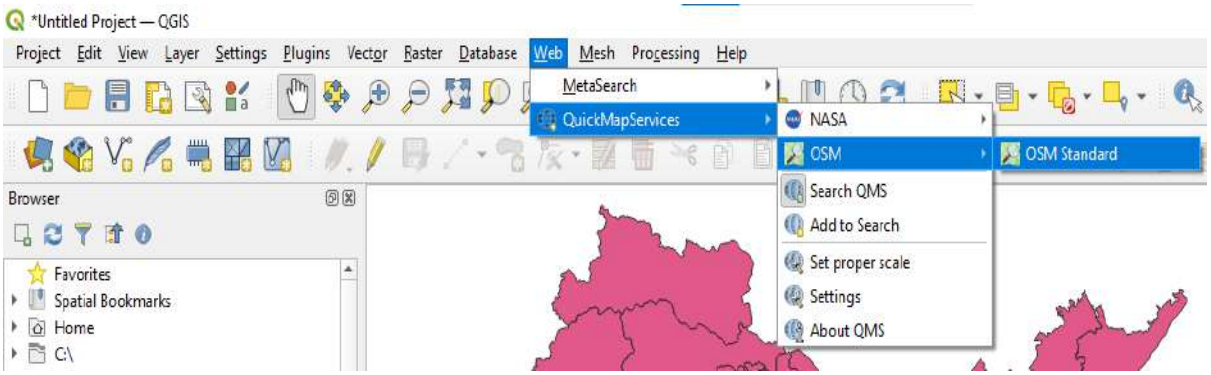


6. **Downloading Plugin in QGIS**
   - From the main QGIS menu, go to 'Plugins' and select 'Manage and Install Plugins.'
   - When the Plugin Manager opens, select the 'All' tab.



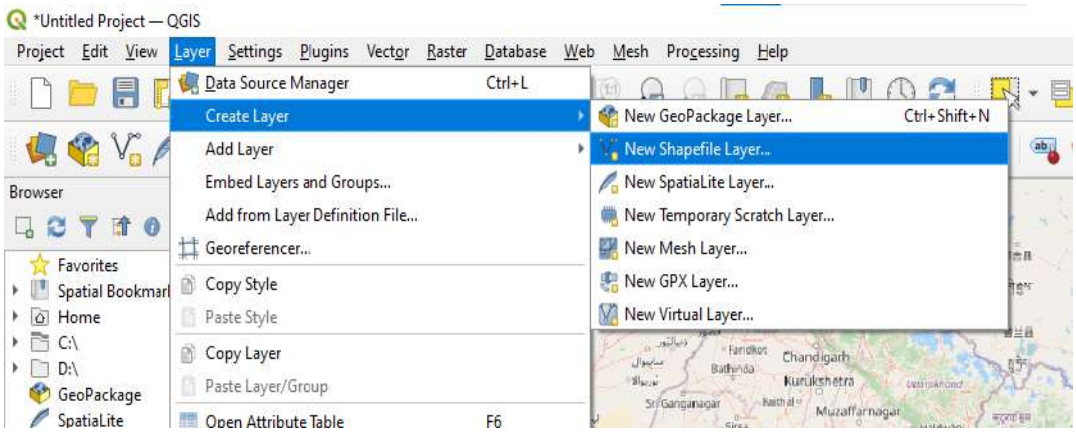
- In the search box, enter the name or description of the external plugin you are looking for. Based on your search, available plugins will appear in the list. Click on the plugin you want to download.
- After the plugin description opens, click the 'Install Plugin' button.

- Once the plugin is successfully installed, it may need to be activated. For this, go to the 'Plugins' menu, select 'Manage and Install Plugins,' and click on the plugin's name to activate it.

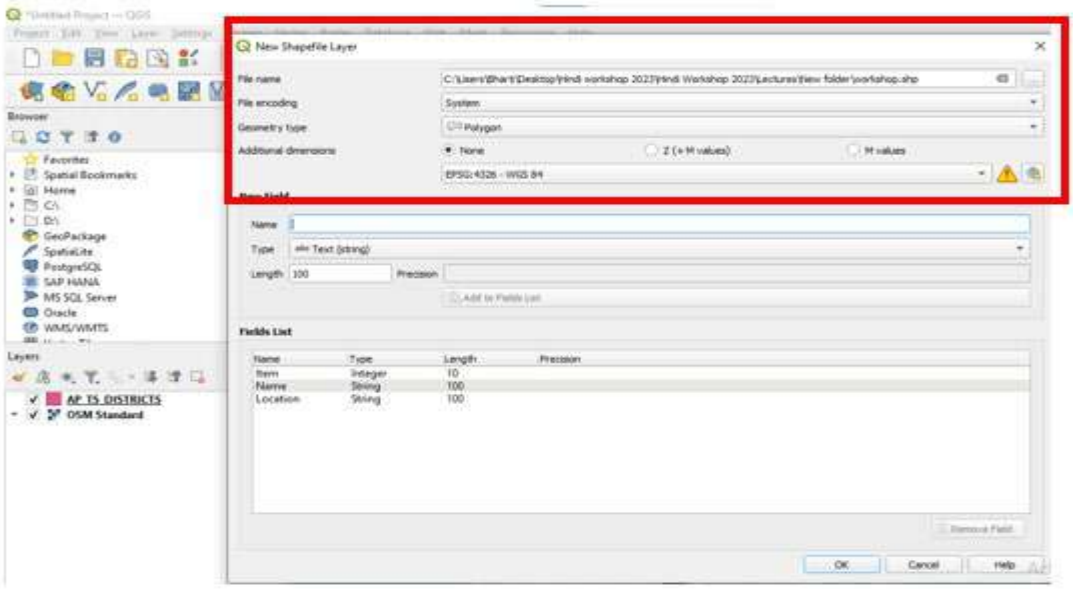


7. Create Layer in QGIS:

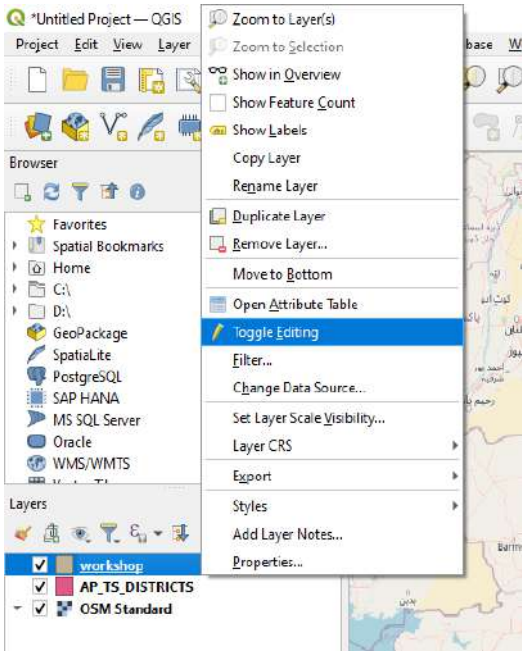
- From the main QGIS menu, go to 'Layer' and select 'Create Layer.'



- You will be shown options to choose the layer type, such as Point, Line, and Polygon. Choose the appropriate type based on your requirements.

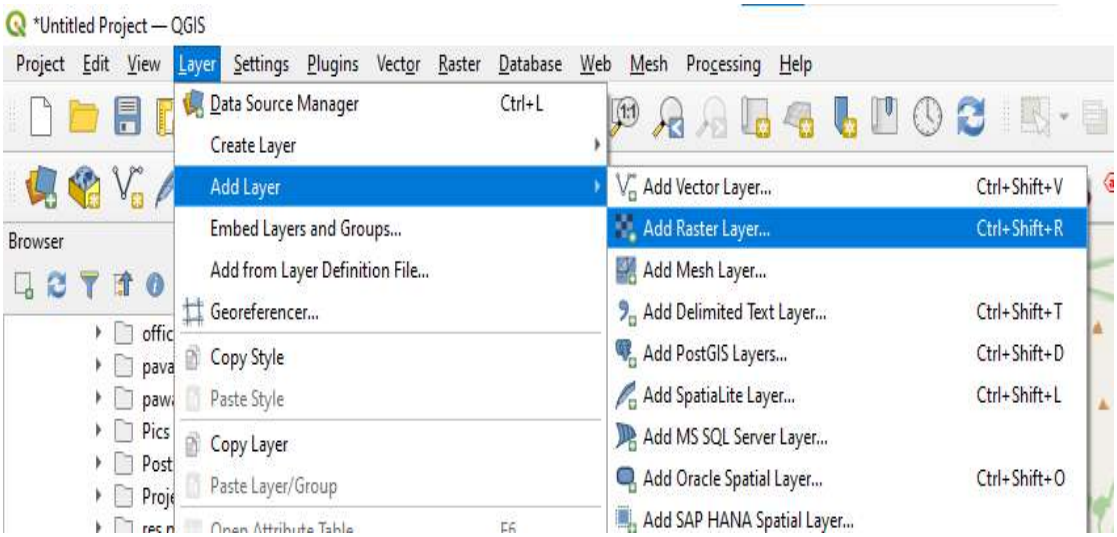


- After selecting the layer type, choose the coordinate system for the layer. Select the correct coordinate system. You may be shown various options for layer configuration, such as layer name, field structure, etc. Fill in the required details for layer configuration and click the 'OK' button.
- The new layer should now appear in QGIS, and you can use it for display and editing in the map.

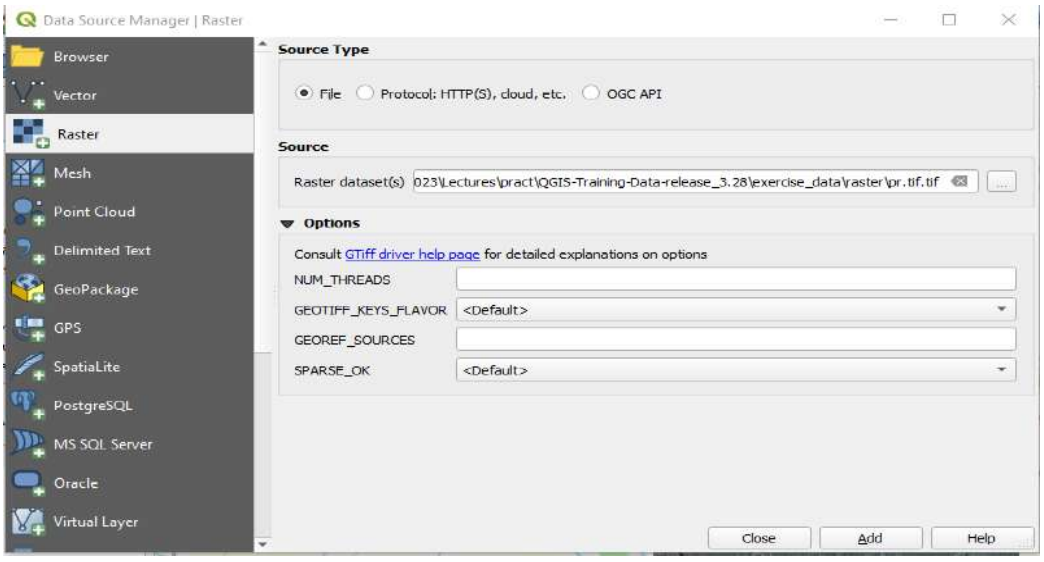


8. Add Raster Layer in QGIS:

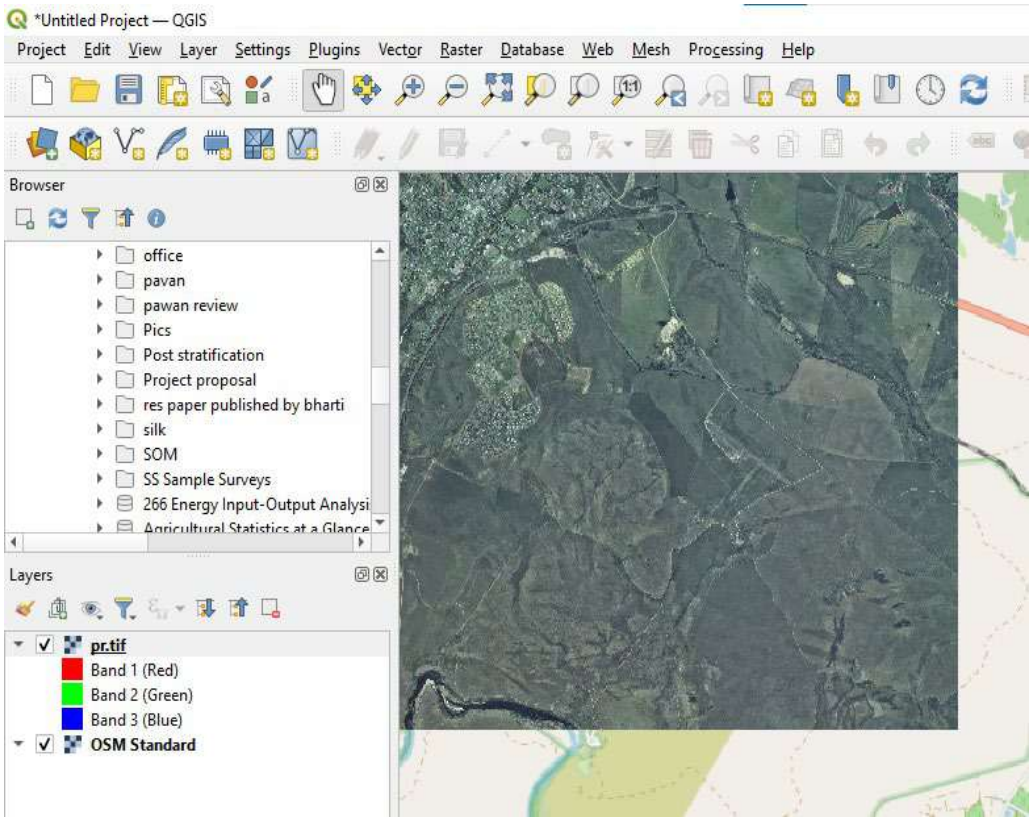
- After opening QGIS, click on the "Layer" option from the main menu and select "Add Raster Layer."



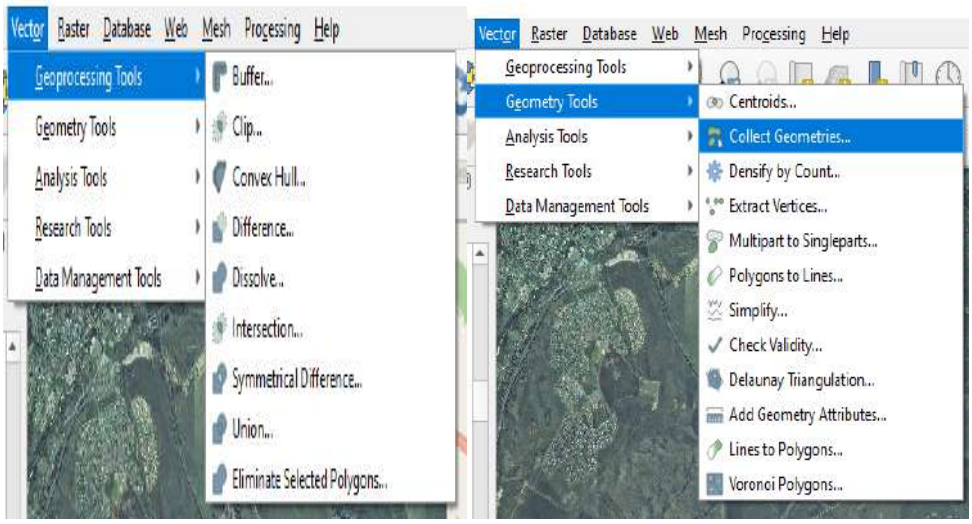
- After selecting "Add Raster Layer," a dialog box will appear asking you to browse and select the raster file stored on your system. Browse to the file path and select the raster file. Once you select the raster file, click the "Add" button.

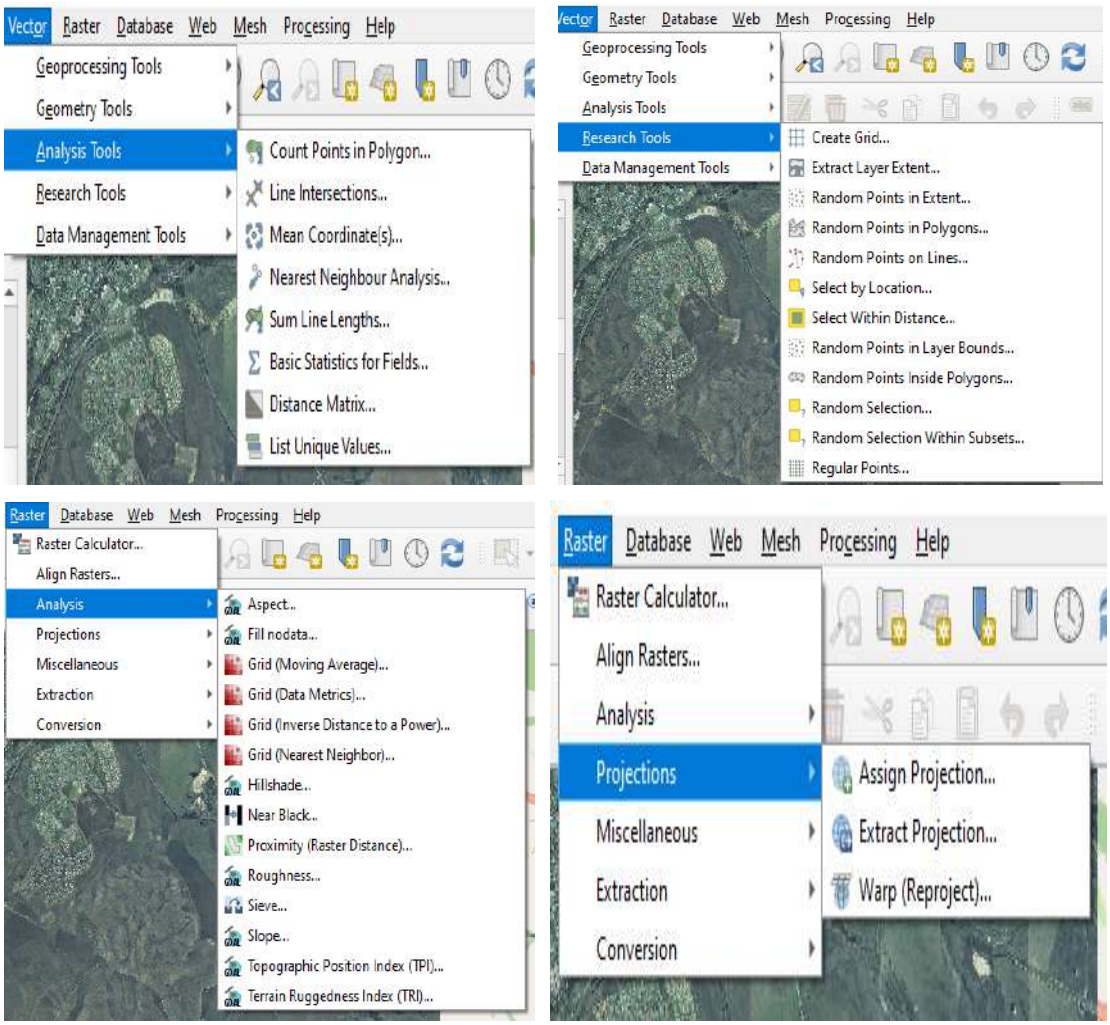


- Based on the selected raster file, a new raster layer will be added to QGIS project.



9. QGIS "Analysis Tools"





10. Application of QGIS in Agriculture:

- **Spatial Analysis:** QGIS can analyze spatial data, such as soil types and land-use patterns. This information can help in making informed decisions about crop selection, planting patterns, and irrigation strategies.
- **Precision Agriculture:** QGIS can create maps of field conditions by collecting satellite data from various sources. Farmers can use these maps to identify areas with varying levels of nutrients, moisture, or pests, allowing for efficient use of fertilizers, pesticides, and water.
- **Yield Prediction:** By analyzing historical yield data along with weather patterns, soil quality, and planting techniques, QGIS can be used to build predictive models for crop yields. This information helps farmers make better decisions about production and marketing.
- **Water Management:** QGIS can gather data related to water sources, irrigation systems, and weather patterns. Farmers can use this information to create optimal irrigation schedules and manage water resources effectively.
- **Environmental Impact Assessment:** When planning new agricultural projects, QGIS can be used to assess potential environmental impacts. It helps identify critical areas, habitats, and water sources that need protection.

OVERVIEW OF ARTIFICIAL INTELLIGENCE/MACHINE LEARNING

Chandan Kumar Deb

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

1. Introduction

Artificial Intelligence (AI) is a branch of computer science focused on creating intelligent machines that can simulate human behaviour. The concept dates back to antiquity, with early mentions in myths and folklore about artificial beings with human-like intelligence. However, the formal foundation of AI as a field began in the 20th century with theoretical and practical advancements in computing.

2. Historical Development of AI

Early Concepts (Antiquity to 19th Century)

The idea of artificial beings with intelligence can be traced to ancient myths and the works of philosophers such as Aristotle, who speculated about "intelligent artifacts." Leonardo da Vinci's designs for mechanical robots were early attempts to conceptualize artificial beings.

Emergence of Modern AI (20th Century)

In 1950, Alan Turing proposed the Turing Test to assess a machine's ability to exhibit intelligent behaviour indistinguishable from humans. Around the same time, McCulloch and Pitts introduced the first mathematical model of a neural network. The Dartmouth Conference of 1956, led by John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon, officially established AI as a field.

[John McCarthy, Source: Wikipedia]

Early Developments (1950s to 1970s)

- **Logic Theorist (1956):** The first AI program, capable of solving mathematical problems.
- **General Problem Solver (1957):** Developed by Newell and Simon to solve various problems.
- **Perceptrons (1950s-1960s):** Frank Rosenblatt's work laid the foundation for neural networks.
- **Expert Systems (1970s):** These used knowledge representation for problem-solving in specific domains.

AI Winter (1970s to 1980s)

AI research faced setbacks due to skepticism, overhyped expectations, and lack of computational power. Funding cuts led to a period of stagnation known as the "AI winter."

Resurgence (1980s to Present)

The 1980s saw renewed interest in AI with advancements in expert systems and neural networks. In the 21st century, machine learning, deep learning, and AI applications in various industries propelled AI into mainstream use. Major breakthroughs include speech recognition, computer vision, and natural language processing.

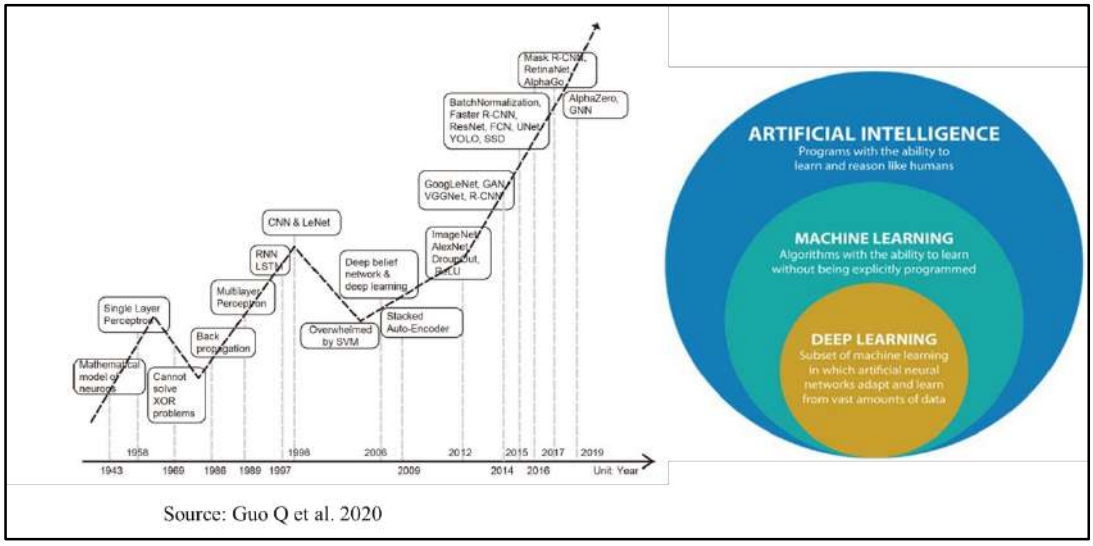


Figure: Deep Learning Timeline

3. Machine Learning

Machine learning (ML) is a subset of AI that enables computers to learn from data without explicit programming. Arthur Samuel defined it as “the field of study that gives computers the ability to learn without being explicitly programmed.” Tom Mitchell later refined this definition by incorporating performance improvement with experience.

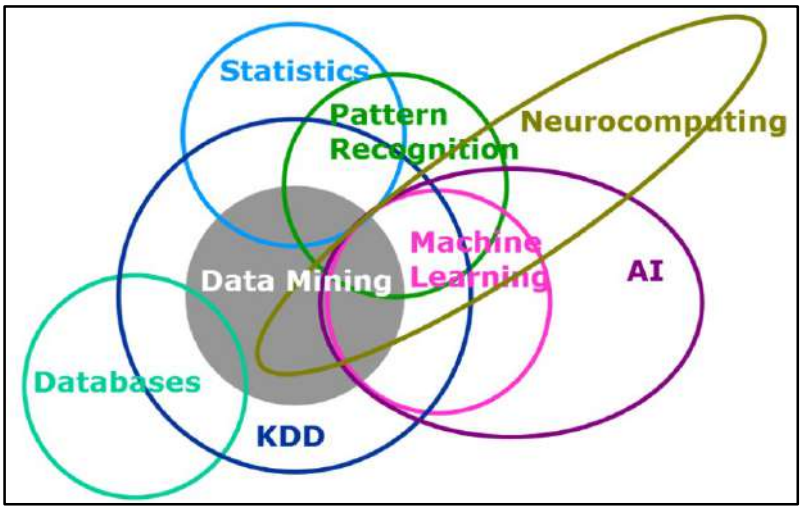


Figure: Machine Learning and other fields

Types of Machine Learning

1. **Supervised Learning** – Uses labeled data to train models.

- Examples: Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Neural Networks

2. **Unsupervised Learning** – Finds hidden patterns in unlabeled data.
   - Examples: Clustering algorithms such as k-Means and Hierarchical Cluster Analysis
3. **Semi-Supervised Learning** – Combines labeled and unlabeled data to improve learning.
4. **Reinforcement Learning** – Uses reward-based training to optimize decision-making.

4. Applications of AI

AI has revolutionized multiple industries by automating tasks, improving efficiency, and providing data-driven insights. Some major areas include:

Computer Vision in Agriculture

- **AI-DISC (Artificial Intelligence-based Disease Identification System for Crops)**
  - An AI-powered mobile app that identifies crop diseases using image processing.
  - Developed under NAHEP Component 2 and NASF Project.
  - Hosted on Krishi-Megh Cloud Infrastructure.
  - Uses deep learning models trained on 1.5 lakh images covering over 20 crops.

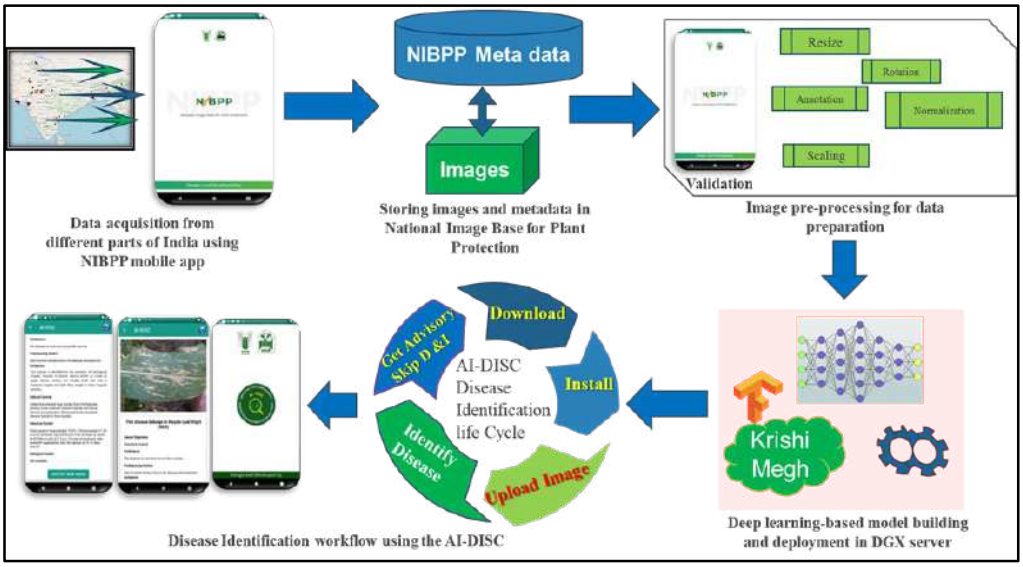


Figure: Life cycle of AI-DISC

AI in Livestock Management

- **AI-DISA (Artificial Intelligence-Based Disease Identification for Animals)**
  - Focuses on detecting diseases in bovines such as FMD, Mastitis, and LSD.
  - Uses deep learning-based object detection models.

  - Implements YOLOv5 models for lesion detection in images.
- **Pig Live Weight Monitoring**
  - CNN-based model trained on image datasets to estimate live pig weight.
  - Provides a mobile-based application for farmers to track livestock growth.

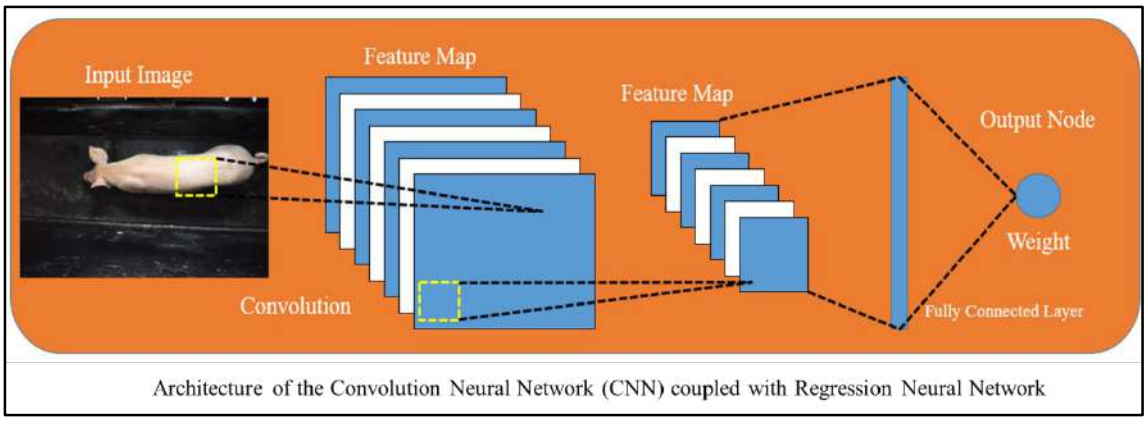


Figure: **A CNN model was developed using Keras, a high level API of Python backed by Tensorflow engine**

AI-Based Crop Protection

AI has contributed to disease and pest management in crops by integrating deep learning models for automated identification and advisory services. Some notable models include:

- **SpikeSegNet:** A deep learning model for wheat spike segmentation.
- **SlypNet:** A deep learning and image processing package for agricultural applications.
- **MuSiC V1.0:** A model used for mustard silique counting.
- **PanicleDet:** Utilized for segmentation in paddy yield estimation.

AI and Deep Learning in Object Detection

Object detection models have been widely adopted in AI-based applications for agriculture and livestock management.

- **Two-stage object detection:** Uses VGGNet for insect pest detection.
- **Single-stage object detection:** YOLO family models (YOLOv5s, YOLOv7, YOLOv8) are employed for rapid and efficient object detection.

Challenges and Future Directions

Despite its success, AI faces several challenges, including:

- **Data Dependency:** AI models require large datasets for training.
- **Computational Costs:** High-performance GPUs and cloud infrastructure are necessary.
- **Ethical Concerns:** AI-driven automation can lead to job displacement and privacy concerns.
- **Bias and Fairness:** AI models can inherit biases from training data, affecting decision-making.

Future Trends

- **Explainable AI (XAI):** Enhancing model interpretability.
- **Federated Learning:** Decentralized ML models that preserve privacy.
- **Edge AI:** Running AI algorithms on edge devices instead of central servers.
- **AI in Climate Resilience:** Using AI for sustainable agriculture and environmental monitoring.

Conclusion

Artificial Intelligence has evolved significantly from early concepts to modern applications in various industries. AI-powered solutions in agriculture, healthcare, finance, and transportation continue to reshape industries and enhance human capabilities. With continued advancements in deep learning, reinforcement learning, and ethical AI, the future of AI remains promising and full of potential for innovation.

Reference

Gole, P., Bedi, P., Marwaha, S., Haque, M. A., & Deb, C. K. (2023). TrIncNet: a lightweight vision transformer network for identification of plant diseases. Frontiers in Plant Science, 14

Nigam, S., Jain, R., Marwaha, S., Arora, A., Haque, M. A., Dheeraj, A., & Singh, V. K. (2023). Deep transfer learning model for disease identification in wheat crop. Ecological Informatics, 75, 102068.

Haque, M. A., Marwaha, S., Arora, A., Deb, C. K., Misra, T., Nigam, S., & Hooda, K. S. (2022). A lightweight convolutional neural network for recognition of severity stages of maydis leaf blight disease of maize. Frontiers in Plant Science, 13, 1077568.

Haque, M. A., Marwaha, S., Deb, C. K., Nigam, S., & Arora, A. (2023). Recognition of diseases of maize crop using deep learning models. Neural Computing and Applications, 35(10), 7407-7421.

Haque, M. A., Marwaha, S., Deb, C. K., Nigam, S., Arora, A., Hooda, K. S., ... & Agrawal, R. C. (2022). Deep learning-based approach for identification of diseases of maize crop. Scientific reports, 12(1), 6334.

Haque, M. A., Marwaha, S., Arora, A., Paul, R. K., Hooda, K. S., Sharma, A., & Grover, M. (2021). Image-based identification of maydis leaf blight disease of maize (Zea mays) using deep learning.

Misra, T., Arora, A., Marwaha, S., Chinnusamy, V., Rao, A. R., Jain, R., ... & Goel, S. (2020). SpikeSegNet-a deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. Plant methods, 16(1), 1-20.

Maji, A. K., Marwaha, S., Kumar, S., Arora, A., Chinnusamy, V., & Islam, S. (2022). SlypNet: Spikelet-based yield prediction of wheat using advanced plant phenotyping and computer vision techniques. Frontiers in Plant Science, 13, 889853.

Misra, T., Arora, A., Marwaha, S., Jha, R. R., Ray, M., Jain, R., ... & Chinnusamy, V. (2021). Web-SpikeSegNet: deep learning framework for recognition and counting of spikes from visual images of wheat plants. IEEE Access, 9, 76235-76247.

NEURAL NETWORK MODELLING

Girish Kumar Jha

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Producers of statistics have always been concerned about the quality of their statistics. Editing and imputation are often necessary to improve the quality of data collected from a survey or a census. The aim of editing is to identify the records that are unacceptable, then identify the values of such records that need to be corrected and then correct those values using imputation. When computers are used in this process, this is called automated editing and imputation. Fellegi and Holt (1976) developed methods for automatic editing and imputation of survey data using high-speed computers. They assumed that the edit specifications would be given explicitly by subject matter expert so that as few values as possible should be changed and that the imputation is of the hot deck type. It is estimated that 20 to 40 per cent of the total cost of a survey or a census are still used for editing even after extensive use of computerized editing. Hence, to look for new opportunities to reduce the resources and time required for editing is a continuous challenge. Recent progress in the field of artificial neural networks (ANN) and continued development in capacity and speed of computers indicate that editing can be reformulated with the help of neural networks which can be trained to edit and impute from a sample of records edited by experts rather than use of explicit edit and imputation rules.

In recent years neural computing has emerged as a practical technology, with successful applications in many fields as diverse as finance, medicine, engineering, geology, physics and biology. The excitement stems from the fact that these networks are attempts to model the capabilities of the human brain. From a statistical perspective neural networks are interesting because of their potential use in prediction and classification problems.

Artificial neural networks (ANNs) are non-linear data driven self adaptive approach as opposed to the traditional model based methods. They are powerful tools for modelling, especially when the underlying data relationship is unknown. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the outcome of new independent input data. ANNs imitate the learning process of the human brain and can process problems involving non-linear and complex data even if the data are imprecise and noisy. Thus they are ideally suited for the modeling of survey data which are known to be complex and often non-linear.

A very important feature of these networks is their adaptive nature, where “learning by example” replaces “programming” in solving problems. This feature makes such computational models very appealing in application domains where one has little or incomplete understanding of the problem to be solved but where training data is readily available.

These networks are “neural” in the sense that they may have been inspired by neuroscience but not necessarily because they are faithful models of biological neural or cognitive phenomena. In fact majority of the network are more closely related to traditional

mathematical and/or statistical models such as non-parametric pattern classifiers, clustering algorithms, nonlinear filters, and statistical regression models than they are to neurobiology models.

Neural networks (NNs) have been used for a wide variety of applications where statistical methods are traditionally employed. They have been used in classification problems, such as identifying underwater sonar currents, recognizing speech, and predicting the secondary structure of globular proteins. In time-series applications, NNs have been used in predicting stock market performance. As statisticians or users of statistics, these problems are normally solved through classical statistical methods, such as discriminant analysis, logistic regression, Bayes analysis, multiple regression, and ARIMA time-series models. It is, therefore, time to recognize neural networks as a powerful tool for data analysis.

The purpose of this lecture note is to provide an overview of ANNs and discuss how neural networks can be used effectively and efficiently in control and imputation of individual records of a data set. A detailed discussion on this topic is given by Roddick (1993) and Nordbotten (1995).

2. Characteristics of Neural Networks

- The NNs exhibit mapping capabilities, that is, they can map input patterns to their associated output patterns.
- The NNs learn by examples. Thus, NN architectures can be 'trained' with known examples of a problem before they are tested for their 'inference' capability on unknown instances of the problem. They can, therefore, identify new objects previously untrained.
- The NNs possess the capability to generalize. Thus, they can predict new outcomes from past trends.
- The NNs are robust systems and are fault tolerant. They can, therefore, recall full patterns from incomplete, partial or noisy patterns.
- The NNs can process information in parallel, at high speed, and in a distributed manner.

3. Basics of artificial neural networks

An artificial neural network is a set of simple computational units that are highly interconnected. The units are also called nodes and loosely represent the biological neuron. A graphical presentation of neuron is given in Figure 1. A neuron is an information processing unit that is fundamental to the operation of a neural network. The connections between nodes are unidirectional and are represented by arrows in the figure. These connections model the synaptic connections in the brain. Each connection has a weight called the synaptic weight, denoted as w_{kj}, associated with it. The synaptic weight, w_{kj}, is interpreted as the strength of the connection from the jth unit to the kth unit. Unlike a synapse in the brain, the synaptic weight of an artificial neuron may lie in a range that includes negative as well as positive values. If a weight is negative, it is termed inhibitory because it decreases the net input. If the weight is positive, the contribution is excitatory because it increases the net input.

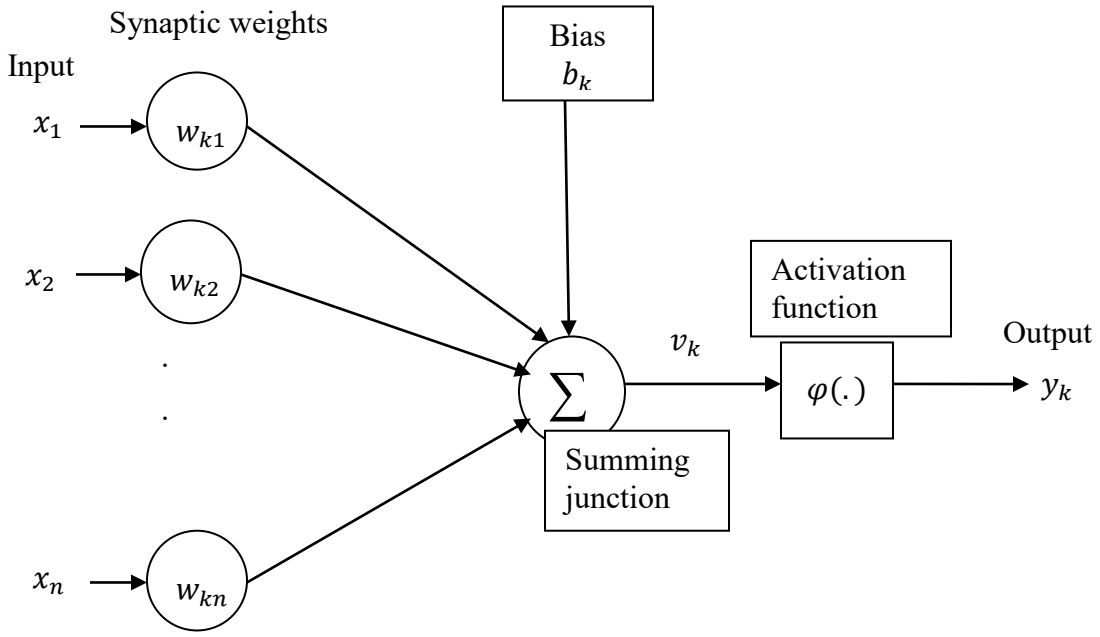


Figure 1: Nonlinear model of a neuron

The input into a node is a weighted sum of the outputs from nodes connected to it. Each unit takes its net input and applies an activation function to it. An activation function which is also known as squashing function, squashes or limits the amplitude range of the output of a neuron. The neuronal model of Figure 1 also includes an externally applied bias, denoted by b_k. The bias b_k has the effect of increasing or lowering the net input of the activation function depending on whether it is positive or negative respectively.

In mathematical terms, we may describe a neuron *k* by the following equations

$$y_k = \varphi(v_k) = \varphi\left(\sum_{j=1}^{n} w_{kj}x_j + b_k\right)$$

where $x_1, x_2, \ldots, x_n$ are the input patterns, $w_{k1}, w_{k2}, \ldots, w_{kn}$ are the synaptic weights of neuron k, b_k is the bias, $\varphi(.)$ is the activation function and y_k is the output of the neuron. The sigmoid function, whose graph is *s*-shaped, is by far the most common form of activation function used in the construction of artificial neural networks. The neural networks are built from layers of neurons connected so that one layer receives input from the preceding layer of neurons and passes the output on to the subsequent layer.

4. Neural networks architectures

An artificial neural network is defined as a data processing system consisting of a large number of simple highly inter connected processing elements (artificial neurons) in an architecture inspired by the structure of the cerebral cortex of the brain. There are several types of architecture of neural networks. However, the two most widely used NNs are discussed below:

Feed forward networks

In a feed forward network, information flows in one direction along connecting pathways, from the input layer via the hidden layers to the final output layer. There is no feedback (loops) i.e., the output of any layer does not affect that same or preceding layer.

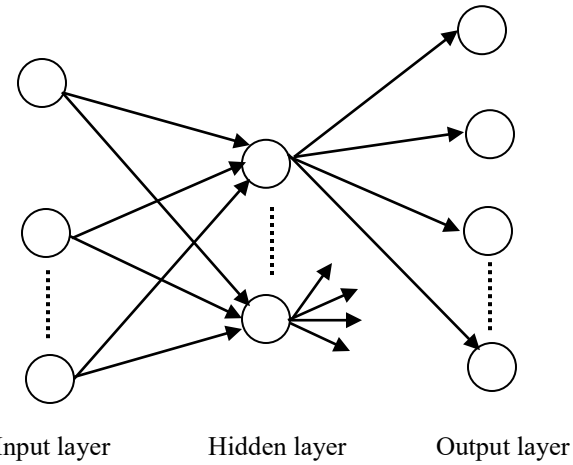


Figure 2: A multi-layer feed forward neural network

Recurrent networks

These networks differ from feed forward network architectures in the sense that there is at least one feedback loop. Thus, in these networks, for example, there could exist one layer with feedback connections as shown in figure below. There could also be neurons with self-feedback links, i.e. the output of a neuron is fed back into itself as input.

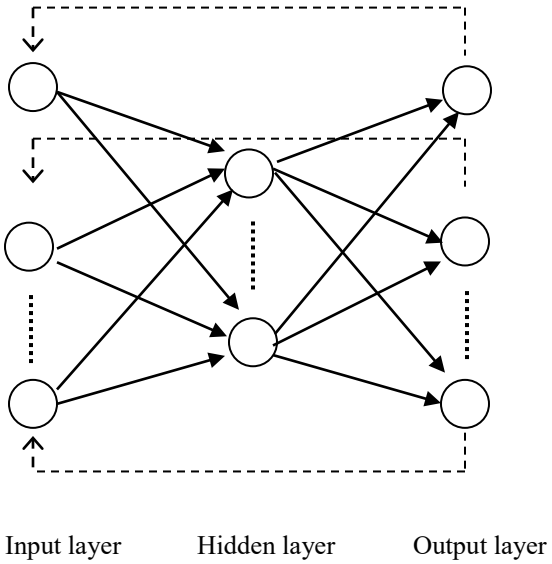


Figure 3: A recurrent neural network

Learning/Training methods

Learning methods in neural networks can be broadly classified into three basic types: supervised, unsupervised and reinforced.

Supervised learning

In this, every input pattern that is used to train the network is associated with an output pattern, which is the target or the desired pattern. A teacher is assumed to be present during the learning process, when a comparison is made between the network's computed output and the correct expected output, to determine the error. The error can then be used to change network parameters, which result in an improvement in performance.

Unsupervised learning

In this learning method, the target output is not presented to the network. It is as if there is no teacher to present the desired patterns and hence, the system learns of its own by discovering and adapting to structural features in the input patterns.

Reinforced learning

In this method, a teacher though available, does not present the expected answer but only indicates if the computed output is correct or incorrect. The information provided helps the network in its learning process. A reward is given for a correct answer computed and a penalty for a wrong answer. But, reinforced learning is not one of the popular forms of learning.

Types of neural networks

The most important class of neural networks for real world problems solving includes

- Multilayer Perceptrons
- Radial Basis Function Networks
- Kohonen Self Organizing Feature Maps

Multilayer Perceptrons

The most popular form of neural network architecture is the multilayer perceptrons (MLP) which is a generalization of the single-layer perceptron. Typically, the MLP network consists of a set of source nodes that constitute the input layer, one or more hidden layers of computation nodes and an output layer of computation nodes. The input signal propagates through the network in a forward direction on a layer by layer basis. MLP have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. A multilayer perceptron has three distinctive characteristics:

- The model of each neuron in the network includes a nonlinear activation function which should also be a differentiable everywhere. A commonly used form of nonlinearity that satisfies this requirement is a sigmoidal nonlinearity. The presence of nonlinearities is important because otherwise the input-output relation of the network could be reduced to that of a single layer perceptron.
- The network contains one or more layers of hidden neurons that are not part of the

input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input pattens.

- The network exhibits a high degree of connectivity determined by the synapses of the network. A change in the connectivity of the network requires a change in the population of synaptic connections or their weights.

Given enough data, enough hidden units, and enough training time, an MLP with just one hidden layer can learn to approximate virtually any function to any degree of accuracy. (A statistical analogy is approximating a function with *n*th order polynomials.) For this reason MLPs are known as universal approximators and can be used when we have little prior knowledge of the relationship between inputs and targets. Although one hidden layer is always sufficient provided we have enough data, there are situations where a network with two or more hidden layers may require fewer hidden units and weights than a network with one hidden layer, so using extra hidden layers sometimes can improve generalization.

5. Radial Basis Function Networks

Radial basis function (RBF) networks have a very strong mathematical foundation rooted in regularization theory for solving ill-conditioned problems. RBF networks, almost invariably, consists of three layers: a transparent input layer, a hidden layer with sufficiently large number of nodes and an output layer. As its name implies, radially symmetric basis function is used as activation function of hidden nodes. The transformation from the input nodes to the hidden nodes is non-linear one and training of this portion of the network is generally accomplished by an unsupervised fashion. The training of the network parameters between the hidden and output layers occurs in asupervised fashion based on target outputs.

MLPs are said to be distributed-processing networks because the effect of a hidden unit can be distributed over the entire input space. On the other hand, Gaussian RBF networks are said to be local-processing networks because the effect of a hidden unit is usually concentrated in a local area centered at the weight vector.

6. Kohonen Neural Network

Self Organizing Feature Map (SOFM, or Kohonen) networks are used quite differently to the other networks. Whereas all the other networks are designed for supervised learning tasks, SOFM networks are designed primarily for unsupervised learning (Patterson, 1996).

The principal goal of the SOFM is to transform an incoming signal pattern of arbitrary dimension into a one-or two dimensional discrete map, and to perform this transformation adaptively in a topologically ordered fashion. Figure 4 shows the schematic diagram of a two-dimensional lattice of neurons commonly used as the discrete map. Each neuron in the lattice is fully connected to all the source nodes in the input layer. This network represents a feedforward structure with a single computational layer consisting of neurons arranged in rows and columns. A one-dimensional lattice is a special case of the configuration depicted in Fig. 4: in this special case the computational layer consists simply of a single column or row of neurons.

Each input pattern presented to the network typically consists of a localized region or “spot” of activity against a quiet background. The location and nature of such a spot usually varies

from one realization of the input pattern to another. All the neurons in the network should therefore be exposed to a sufficient number of different realizations of the input pattern to ensure that the self-organization process has a chance to mature properly.

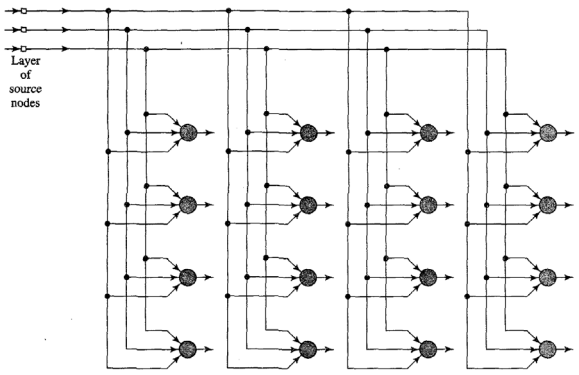


Fig. 4: Two-dimensional lattice of neurons

The algorithm responsible for the formation of the self-organizing map proceeds first by initialing the synaptic weights in the network. This can be done by assigning them small values picked from a random number generator, in so doing, no prior order is imposed on the feature map. Once the network has been properly initialized, there are three essential processes involved in the formation of the self-organizing map, as summarized here:

- *Competition*: For each input pattern, the neurons in the network compute their respective values of a discriminant function. This discriminant function provides the basis for competition among the neurons. The particular neuron with the largest value of discriminant function is declared winner of the competition.
- *Cooperation*: The winning neuron determines the spatial location of a topological neighborhood of excited neurons, thereby providing the basis for cooperation among such neighboring neurons.
- *Synaptic Adaptation*: This last mechanism enables the excited neurons to increase their individual values of the discriminant function in relation to the input pattern through suitable adjustments applied to their synaptic weights. The adjustments made are such that the response of winning neuron to the subsequent application of similar input pattern in enhanced.

The processes of competition and cooperation are in accordance with two of the four principles of self-organization. As for the principle of self-amplification, it comes in a modified form of Hebbian learning in the adaptive process. The presence of redundancy in the input data is needed for learning since it provides knowledge.

One possible use of SOFM is in exploratory data analysis. A second possible use is in novelty detection. SOFM networks can learn to recognize clusters in the training data, and respond to it. If new data, unlike previous cases, is encountered, the network fails to recognize it and this indicates novelty.

7. Neural networks for editing and imputation

Neural networks have proven to be effective mapping tools for a wide variety of problems and consequently they have been used extensively by practitioners in almost every application domain ranging from agriculture to zoology. Since neural networks are best at

identifying patterns or trends in data, they are well suited for editing and imputation applications.

Different types of sampling and non-sampling errors are introduced in preparation of statistics. The survey design introduces design errors, data collection introduces register, sampling, interviewer and response errors while the aggregation and dissemination processes are the sources of processing and presentation errors. Statistical data editing plays a special role in preparation of statistics because it aims at reducing the effect of errors in other statistical processes. Editing can be of two different types, logical (pre-defined rules must be obeyed) and statistical (a value is unlikely). In this case, we are concerned with evaluation of overall editing performance that is detection of data fields with errors. There are two performance requirements for editing. They include efficient error detection and influential error detection. Error detection should be evaluated in terms of both the number of errors correctly identified and the number of incorrect detections it makes. Imputation is the process by which missing or suspicious values are replaced. Here we concern ourselves with assessing the imputation of identifiable missing values. Ideally an imputation procedure should be capable of effectively reproducing the key outputs that would have been obtained from “complete data”.

The multi-layer feed forward neural networks are mainly used for editing and imputation. This network can be used for both detecting errors and imputing missing values. There are two approaches for error detection. The first one considers the presence or absence of an error as target variable. For this approach the presence of both clean and perturbed datasets are required for training the networks. By comparing clean and perturbed data, an indicator of presence/absence of errors for each variable is calculated. The network is trained on the perturbed data with the indicator variable. The other approach consists in considering as target variable the variable itself. If the predicted value differs from the actual value then it can be considered erroneous. As far as the imputation process is concerned, neural networks model with target variable equal to the variable itself are trained on those records for which the target value is not missing, and the networks thus generated are applied for imputing missing values.

The large number of parameters that must be selected to develop a neural network model for any application indicates that the design process still involves much trial and error. The next section provides a practical introductory guide for designing a neural network model.

8. Development of an ANN model

The various steps in developing a neural network model are

A. Variable selection

The input variables important for modeling variable(s) under study are selected by suitable variable selection procedures.

B. Formation of training, testing and validation sets

The data set is divided into three distinct sets called training, testing and validation sets. The training set is the largest set and is used by neural network to learn patterns present in the data. The testing set is used to evaluate the generalization ability of a supposedly trained network. A final check on the performance of the trained network is made using validation set.

C. Neural network architecture

Neural network architecture defines its structure including number of hidden layers, number of hidden nodes and number of output nodes etc.

- Number of hidden layers: The hidden layer(s) provide the network with its ability to generalize. In theory, a neural network with one hidden layer with a sufficient number of hidden neurons is capable of approximating any continuous function. In practice, neural network with one and occasionally two hidden layers are widely used and have to perform very well.
- Number of hidden nodes: There is no magic formula for selecting the optimum number of hidden neurons. However, some thumb rules are available for calculating number of hidden neurons. A rough approximation can be obtained by the geometric pyramid rule proposed by Masters (1993). For a three layer network with n input and m output neurons, the hidden layer would have *sqrt(n*m)* neurons.
- Number of output nodes: Neural networks with multiple outputs, especially if these outputs are widely spaced, will produce inferior results as compared to a network with a single output.
- Activation function: Activation functions are mathematical formulae that determine the output of a processing node. Each unit takes its net input and applies an activation function to it. Non linear functions have been used as activation functions such as logistic, tanh etc. The purpose of the transfer function is to prevent output from reaching very large value which can 'paralyze' neural networks and thereby inhibit training. Transfer functions such as sigmoid are commonly used because they are nonlinear and continuously differentiable which are desirable for network learning.

D. Evaluation criteria

The most common error function minimized in neural networks is the sum of squared errors. Other error functions offered by different software include least absolute deviations, least fourth powers, asymmetric least squares and percentage differences.

E. Neural network training

Training a neural network to learn patterns in the data involves iteratively presenting it with examples of the correct known answers. The objective of training is to find the set of weights between the neurons that determine the global minimum of error function. This involves decision regarding the number of iteration i.e., when to stop training a neural network and the selection of learning rate (a constant of proportionality which determines the size of the weight adjustments made at each iteration) and momentum values (how past weight changes affect current weight changes).

9. Conclusion

The computing world has a lot to gain from neural networks. Their ability to learn by example makes them very flexible and powerful. A large number of claims have been made about the modeling capabilities of neural networks, some exaggerated and some justified. Hence, to best utilize ANNs for different problems, it is essential to understand the potential as well as pitfalls of neural networks. Despite experiments with change of parameters and

topologies of the networks employed, it is difficult to suggest general approach for optimizing the networks. The same type of network may have relatively good performance with respect to some variables and very poor performance with respect to others for editing and imputation purposes. For some tasks, neural networks will never replace conventional methods, but for a growing list of applications, the neural architecture will provide either an alternative or a complement to these existing techniques. Finally, I would like to state that even though neural networks have a huge potential we will only get the best of them when they are integrated with Artificial Intelligence, Fuzzy Logic, Particle Swarm Optimization and related subjects.

10. Practical Exercise using SPSS

The Neural Network add-on module of SPSS provides two type of network architecture namely multilayer perceptron (MLP) and radial basis function (RBF) networks.

Creating a Multilayer Perceptron Network

From the menus choose:

Analyze

Neural Networks

Multilayer Perceptron...

Creating a Radial Basis Function Network

From the menus choose:

Analyze

Neural Networks

Radial Basis Function...

The details of construction and training of a feed forward neural network will be illustrated through a predictive application using the Neural Network add-on module of SPSS in the class.

References

Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, 9, 2-54.

Fellegi, I. P. & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Jha, G. K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. *Neural Computing and Application*s, 24, 563-571.

Haykin, S. (2006). Neural Networks: A comprehensive foundation, Pearson Prentice Hall.

Kaastra, I. and Boyd, M.(1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10, 215-236.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data, Survey Methodology, 12, 1-16.

Nordbotten, S. (1995). Editing statistical records by neural networks, Jo*urnal of Official Statistics*, 11, 391-411.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*, 12, 385-401.

Rao, J.N.K (1999): Some current trends in sample survey and methods. *Sankhya,* 61, Series B, 1-57.

Roddick, L.H. (1993). Data editing using neural networks. Technical Report, Systems Development Division, Statistics Canada.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage ang organization in the brain. *Psychological review*, 65, 386-408.

Rumelhart, D.E., Hinton, G.E and Williams, R.J. (1986). "Learning internal representation by error propagation", in Parallel distributed processing: Exploration in microstructure of cognition, Vol. (1) *(D.E. Rumelhart, J.L. McClelland and the PDP research gropus, edn.) Cambridge, MA: MIT Press*, 318-362.

Warner, B. and Misra, M. (1996). Understanding neural networks as statistical Tools. *The American Statistician*, 50, 284-293.

Zhang, G., Patuwo, B. E. and Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* 14, 35-62.

APPLICATION OF AI/ML IN CROP YIELD ESTIMATION

Pankaj Das

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

Abstract

Estimating crop yield is a critical aspect of modern agriculture, and the integration of machine learning models has proven to be instrumental in enhancing the accuracy and efficiency of these predictions. Various machine learning (ML) models are employed to analyze a multitude of factors that affect crop yield, enabling farmers to make informed decisions and optimize their agricultural practices. The machine learning models play a pivotal role in crop yield estimation by leveraging historical data, satellite imagery, and real-time sensor data. The integration of these models empowers farmers with accurate predictions, allowing for timely interventions and optimized resource allocation, ultimately contributing to improved agricultural productivity and sustainability.

Keyword: Crop yield; Machine Learning; Crop Yield Estimation; Agricultural Productivity; Sustainability.

Introduction

Agriculture is a cornerstone of the global economy, providing livelihoods for billions of people and ensuring the availability of food worldwide. In the agrarian nation of India, almost 60% of rural households depend on agriculture as their primary source of income. Food consumption is predicted to rise sharply in conjunction with the world's population growth, increasing pressure on agricultural systems to improve sustainability and productivity. Accurately predicting crop yields has become a critical need for addressing these challenges, enabling stakeholders to make informed decisions and plan effectively.

Crop Yield and Importance of Crop Yield Estimation

Crop yield refers to the amount of agricultural produce harvested per unit area of land. It is a critical measure of agricultural productivity and serves as an essential indicator of food security, economic stability, and environmental sustainability. High crop yields contribute to reducing hunger and poverty, enhancing farmers' income, and supporting global trade. Conversely, low yields can lead to food shortages, economic instability, and increased pressure on natural resources as efforts are made to cultivate more land to meet demand. Geographical variations in weather, soil, and agro management techniques cause crop yield variability. Therefore, an in-season crop estimation can help farmers adopt timely actions to increase productivity and is beneficial for government organizations to develop effective planning.

Accurate crop yield estimation benefits various stakeholders, including:

- **Economic Growth:** Agriculture is a significant contributor to the GDP of many countries. Higher yields lead to increased income for farmers and a robust agricultural economy. Increased income for farmers leads to better living standards in rural areas, reducing poverty and migration to cities. A thriving agricultural sector creates jobs not only on farms but also in related industries like transportation, food processing, and agricultural machinery. Countries with high agricultural productivity can export surplus crops, boosting their economy and foreign exchange reserves.
- **Food Security:** Accurate yield estimation ensures adequate food supply, helps prevent shortages, and enables effective distribution of resources. Increased yields lead into

more food for the population now as well as in the future. Accurate yield prediction helps prevent sudden food shortages or price spikes due to unexpected or epidemic events like droughts or disease pest infestation etc.

- **Environmental Sustainability:** Optimization of crop yields can reduce the need for excessive use of water, fertilizers, and pesticides, minimizing environmental pollution and degradation. The conversion of new forests, grasslands, or wetlands into agricultural land can be avoided by increasing the yields on the cropland that is already there. Sustainable agricultural practices that improve yields can also help reduce greenhouse gas emissions and enhance carbon sequestration in the soil. This protects biodiversity and crucial ecosystems.
- **Policy Planning:** Governments and organizations rely on yield estimates to plan for emergencies, allocate resources, and develop agricultural policies. Accurate yield data allows governments to make informed decisions about agricultural subsidies, trade policies, and food reserves. Yield estimates help in planning for potential food shortages due to natural disasters or conflicts, enabling timely interventions and aid distribution. Understanding yield trends helps prioritize research and development efforts in agriculture, leading to further improvements in productivity and sustainability.

In short, Crop yield is a critical factor that affects not only our food supply but also our economy, environment, and overall well-being. By focusing on increasing and accurately predicting crop yields, we can create a more food-secure, prosperous, and sustainable future for all.

Machine Learning:

Machine learning (ML) is a branch of artificial intelligence where computers learn from data without explicit programming. Machine learning is branch of applied science that improves the performance of a machine or model by providing the power to mimic like human brain for solving complex problems. The ML models are mostly data driven, self-adaptive and nonlinear in nature. These models gained significance importance in various fields like medical, data science etc. due to its ability of identifying hidden data patterns and improving with experience. The machine learning algorithms identify patterns, make predictions, and improve their performance over time through experience and exposure to more data. This learning process involves training a model on a dataset, allowing it to recognize underlying relationships and make informed decisions on new, unseen data. ML models make predictions and automate tasks like image recognition, language processing, and fraud detection. Key types include supervised learning (using labeled data), unsupervised learning (finding patterns in unlabeled data), and reinforcement learning (learning through trial and error).

Process of ML Model Building:

Building a machine learning model is a multi-stage process (Figure 1). It begins with clearly defining the problem and setting measurable objectives. Next, relevant data is collected, cleaned, explored, and preprocessed to handle inconsistencies and extract meaningful insights. Feature engineering follows, where relevant features are selected and transformed to optimize model performance. An appropriate machine learning model is then chosen, trained on a portion of the data, and its hyperparameters are tuned for optimal results. Each model has different hyperparameters. The model's performance is rigorously evaluated and validated on separate data to ensure accuracy and generalization. The model accuracy measures for root mean squared error (RMSE), mean absolute deviation (MAD), mean

absolute percentage error (MAPE) and maximum error (ME) were used to select the best model. Finally, the trained model is deployed to a production environment and continuously monitored to maintain its effectiveness over time, often requiring retraining with new data and updates as needed.

Key Features for Crop Yield Prediction

Machine learning models for crop yield estimation rely on a variety of features, including:

1. **Weather Data:** Temperature, precipitation, humidity, and solar radiation.
2. **Soil Properties:** pH, organic matter, nutrient levels, and soil type.
3. **Crop Management Practices:** Irrigation, fertilizer use, and pest control methods.
4. **Remote Sensing Data:** Satellite imagery and drone data providing vegetation indices like NDVI (Normalized Difference Vegetation Index).
5. **Historical Yield Data:** Past crop yield information for trend analysis.

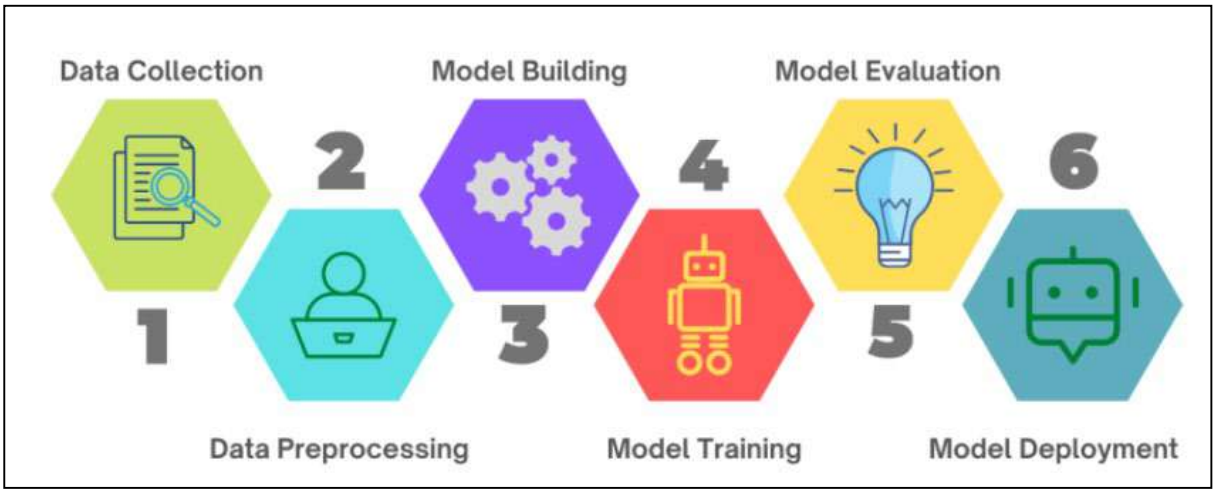


Figure 1: Process of ML Model Building

Machine Learning Techniques for Crop Yield Estimation

The crop yield is affected by multiple factors such as physical, economic and technological. These play a crucial role in yield enhancement as well as reduction. (Das et al., 2023). A good prediction model explores the complex relationship between different factors and yield. It helps to improve management techniques and boost actual yields. A good prediction model should be reliable, consistent, object-oriented, cost effective and sensitive to extreme events (Das et al., 2023; Bharti et al., 2023). Crop yield prediction is complex process due to its multifactorial behaviour. Several researchers have attempted to model the crop yield using different models such as simple correlation, path analysis, multiple linear regression, stepwise regression, factorial analysis and principle component analysis. These studies assumed the linear relationship between plant characters and crop yield. However, these models have not been successful in capturing the nonlinear relationship between crop yield and its factors. Traditional methods for estimating crop yields often rely on field surveys and expert assessments, which can be time-consuming, labor-intensive, and prone to error. Machine learning (ML) has emerged as a powerful tool to overcome these challenges by analyzing vast datasets to make accurate predictions. By leveraging diverse data sources such as weather patterns, soil conditions, and remote sensing imagery, ML techniques can uncover complex patterns and relationships that were previously inaccessible.

Some commonly used ML models in crop yield estimation are as follows-

Random Forest: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. The final prediction is then made by averaging the predictions of individual trees. Random Forest models are adept at handling large datasets and capturing complex relationships between various factors influencing crop yield, such as weather conditions, soil quality, and historical crop data. (Parsad et al., 2021; Dhillon et al., 2023).

Support Vector Machines (SVM): SVM is a supervised learning model that analyzes data for classification and regression analysis. It works by finding the hyperplane that best divides a dataset into classes or predicts a continuous outcome. SVM can be employed for crop yield estimation by mapping various input features like temperature, precipitation, and nutrient levels to predict crop output. (Das et al., 2023; Dang et al., 2021).

Neural Networks: Neural networks, especially deep learning models, consist of layers of interconnected nodes that mimic the structure of the human brain. They can learn intricate patterns and relationships in data. Neural networks are valuable for crop yield estimation due to their ability to handle non-linear relationships and process vast amounts of data. They excel when there are complex interactions between different variables. (Das et al., 2023; Bharti et al., 2023).

Decision Trees: Decision trees break down a dataset into smaller subsets based on various conditions. They are used for both classification and regression tasks.Decision trees can be applied to crop yield estimation by considering factors like weather patterns, soil health, and crop type, providing a clear and interpretable model for farmers. (Bhatnagar & Gohain, 2020).

Linear Regression: Linear regression models establish a linear relationship between input features and the target variable. They are particularly useful for understanding the impact of individual factors on crop yield. Linear regression can be applied to predict crop yield by analyzing the linear correlation between variables such as temperature, precipitation, and the historical yield of a specific crop. (Das et al., 2023; Bharti et al., 2023).

Gradient Boosting: Gradient Boosting is an ensemble learning technique that builds a series of weak learners (usually decision trees) sequentially, with each one correcting the errors of its predecessor.Gradient Boosting models are effective for crop yield estimation when there is a need to capture the influence of multiple factors, refining predictions with each iteration.(Arumugam et al., 2021).

K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm that classifies a data point based on how its neighbors are classified. KNN can be utilized for crop yield estimation by considering the similarity between the current growing conditions and those observed in historical data. (Medar, & Rajpurohit, 2014).

Deep learning models: Deep learning models, a subset of machine learning, have shown remarkable capabilities in handling complex patterns and relationships within data. When applied to crop yield estimation, deep learning models can leverage their ability to learn hierarchical representations, capturing intricate dependencies among various factors influencing crop production. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Autoencoders, Generative Adversarial Networks (GANs) are few models that can be used for crop yield estimation (Khaki & Wang, 2019; Huber et al., 2022).

Applications of Machine Learning in Crop Yield Estimation

1. **Precision Agriculture:** Machine learning is revolutionizing precision agriculture by enabling data-driven decision-making and automation across various farming practices. The early identification of crop diseases and pests is made possible by image recognition and predictive modeling, which improves interventions and minimizes losses. Machine learning algorithms can effectively predict crop yields through analyzing a variety of data sources, such as soil conditions and weather patterns, which makes resource allocation more efficient. In automated irrigation systems, real-time data analysis ensures precise water usage, preventing overwatering and conserving resources. Furthermore, machine learning facilitates targeted weed and pest management, reducing the need for widespread chemical applications and minimizing environmental impact. Finally, soil data analysis enables detailed soil mapping and variable rate application of fertilizers, ensuring optimal nutrient distribution for enhanced crop growth.
2. **Disaster Management:**
    - Estimating the impact of natural disasters, such as droughts and floods, on crop yields.
3. **Market Forecasting:**
    - Helping stakeholders predict supply trends and adjust pricing strategies.
4. **Climate Change Studies:**
    - Analyzing the impact of changing weather patterns on agricultural productivity.

Challenges in Crop Yield Prediction using ML models

1. **Data Availability and Quality:**
    - Limited access to high-quality, annotated datasets can hinder model training.
2. **Model Interpretability:**
    - Complex models, especially deep learning algorithms, often operate as "black boxes," making it hard to interpret their decisions.
3. **Scalability:**
    - Adapting models to diverse regions with varying climates, soils and crops can be challenging.
4. **Integration with Existing Systems:**
    - Incorporating ML models into traditional agricultural workflows requires significant effort.

Future Directions

1. **Integration of IoT and Edge Computing:**
    - Real-time data from IoT devices like soil sensors and weather stations can enhance model predictions.

2. **Federated Learning:**
    - Collaborative ML approaches allow multiple stakeholders to train models without sharing sensitive data.
3. **Explainable AI (XAI):**
    - Developing interpretable models to gain insights into the factors affecting crop yields.
4. **Open Data Initiatives:**
    - Encouraging data sharing across institutions to build comprehensive datasets for ML training.

Conclusion

Crop yield estimation is vital for sustainable and profitable agriculture. Machine learning offers tremendous potential to transform crop yield estimation, making it more accurate, efficient, and scalable. By leveraging diverse datasets and advanced algorithms, stakeholders can address challenges in agriculture more effectively. However, the success of these technologies depends on overcoming barriers related to data availability, model interpretability and integration into existing systems. With continued advancements, machine learning will play a pivotal role in ensuring global food security and sustainable agricultural practices. Further, Integration of remote sensing, artificial intelligence and machine learning algorithms into agriculture will enhance prediction accuracy and revolutionize farming practices and support global food security.

Reference:

Arumugam, P., Chemura, A., Schauberger, B., & Gornott, C. (2021). Remote sensing based yield estimation of Rice (Oryza sativa L.) using gradient boosted regression in India. *Remote Sensing,* **13(12)**, 2379.

Bharti, Das, P., Banerjee, R., Ahmad, T., Devi, S., & Verma, G. (2023). Artificial Neural Network Based Apple Yield Prediction Using Morphological Characters. *Horticulturae*, **9(4)**, 436.

Bhatnagar, R., & Gohain, G. B. (2020). Crop yield estimation using decision trees and random forest machine learning algorithms on data from terra (EOS AM-1) & Aqua (EOS PM-1) satellite data. *Machine Learning and Data Mining in Aerospace Technology*, 107-124.

Dang, C., Liu, Y., Yue, H., Qian, J., & Zhu, R. (2021). Autumn crop yield prediction using data-driven approaches:-support vector machines, random forest, and deep neural network methods. *Canadian Journal of Remote Sensing*, **47(2)**, 162-181.

Das, P., Jha, G. K., Lama, A., & Parsad, R. (2023). Crop Yield Prediction Using Hybrid Machine Learning Approach: A Case Study of Lentil (Lens culinaris Medik.). *Agriculture*, **13(3)**, 596.

Dhillon, M. S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I., & Ullmann, T. (2023). Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. *Frontiers in Remote Sensing*, **3**, 1010978.

Huber, F., Yushchenko, A., Stratmann, B., & Steinhage, V. (2022). Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. *Computers and Electronics in Agriculture*, **202**, 107346.

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, **10**, 621.

Medar, R. A., & Rajpurohit, V. S. (2014). A survey on data mining techniques for crop yield prediction. *International Journal of Advance Research in Computer Science and Management Studies*, **2(9)**, 59-64.

Prasad, N. R., Patel, N. R., & Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*, **29**, 195-206.

OVERVIEW OF DESIGN OF EXPERIMENTS

Eldho Varghese*, Rajender Parsad, Seema Jaggi and Cini Varghese

*ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012, *ICAR-CMFRI, Kochi*

1. Introduction

In this chapter, three basic designs viz., Completely randomized design (CRD), Randomized Complete Block Design (RCBD) and Latin Square Design (LSD) are explained in detail.

2. Completely Randomized Design

Designs are usually characterized by the nature of grouping of experimental units and the procedure of random allocation of treatments to the experimental units. In a completely randomized design the units are taken in a single group. As far as possible the units forming the group are homogeneous. This is a design in which only randomization and replication are used. There is no use of local control here.

Let there be v treatments in an experiment and n homogeneous experimental units. Let the i^{th} treatment be replicated r_i times $(i = 1,2,\ldots, v)$ such that $\sum_{i=1}^{v} r_i = n$. The treatments are allotted at random to the units.

Normally the number of replications for different treatments should be equal as it ensures equal precision of estimates of the treatment effects. The actual number of replications is, however, determined by the availability of experimental resources and the requirement of precision and sensitivity of comparisons. If the experimental material for some treatments is available in limited quantities, the numbers of their replication are reduced. If the estimates of certain treatment effects are required with more precision, the numbers of their replication are increased.

Randomization

There are several methods of random allocation of treatments to the experimental units. The v treatments are first numbered in any order from 1 to v. The n experimental units are also numbered suitably. One of the methods uses the random number tables. Any page of a random number table is taken. If v is a one-digit number, then the table is consulted digit by digit. If v is a two-digit number, then two-digit random numbers are consulted. All numbers greater than v including zero are ignored.

Let the first number chosen be n_1; then the treatment numbered n_1 is allotted to the first unit. If the second number is n_2 which may or may not be equal to n_1 then the treatment numbered n_2 is allotted to the second unit. This procedure is continued. When the i^{th}

treatment number has occurred r_i times, $(i = 1,2,...,v)$ this treatment is ignored subsequently. This process terminates when all the units are exhausted.

One drawback of the above procedure is that sometimes a very large number of random numbers may have to be ignored because they are greater than *v*. It may even happen that the random number table is exhausted before the allocation is complete. To avoid this difficulty the following procedure is adopted. We have described the procedure by taking *v* to be a two-digit number.

Let *P* be the highest two-digit number divisible by *v*. Then all numbers greater than *P* and zero are ignored. If a selected random number is less than *v*, then it is used as such. If it is greater than or equal to *v*, then it is divided by *v* and the remainder is taken to the random number. When a number is completely divisible by *v*, then the random number is *v*. If *v* is an *n*-digit number, then *P* is taken to be the highest *n*-digit number divisible by *v*. The rest of the procedure is the same as above.

Analysis

This design provides a one-way classified data according to levels of a single factor. For its analysis the following model is taken:

$$y_{ij} = \mu + t_i + e_{ij}\,, \qquad i = 1,\cdots,v;\ j = 1,\cdots r_i,$$

where y_{ij} is the random variable corresponding to the observation y_{ij} obtained from the j^{th} replicate of the i^{th} treatment, μ is the general mean, t_i is the fixed effect of the i^{th} treatment and e_{ij} is the error component which is a random variable assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Let $\sum_j y_{ij} = T_i \quad (i = 1,2,...,v)$ be the total of observations from i^{th} treatment. Let further $\sum_i T_i = G$. Correction factor $(C.F.) \ = G^2/n.$

Sum of squares due to treatments $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$

Total sum of squares $= \sum_{i=1}^{v}\sum_{j=1}^{r_i} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Treatments	$v - 1$	$SST = \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$n - v$	$SSE = by\ subtraction$	$MSE = SSE / (n - v)$	
Total	$n - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test where F is the ratio *MST / MSE* with *(v - 1)* and *(n - v)* degrees of freedom.

3. Randomized Complete Block Design

It has been seen that when the experimental units are homogeneous then a CRD should be adopted. In any experiment, however, besides treatments the experimental material is a major source of variability in the data. When experiments require a large number of experimental units, the experimental units may not be homogeneous, and in such situations CRD can not be recommended. When the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or blocks). The principle of first forming homogeneous groups of the experimental units and then allotting at random each treatment once in each group is known as local control. This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks, is smaller because of homogeneous blocks. This type of allocation makes it possible to eliminate from error variance a portion of variation attributable to block differences. If, however, variation between the blocks is not significantly large, this type of grouping of the units does not lead to any advantage; rather some degrees of freedom of the error variance is lost without any consequent decrease in the error variance. In such situations it is not desirable to adopt randomized complete block designs in preference to completely randomized designs.

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group then such an arrangement is called a ***randomized complete block design.***

Suppose the experimenter wants to study v treatments. Each of the treatments is replicated r times (the number of blocks) in the design. The total number of experimental units is,

therefore, *vr*. These units are arranged into *r* groups of size *v* each. The error control measure in this design consists of making the units in each of these groups homogeneous.

The number of blocks in the design is the same as the number of replications. The *v* treatments are allotted at random to the *v* plots in each block. This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block designs. The availability of resources and considerations of cost and precision determine actual number of replications in the design.

Analysis

The data collected from experiments with randomized block designs form a two-way classification, that is, classified according to the levels of two factors, *viz.,* blocks and treatments. There are *vr* cells in the two-way table with one observation in each cell. The data are orthogonal and therefore the design is called an *orthogonal design.* We take the following model:

$$y_{ij} = \mu + t_i + b_j + e_{ij}\ , \qquad \begin{pmatrix} i = 1,2,...,v; \\ j = 1,2,...,r \end{pmatrix},$$

where y_{ij} denotes the observation from i^{th} treatment in j^{th} block. The fixed effects μ, t_i, b_j denote respectively the general mean, effect of the i^{th} treatment and effect of the j^{th} block. The random variable e_{ij} is the error component associated with y_{ij}. These are assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Following the method of analysis of variance for finding sums of squares due to blocks, treatments and error for the two-way classification, the different sums of squares are obtained as follows: Let $\sum_j y_{ij} = T_i \ \ (i = 1,2,...,v)$ = total of observations from i^{th} treatment and $\sum_j y_{ij} = B_j \qquad j = 1,\cdots,r$ = total of observations from j^{th} block. These are the marginal totals of the two-way data table. Let further, $\sum_i T_i = \sum_j B_j = G.$

Correction factor ($C.F.$) = G^2/rv, Sum of squares due to treatments $= \sum_i \frac{T_i^2}{r} - C.F.$,

Sum of squares due to blocks $= \sum_j \frac{B_j^2}{v} - C.F.$, Total sum of squares $= \sum_{ij} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Blocks	*r - 1*	$SSB = \sum_j \frac{B_j^2}{v} - C.F.$	*MSB = SSB / (r - 1)*	*MSB/MSE*
Treatments	*v - 1*	$SST = \sum_i \frac{T_i^2}{r} - C.F.$	*MST = SST / (v - 1)*	*MST/MSE*
Error	*(r - 1)(v - 1)*	*SSE = by subtraction*	*MSE = SSE / (v - 1)(r - 1)*	
Total	*vr - 1*	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test, where F is the ratio *MST / MSE* with *(v - 1)* and *(v - 1)(r - 1)* degrees of freedom. We may then be interested to either compare the treatments in pairs or evaluate special contrasts depending upon the objectives of the experiment. This is done as follows:

The critical difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $C.D. = t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE/r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's *t* at the level of significance α and degree of freedom *(v - 1)(r - 1)*. If the difference of any two-treatment means is greater than the C.D. value, the corresponding treatment effects are significantly different.

4. Latin Square Design

Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. These designs require that the number of replications equal the number of ***treatments*** or ***varieties***.

Definition 1. A Latin square arrangement is an arrangement of *v* symbols in v^2 cells arranged in *v* rows and *v* columns, such that every symbol occurs precisely once in each row and precisely once in each column. The term *v* is known as the **order** of the Latin square.

If the symbols are taken as *A, B, C, D,* a Latin square arrangement of order 4 is as follows:

A	*B*	*C*	*D*
B	*C*	*D*	*A*
C	*D*	*A*	*B*
D	*A*	*B*	*C*

A Latin square is said to be in the ***standard form*** if the symbols in the first row and first column are in natural order, and it is said to be in the ***semi-standard form*** if the symbols of the first row are in natural order. Some authors denote both of these concepts by the term ***standard form***. However, there is a need to distinguish between these two concepts. The standard form is used for randomizing the Latin-square designs, and the semi-standard form is needed for studying the properties of the orthogonal Latin squares.

Definition 2. If in two Latin squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, the two Latin squares are said to be **orthogonal**. If the symbols of one Latin square are denoted by Latin letters and the symbols of the other are denoted by Greek letters, the pair of orthogonal Latin squares is also called a **graeco-latin square**.

Definition 3. If in a set of Latin squares every pair is orthogonal, the set is called a set of **mutually orthogonal latin squares (MOLS)**. It is also called a **hypergraeco latin square.**

The following is an example of graeco latin square:

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

α	γ	δ	β
β	δ	γ	α
γ	α	β	δ
δ	β	α	γ

$A\alpha$	$B\gamma$	$C\delta$	$D\beta$
$B\beta$	$A\delta$	$D\gamma$	$C\alpha$
$C\gamma$	$D\alpha$	$A\beta$	$B\delta$
$D\delta$	$C\beta$	$B\alpha$	$A\gamma$

We can verify that in the above arrangement every pair of ordered Latin and Greek symbols occurs exactly once, and hence the two latin squares under consideration constitute a graecolatin square.

It is well known that the maximum number of MOLS possible of order v is $v - 1$. A set of $v - 1$ MOLS is known as a complete set of MOLS. Complete sets of MOLS of order v exist when v is a ***prime or prime power.***

Randomization

According to the definition of a Latin square design, treatments can be allocated to the v^2 experimental units (may be animal or plots) in a number of ways. There are, therefore, a number of Latin squares of a given order. The purpose of randomization is to select one of these squares at random. The following is one of the methods of random selection of Latin squares.

Let a $v \times v$ Latin square arrangement be first written by denoting treatments by Latin letters *A, B, C*, *etc.* or by numbers *1, 2, 3, etc.* Such arrangements are readily available in the ***Tables for Statisticians and Biometricians*** (Fisher and Yates, 1974). One of these squares of any order can be written systematically as shown below for a 5×5 Latin square:

A	*B*	*C*	*D*	*E*
B	*C*	*D*	*E*	*A*
C	*D*	*E*	*A*	*B*
D	*E*	*A*	*B*	*C*
E	*A*	*B*	*C*	*D*

For the purpose of randomization rows and columns of the Latin square are rearranged randomly. There is no randomization possible within the rows and/or columns. For example, the following is a row randomized square of the above *5×5* Latin square;

A	*B*	*C*	*D*	*E*
B	*C*	*D*	*E*	*A*
E	*A*	*B*	*C*	*D*
D	*E*	*A*	*B*	*C*
C	*D*	*E*	*A*	*B*

Next, the columns of the above row randomized square have been rearranged randomly to give the following random square:

E	*B*	*C*	*A*	*D*
A	*C*	*D*	*B*	*E*
D	*A*	*B*	*E*	*C*
C	*E*	*A*	*D*	*B*
B	*D*	*E*	*C*	*A*

As a result of row and column randomization, but not the randomization of the individual units, the whole arrangement remains a Latin square.

Analysis of Latin Square Designs

In Latin square designs there are three factors. These are the factors *P, Q,* and treatments. The data collected from this design are, therefore, analyzed as a three-way classified data. Actually, there should have been v^3 observations as there are three factors each at v levels. But because of the particular allocation of treatments to the cells, there is only one observation per cell instead of v in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained. Accordingly, we take the model

$$Y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} denotes the observation in the i^{th} row, j^{th} column and under the s^{th} treatment; $\mu, r_i, c_j, t_s (i, j, s = 1,2,...,v)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component, assumed to be independently and normally distributed with zero mean and a constant variance, σ^2.

The analysis is conducted by following a similar procedure as described for the analysis of two-way classified data. The different sums of squares are obtained as below: Let the data be arranged first in a *row* $\times$ *column* table such that y_{ij} denotes the observation of (*i, j)*th cell of table.

Let $R_i = \sum_j y_{ij} = i^{th}\ row\ total\ (i = 1,2,...,v)$, $C_j = \sum_i y_{ij} = j^{th}\ column\ total\ (j = 1,2,...,v)$, $T_s =$ sum of those observations which come from s^{th} treatment *(s= 1,2,...,v)*, $G = \sum_i R_i = grand\ total$. Correction factor, $C.F. = \frac{G^2}{v^2}$. Treatment sum of squares = $\sum_s \frac{T_s^2}{v} - C.F.$, Row sum of squares = $\sum_i \frac{R_i^2}{v} - C.F.$, Column sum of squares = $\sum_j \frac{C_j^2}{v} - C.F.$

Analysis of Variance of $v \times v$ Latin Square Design

Sources of Variation	D.F.	S.S.	M.S.	F
Rows	*v -1*	$\sum_i \frac{R_i^2}{v} - C.F.$		
Columns	*v - 1*	$\sum_j \frac{C_j^2}{v} - C.F.$		
Treatments	*v - 1*	$\sum_s \frac{T_s^2}{v} - C.F.$	s_t^2	s_t^2 / s_e^2
Error	*(v - 1)(v - 2)*	By subtraction	s_e^2	
Total	v^2-1	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis of equal treatment effects is tested by *F*-test, where *F* is the ratio of treatment mean squares to error mean squares. If *F* is not significant, treatment effects do not differ significantly among themselves. If *F* is significant, further studies to test the significance of

any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

Contrasts Analysis

The main technique adopted for the analysis and interpretation of the data collected from an experiment is the analysis of variance technique that essentially consists of partitioning the total variation in an experiment into components ascribable to different sources of variation due to the controlled factors and error. Analysis of variance clearly indicates a difference among the treatment means. The objective of an experiment is often much more specific than merely determining whether or not all of the treatments give rise to similar responses. For examples, a chemical experiment might be run primarily to determine whether or not the yield of the chemical process increases as the amount of the catalyst is increased. A medical experimenter might be concerned with the efficacy of each of several new drugs as compared to a standard drug. A nutrition experiment may be run to compare high fiber diets with low fiber diets. A plant breeder may be interested in comparing exotic collections with indigenous cultivars. An agronomist may be interested in comparing the effects of biofertilisers and chemical fertilisers. A water technologist may be interested in studying the effect of nitrogen with Farm Yard Manure over the nitrogen levels without farm yard manure in presence of irrigation.

2.1 Contrasts

Let $y_1, y_2, \ldots, y_n$ denote n observations or any other quantities. The linear function $C = \sum_{i=1}^{n} l_i y_i$, where l_i's are given number such that $\sum_{i=1}^{n} l_i = 0$, is called a *contrast* of $y_i's$.

Let $y_1, y_2, \ldots, y_n$ be independent random variables with a common mean μ and variance σ^2. The expected value of the random variable C is zero and its variance is $\sigma^2 \sum_{i-1}^{n} l_i^2$. In what follows we shall not distinguish between a contrast and its corresponding random variable.

Sum of squares (s.s.) of contrasts. The sum of squares due to the contrast C is defined as $C^2 / \sigma^{-2} Var(C) = C^2 / \left(\sum_{i=1}^{n} l_i^2 \right)$. Here σ^2 is unknown and is replaced by its unbiased estimate, *i.e. mean square error*. It is known that this square has a $\sigma^2 \chi^2$ distribution with one degree of freedom when the $y_i's$ are normally distributed. Thus the sum of squares due to two or more contrasts has also a $\sigma^2 \chi^2$ distribution if the contrasts are independent. Multiplication of any contrast by a constant does not change the contrast. The sum of squares due to a contrast as defined above is not evidently changed by such multiplication.

Orthogonal contrasts. Two contrasts, $C_1 = \sum_{i=1}^{n} l_i y_i$ and $C_2 = \sum_{i=1}^{n} l_i y_i$ are said to be orthogonal if and only if $\sum_{i=1}^{n} l_i m_i = 0$. This condition ensures that the covariance between C_1 and C_2 is zero.

When there are more than two contrasts, they are said to be mutually orthogonal if they are orthogonal pair wise. For example, with four observations y_1, y_2, y_3, y_4, we may write the following three mutually orthogonal contrasts:

(i) $y_1 + y_2 - y_3 - y_4$

(ii) $y_1 - y_2 - y_3 + y_4$

(iii) $y_1 - y_2 + y_3 - y_4$

The sum of squares due to a set of mutually orthogonal contrasts has a $\sigma^2 \chi^2$ distribution with as many degrees of freedom as the number of contrasts in the set.

Multiple Comparison Procedures

Duncan's Multiple Range Test

A widely used procedure for comparing all pairs of means is the multiple range test developed by Duncan (1955). The application of Duncan's multiple range test (*DMRT*) is similar to that of *lsd* test. *DMRT* involves the computation of numerical boundaries that allow for the classification of the difference between any two treatment means as significant or non-significant. *DMRT* requires computation of a series of values each corresponding to a specific set of pair comparisons unlike a single value for all pairwise comparisons in case of *lsd.* It primarily depends on the standard error of the mean difference as in case of *lsd.* This can easily be worked out using the estimate of variance of an estimated elementary treatment contrast through the design.

For application of the *DMRT* rank all the treatment means in decreasing or increasing order based on the preference of the character under study.

Tukey Method for All Pairwise Comparisons

Tukey (1953) proposed a method for making all possible pairwise treatment comparisons. The test compares the difference between each pair of treatment effects with appropriate adjustment for multiple testing. This test is also known as Tukey's honestly significant difference test or Tukey's HSD. It may be mentioned here that Tukey's method is the best for

all pairwise treatment comparisons. It can be used for completely randomized designs, randomized complete block designs and balanced incomplete block designs. It is believed to be applicable (conservative, true α level lower than stated) for other incomplete block designs as well, but this has not yet been proven. It can be extended to include all contrasts but Scheffe's method is generally better for these types of contrasts.

Dunnett Method for Treatment-Versus-Control Comparisons

Dunnett (1955) developed a method of multiple comparisons for obtaining a set of simultaneous confidence intervals for preplanned treatment-versus-control contrasts $t_i - t_1 (i = 2,..., v)$ where level 1 corresponds to the control treatment. The intervals are shorter than those given by the Scheffe, Tukey and Bonferroni methods, but the method should not be used for any other type of contrasts. For details on this method, a reference may be made to Dunnett (1955, 1964) and Hochberg and Tamhane (1987). In general this procedure is, therefore, best for all treatment-versus-control comparisons. It can be used for completely randomized designs, randomized complete block designs. It can also be used for balanced incomplete block designs but not in other incomplete block designs without modifications to the corresponding multivariate t-distribution tables given in Hochberg and Tamhane (1987).

SAS Code

Analysis of data obtained from experiment conducted under CRD setup

```
data crd;
input trt yld;
cards;
1	850.5
1	453.6
1	878.85
1	623.7
1	510.3
1	765.45
1	680.4
1	595.35
1	538.65
1	850.5
1	850.5
1	793.8
1	1020.6
1	708.75
1	652.05
1	623.7
1	396.9
1	822.15
1	680.4
1	652.05
1	538.65
```

1	850.5
1	680.4
1	.
1	.
2	510.3
2	963.9
2	652.05
2	1020.6
2	878.85
2	567
2	680.4
2	538.65
2	567
2	510.3
2	425.25
2	567
2	623.7
2	538.65
2	737.1
2	453.6
2	481.95
2	368.55
2	567
2	595.35
2	567
2	595.35
2	.
2	.
2	.
3	992.25
3	850.5
3	1474.2
3	510.3
3	850.5
3	793.8
3	453.6
3	935.55
3	1190.7
3	481.95
3	623.7
3	878.85
3	1077.3
3	850.5
3	680.4
3	737.1
3	737.1

```
3       708.75
3       708.75
3       652.05
3       567
3       453.6
3       652.05
3       567
3       .
;

proc glm;
class trt;
model yld = trt;
means  trt;
means trt/lsd;
run;
```

Analysis of data obtained from experiment conducted under RBD setup

```
data rbd;
input blk  trt yld;
cards;
1       1       6.9
1       2       6.48
1       3       6.52
1       4       6.9
1       5       6
1       6       7.9
2       1       4.6
2       2       5.57
2       3       7.6
2       4       6.65
2       5       6.18
2       6       7.57
3       1       4.4
3       2       4.28
3       3       5.3
3       4       6.75
3       5       5.5
3       6       6.8
4       1       4.81
4       2       4.45
4       3       5.3
4       4       7.75
4       5       5.5
4       6       6.62
```

```
;
proc glm;
class  blk trt;
model yld = blk trt;
means  trt;
means trt/tukey;
run;
```

Analysis of data obtained from experiment conducted under LSD setup

```
data lsd;
input row column trt yld;
cards;
1	1	3	3.1
1	2	6	5.95
1	3	1	1.75
1	4	5	6.4
1	5	2	3.85
1	6	4	5.3
2	1	2	4.8
2	2	1	2.7
2	3	3	3.3
2	4	6	5.95
2	5	4	3.7
2	6	5	5.4
3	1	1	3
3	2	2	2.95
3	3	5	6.7
3	4	4	5.95
3	5	6	7.75
3	6	3	7.1
4	1	5	6.4
4	2	4	5.8
4	3	2	3.8
4	4	3	6.55
4	5	1	4.8
4	6	6	9.4
5	1	6	5.2
5	2	3	4.85
5	3	4	6.6
5	4	2	4.6
5	5	5	7
5	6	1	5
6	1	4	4.25
6	2	5	6.65
6	3	6	9.3
6	4	1	4.95
```

```
6       5       3       9.3
6       6       2       8.4
;
proc glm;
class  blk trt;
model yld = blk trt;
means  trt;
means trt/tukey;
run;
```

Analysis of data obtained from experiment conducted under CRD setup with a control

```
data;
input trt       rep     plant_height_30       plant_height_60       plant_height_90
      plant_height_120;
/*trt 1 is the control*/
cards;
1       1       19.67   112.67 153.33 178.33
2       1       22.00   126.67 180.33 198.33
3       1       22.67   118.33 162.33 198.33
4       1       22.67   123.33 156.67 203.33
5       1       22.67   113.33 183.00 193.33
6       1       25.33   118.33 182.00 204.33
7       1       25.67   136.67 167.67 186.67
8       1       24.00   127.67 174.00 181.67
9       1       30.33   126.00 175.67 188.33
10      1       32.67   140.00 146.67 187.33
1       2       21.67   102.67 180.67 182.33
2       2       20.67   106.67 171.00 176.00
3       2       21.67   120.00 167.00 191.33
4       2       22.00   130.00 190.00 196.33
5       2       23.67   111.67 200.67 191.67
6       2       23.33   123.33 175.67 191.67
7       2       24.33   114.33 182.33 188.33
8       2       24.00   123.33 178.33 190.00
9       2       26.33   129.00 190.33 200.00
10      2       22.33   115.00 163.33 191.00
1       3       19.33   96.67   148.00 163.33
2       3       19.33   110.00 171.33 164.67
3       3       20.00   120.00 193.00 165.67
4       3       20.67   113.33 171.00 181.67
5       3       21.33   123.33 156.67 181.33
6       3       23.00   116.67 180.33 178.00
7       3       27.33   133.33 169.33 181.67
8       3       21.00   113.33 184.00 183.33
9       3       23.67   132.33 173.67 181.33
```

```
10      3       27.33   136.67 145.33 194.33;
ods rtf file="anova.rtf" startpage=no;
proc glm;
class rep trt;
model plant_height_30       plant_height_60       plant_height_90
      plant_height_120= rep trt/SS3;
*lsmeans trt/pdiff;
Contrast 'CONTROL vs REST' trt -9 1 1 1 1 1 1 1 1 1;
run;
ods rtf close;
```

REFERENCES

Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference*, **1**, 1-25.

Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer Text in Statistics, New York.

Fisher, R.A. and Yates, F. (1963). *Statistical Tables For Biological, Agricultural and Medical Research*. Longman Group Ltd., England.

Parsad, Rajender and Gupta, V.K. Basic Experimental Designs. E book chapter available at http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%20files/Manual%20III/2-Basic%20Experiments.pdf

OVERVIEW OF TIME SERIES ANALYSIS AND ITS APPLICATIONS IN AGRICULTURE

Achal Lama and K N Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Time series analysis deals with observations that are collected over time in a definite order. As the ordering of the data is important time series observations are usually dependent. Time series has diverse applications, depending on which observations may be collected hourly, daily, weekly, monthly, or yearly, and so on. The notation such as $\{X_t\}$ or $\{Y_t\}$ (t =1,…,T) is used to denote a time series of length T.

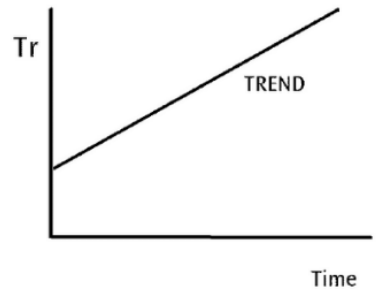
Time series can be viewed as a realization from a stochastic process and efforts are made to understand the probability law that governs the observed time series. So that we can understand the underlying dynamics, forecast future events, and control future events through suitable interventions. As the time series contain finite number of observations, there can be infinite number of stochastic processes that can generate the same observed data. However, some of these processes are more plausible and admit better interpretation than others. As we have finite number of observations, it is practically impossible to understand the underlying process without imposing some constraints. The best possible way is to confine the probability law or the data generating process to a specified family and then to select a member in that family that is most plausible. The former is called modelling and the latter is called estimation, or more generally statistical inference. When the form of the probability laws in a family is specified except for some finite-dimensional defining parameters, such a model is referred to as a parametric model. When the defining parameters lie in a subset of an infinite dimensional space or the form of probability laws is not completely specified, such a model is often called a nonparametric model. Time series analysis depends largely on proper statistical modelling. While selecting a model one should take into account its, interpretability, simplicity, and feasibility. The selected model should satisfactorily reflect the physical law that governs the data. In time series analysis a simple model is usually preferable for explaining the data generating process. The family of probability models should be reasonably large to include the underlying probability law that has generated the data but should not be so large that defining parameters can no longer be estimated with reasonably good accuracy. While choosing a probability model, we should first extract the salient features from the observed data and then chooses an appropriate model that possesses such features. After estimating parameters or functions in the model, it is to be verified whether the model fits the data reasonably well and looks for further improvement whenever possible. The use of different models is totally dependent upon the purpose of analysis. For example, a model that provides a good fitting and admits nice interpretation is not necessarily good for forecasting.

Components of time series

- Trend
- Seasonality
- Cyclic Component
- Random Component

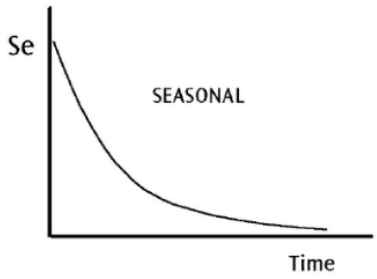
Trend

The trend component represents the long-term movement in a time series, indicating the general direction in which the data is moving over time. A trend can be upward, downward, or stable, depending on whether the values are increasing, decreasing, or remaining relatively constant. Trends can be influenced by economic growth, technological advancements, demographic changes, or other underlying factors that drive the overall pattern of the data.



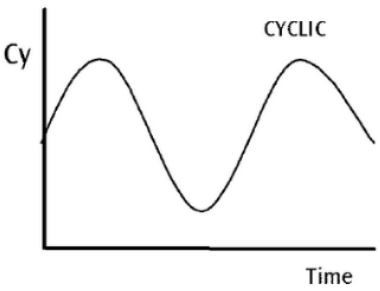
Seasonality

Seasonality refers to periodic fluctuations in a time series that occur at regular intervals, typically due to seasonal factors such as weather, holidays, or cultural events. These patterns repeat over a fixed period, such as daily, weekly, monthly, or annually. For example, retail sales tend to increase during festive seasons, and agricultural yields may follow a seasonal cycle based on planting and harvesting times.



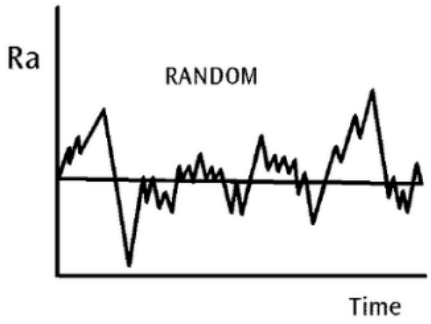
Cyclic Component

The cyclic component captures fluctuations in a time series that occur over longer, irregular intervals, often influenced by economic or business cycles. Unlike seasonality, cyclic patterns do not have a fixed period but instead vary in duration depending on external factors such as market conditions, policy changes, or economic crises. These cycles typically consist of phases such as expansion, peak, contraction, and trough.



Random Component (Residual or Random Variation)

The irregular component, also known as the residual or random variation, represents unpredictable fluctuations in a time series that cannot be attributed to trend, seasonality, or cyclic patterns. These variations are typically caused by unexpected events such as natural disasters, political instability, or sudden market shocks. Since these fluctuations are random, they are often treated as noise in time series analysis and are minimized through smoothing or modeling techniques.



Trend Estimation

Trend estimation is an essential aspect of time series analysis, used to identify the long-term direction of data while filtering out short-term fluctuations. One of the most commonly used techniques for trend estimation is the Moving Average (MA), which smooths data by averaging a fixed number of past observations.

Moving Averages

Simple Moving Average (SMA)

A Simple Moving Average (SMA) calculates the average of a specified number of past data points, shifting the window forward with each new observation. It is mathematically represented as:

$$SMA_t = \frac{X_t + X_{t-1} + \cdots + X_{t-(N-1)}}{N}$$

where X_t represents the observed values, and N is the window size. SMA smooths fluctuations and highlights the overall trend.

3-Period Moving Average (3MA)

A 3-period moving average (3MA) takes the average of the last three observations to estimate the trend. It is given by:

$$3MA_t = \frac{X_t + X_{t-1} + X_{t-2}}{3}$$

This method reacts quickly to changes in data but is more sensitive to short-term fluctuations compared to longer-period MAs.

5-Period Moving Average (5MA)

A 5-period moving average (5MA) smooths variations by averaging the last five observations at each time step. The formula is:

$$5MA_t = \frac{X_t + X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}}{5}$$

Compared to 3MA, the 5MA provides a more stable trend estimation by reducing the impact of short-term variations, making it suitable for detecting medium-term trends.

Exponential Smoothing

Exponential smoothing is a widely used technique for trend estimation in time series analysis, offering a flexible approach to smoothing fluctuations while giving more weight to recent observations. Unlike simple moving averages, exponential smoothing applies exponentially decreasing weights to past data points, making it highly effective for short-term forecasting. Different types of exponential smoothing methods include Simple Exponential Smoothing, Holt's Linear Trend Method, and Holt-Winters Method, each suitable for different types of time series patterns.

Simple Exponential Smoothing (SES)

Simple Exponential Smoothing (SES) is used for time series data without a trend or seasonality. It applies a smoothing factor α (where $0 < \alpha < 1$) to control the rate of decay for past observations. The formula for updating the smoothed value is:

$$S_t = \alpha X_t + (1 - \alpha) S_{t-1}$$

where:

- S_t is the smoothed value at time t,
- X_t is the actual observation at time t,
- $S_{t-1}$is the previous smoothed value,
- α is the smoothing parameter.

A higher α gives more weight to recent observations, making the model more responsive to changes, while a lower α results in smoother trends.

Holt's Linear Trend Method

Holt's method extends simple exponential smoothing by incorporating a trend component, making it suitable for data with a linear trend. It uses two equations: one for the level (S_t) and another for the trend (T_t). The update equations are:

$$S_t = \alpha X_t + (1 - \alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta (S_t - S_{t-1}) + (1 - \beta) T_{t-1}$$
$$\hat{X}_{t+h} = S_t + hT_t$$

where:

- S_t is the estimated level at time t,
- T_t is the estimated trend at time t,
- $\hat{X}_{t+h}$ is the forecast for h periods ahead,
- α and β are smoothing parameters for level and trend, respectively.

Holt-Winters Method (Triple Exponential Smoothing)

Holt-Winters extends Holt's method by incorporating a seasonal component, making it suitable for time series data with both trend and seasonality. It consists of three equations for level, trend, and seasonality. The choice between additive and multiplicative

seasonality depends on whether seasonal effects remain constant or vary proportionally with the trend.

Additive Holt-Winters Model (for constant seasonal variation)

$$S_t = \alpha(X_t - C_{t-m}) + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$
$$C_t = \gamma(X_t - S_t) + (1-\gamma)C_{t-m}$$

Forecast Equation: $\hat{X}_{t+h} = S_t + hT_t + C_{t-m+h}$

where:

- S_t level at time t
- T_t is trend estimate at time t
- C_t is seasonal component at time t
- $X_t$is observed value at time t
- m is Length of the seasonal cycle
- α is smoothing parameter for level
- β is smoothing parameter for trend
- γ is smoothing parameter for seasonality
- h is forecast horizon

2. Linear Time Series Models

The most popular class of linear time series models consists of autoregressive moving average (ARMA) models, including purely autoregressive (AR) and purely moving-average (MA) models as special cases. ARMA models are frequently used to model linear dynamic structures, to depict linear relationships among lagged variables, and to serve as vehicles for linear forecasting. A particularly useful class of models contains the so-called autoregressive integrated moving average (ARIMA) models, which includes stationary ARMA - processes as a subclass. We have tried to briefly introduce these linear models in the subsequent sub-sections.

2.1 Autoregressive (AR) Model

A stochastic model that can be extremely useful in the representation of certain practically occurring series is the autoregressive model. In this model, current value of the process is expressed as a finite, linear aggregate of previous values of the process and a shock ε_t . Let us denote the values of a process at equally spaced time epochs $t, t-1, t-2, \ldots$ by $y_t, y_{t-1}, y_{t-2}, \ldots$ then y_t can be described as

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

If we define an autoregressive operator of order p by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

where B is the backshift operator such that $By_t = y_{t-1}$, autoregressive model can be written as $\varphi(B)\, y_t = \varepsilon_t$.

2.2 Moving Average (MA) Model

Another kind of model of great practical importance in the representation of observed

time-series is finite moving average process. MA (q) model is defined as

$$y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$$

If we define a moving average operator of order q by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

where B is the backshift operator such that By_t □ y_{t-1} , moving average model can be written as y_t □ □(B)□$_t$.

2.3 Autoregressive Moving Average (ARMA) Model

To achieve greater flexibility in fitting of actual time-series data, it is sometimes advantageous to include both autoregressive and moving average processes. This leads to mixed autoregressive-moving average model

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots \theta_q\varepsilon_{t-q}$$

or

$$\square(B)\, y_t \;\square\; \square(B)\square_t$$

and is written as ARMA(p, q). In practice, it is quite often adequate representation of actually occurring stationary time-series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than *2.*

2.4 Autoregressive Integrated Moving Average (ARIMA) Model

A generalization of ARMA models which incorporates a wide class of non-stationary time-series is obtained by introducing the differencing into the model. The simplest example of a non-stationary process which reduces to a stationary one after differencing is Random Walk. A process { y_t } is said to follow an Integrated ARMA model, denoted by ARIMA (p, d, q), if □^{d} y_t □ (1 □ B)^d □$_t$ is ARMA (p, q). The model is written as

$$\varphi(B)(1-B)^d\, y_t = \theta(B)\varepsilon_t$$

ε_t are assumed to be independently and identically distributed with a mean zero and a constant variance of σ^2 .

3. Non-linear models: ARCH and GARCH models

After the dominance of the ARIMA model for over two decades, the need of such model was felt which could predict with varying variance of the error term. The solution was provided by Engle (1982) when he developed ARCH model to estimate the mean and variance of the United Kingdom inflation. This model has few interesting characteristics; it models the conditional variance as the square of the function of the previous error term and assumes the unconditional variance to be constant. Along with the ARCH models can model heavy tail data which are common in financial market. Besides these, Bera and Higgins (1993) pointed out that ARCH models are easy and simple to handle, can take care of clustered errors, non-linearity and importantly takes care of changes in the econometrician's ability to forecast.

The ARCH (q) model for the series $\{\varepsilon_t\}$ is defined by specifying the conditional distribution of given the information available up to time $t-1$. Let denote this information. ARCH (q) model for the series is given by

$$\varepsilon_t|\psi_{t-1} \sim N(0,h_t)$$

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2$$

where, $a_0 > 0$, $a_i \geq 0$, for all i and $\sum_{i=1}^{q} a_i < 1$ are required to be satisfied to ensure non-negativity and finite unconditional variance of stationary $\{\varepsilon_t\}$ series. Bollerslev (1986) and Taylor (1986) proposed the Generalized ARCH (GARCH) model independently of each other, in which conditional variance is also a linear function of its own lags and has the following form

$$\varepsilon_t = \xi_t h_t^{1/2} \tag{1}$$

where $\xi_t \sim$ N (0,1). A sufficient condition for the conditional variance to be positive is

$$a_0 > 0,\ a_i \geq 0,\ i = 1,2,...,q.\ \ b_j \geq 0,\ \ j = 1,2,...,p$$

The GARCH (p, q) process is weakly stationary if and only if

$$. \sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$$

The conditional variance defined by (1) has the property that the unconditional autocorrelation function of ε_t^2 ; if it exists, can decay slowly. For the ARCH family, the decay rate is too rapid compared to what is typically observed in financial time-series, unless the maximum lag q is long. As (1) is a more parsimonious model of the conditional variance than a high-order ARCH model, most users prefer it to the simpler ARCH alternative. The most popular GARCH model in applications is the GARCH (*1,1*) model.

Step 1: Determine whether the time series is stationary.

The series being analysed must be stationary. A stationary time series has the property that its statistical properties such as the mean and variance are constant over time. The presence of stationarity in the data can be obtained by simply plotting the raw data or by plotting the autocorrelation and partial autocorrelation function. Statistical tests like Dickey- Fuller test, augmented Dickey-Fuller test, KPSS (Kwiatkowski, Phillips, Schmidt, and Shin) test, Philips-Perron test are also available to test the stationarity.

Step 2: Identify the model.

After the time-series is stationary we go for identifying the mean model for the series. This is done by fitting the simple ARIMA (Autoregressive integrated moving average) model. The ARIMA (p,d,q) is determined by the ACF (Autocorrelation function) and PACF (Partial autocorrelation function) values of the stationary series. The parameter p is determined by the ACF value and q by the PACF value and d refers to order of differencing done to the original series to make it stationary.

Step 3: Estimate the model parameters and diagnostic checking.

Once few tentative models are specified, estimation of the model parameters is straightforward. The parameters are estimated through maximum likelihood function such that an overall measure of errors is minimized or the likelihood function is maximized. This step is basically to check if the model assumptions about the errors are satisfied. This is achieved by performing portmanteau test. The test is utilized to see whether the model residuals are white noise. The null hypothesis tested is that the current set of residual is white

noise.

The Ljung-Box statistic is given by:

$$Q = n(n+2)\sum_{k=1}^{h}(n-k)^{-1}r_k^{\,2}$$

where, h is the maximum lag, n is the number of observations, k is the number of parameters in the model. If the data are white noise, the Ljung-Box Q statistics has a chi-square distribution with *(h-k)* degrees of freedom.

Step 4: Select the most suitable ARIMA model

The most suitable ARIMA model is selected using the smallest Akaike Information Criterion (AIC) or Schwarz-Bayesian Criterion (SBC). AIC is given by

$$AIC = (-2\log L + 2m)$$

where, $m = p+q$ and L is the likelihood function. SBC is also used as an alternative to AIC which is given by

$$SBC = \log\sigma^2 + (m\log n)/n$$

If the model is not adequate, a new tentative model should be identified, which is again followed by the parameter estimation and model verification. Diagnostic information may help suggest alternative model(s). The steps of model building process are typically repeated several times until a satisfactory mean model is finally selected. The final model can then be used for prediction purposes.

Step 5: Determination of residuals and heteroscedasticity test.

After finding the mean model now the residuals are to be determined. And we create a new variable called ‘rsquare’ by squaring the residuals. Then the ACF and PACF values of the ‘rsquare’ are determined and the lags in which these values are found to be significant are identified. The test for heteroscedasticity is done at identified significant lags. The test employed is the ARCH-LM test.

Step 6: Residuals and diagnostic checking.

The residuals obtained from the mean model used for fitting the different GARCH models were squared and stored in a new variable called ‘esquare’. As already mentioned previously, the diagnostic tests are employed to check whether the residuals are white noise or not.

Step 7: Estimation of parameters.

The parameters of the obtained model are estimated using method of maximum likelihood (MLE). And then forecasting is done using the selecting model.

5. Illustration

In this illustration Cotlook A index data is used and was collected from the commodity price bulletin, published by the United Nations Convention of Trade and Development (UNCTAD). The series contains 360 data pints, 346 data points are used for modelling and remaining 14 points for forecasting. At first the ARIMA model was applied to the data set and on unsatisfactory performance of the model, the GARCH model was used.

5.1 Fitting of the Cotlook A index

Various combinations of the ARIMA models were tried, among all, the AR (1) model had minimum AIC and BIC values. The AIC value for fitted GARCH model has been found to be minimum when the mean equation depends on two recent pasts only. Investigating the autocorrelation function (Acf) of squared residuals of AR (2) model, it is found that the Acf and Pacf are maximum at lag 3, which is 0.226 and 0.221 respectively. But if we go for AR (2)-ARCH (3) model, a large number of parameters are needed to be estimated. So, to get a parsimonious model, the AR (2)-GARCH (1, 1) model is selected.

The mean and conditional variance for fitted AR (2)-GARCH (1, 1) model is computed as follows:

$$y_t = 141.9264 \; -1.3905 \; y_{t-1} + 0.4538 \, y_{t-2} + \varepsilon_t$$

(3.94) (0.05) (0.05)

where

$$\varepsilon_t = h_t^{1/2} \xi_t \,,$$

and h_t satisfies the variance equation

$$h_t = 8.470 + 0.208 \; \varepsilon_{t-1}^2 + 0.215 \, h_{t-1}$$

(1.97) (0.09) (0.079)

The values within brackets denote corresponding standard errors of the estimates. The AIC value, for fitted GARCH model is 2288.88.

Table 1. Forecast of the Cotlook A index series

MONTH	ACTUAL VALUE	FORECAST ARIMA(1,1,0)	FORECAST AR(2)-GARCH(1,1)
Feb-11	469.98	408.34(8.30)	389.59(26.46)
Mar-11	506.34	416.47(15.56)	371.55(25.74)
Apr-11	477.56	421.40(22.35)	348.54(25.05)
May-11	364.91	424.53(28.55)	324.69(24.39)
Jun-11	317.75	426.66(34.17)	301.98(23.75)
Jul-11	268.96	428.23(39.29)	281.25(23.13)
Aug-11	251.55	429.49(43.97)	262.76(22.54)

Sep-11	257.63	430.57(48.29)	246.50(21.97)
Oct-11	243.85	431.55(52.30)	232.32(21.42)
Nov-11	230.78	432.48(56.05)	220.01(20.90)
Dec-11	210.43	433.37(59.58)	209.35(20.39)
Jan-12	222.91	434.25(54.45)	200.15(19.91)
Feb-12	222.12	435.12(57.13)	192.21(19.44)
Mar-12	219.36	435.99(59.68)	185.37(19.01)

Table 2. Forecast evaluation of the Cotlook A index series

MODEL	RMSE	RMAPE (%)
ARIMA(1,1,0)	44.03	60.72
AR(2)-GARCH(1,1)	15.38	9.36

6. R code for analysing a time series data

```
library(“tseries”)
library(“forecast”)
library(“fgarch”)
setwd("C:/Users/BISHAL/Desktop") # Setting of the work directory
data<-read.table("bishal.txt") # Importing data
datats<-ts(data,frequency=12,start=c(1982,4)) # Converting data set into time series
plot.ts(datats) # Plot of the data set
adf.test(datats) # Test for stationarity
diffdatats<-diff(datats,differences=1) # Differencing the series
datatsacf<-acf(datats,lag.max=12) # Obtaining the ACF plot
datapacf<-pacf(datats,lag.max=12) # Obtaining the PACF plot
auto.arima(diffdatats) # Finding the order of ARIMA model
datatsarima<-arima(diffdatats,order=c(1,0,1),include.mean=TRUE) # Fitting of ARIMA model
forearimadatats<-forecast.Arima(datatsarima,h=12) # Forecasting using ARIMA model
plot.forecast(forearimadatats) # Plot of the forecast
residualarima<-resid(datatsarima) # Obtaining residuals
```

```
archTest(residualarima,lag=12) # Test for heteroscedascity


# Fitting of AR-GARCH model

garchdatats<-garchFit(formula = ~ arma(2)+garch(1, 1), data = datats, cond.dist = c("norm"), include.mean = TRUE, include.delta = NULL, include.skew = NULL, include.shape = NULL, leverage = NULL, trace = TRUE,algorithm = c("nlminb"))

# Forecasting using AR-GARCH model

forecastgarch<-predict(garchdatats, n.ahead = 12, trace = FALSE, mse = c("uncond"), plot=FALSE, nx=NULL, crit_val=NULL, conf=NULL)

plot.ts(forecastgarch) # Plot of the forecast
```

References:

Bera, A. K., and Higgins, M. L. (1993), ARCH Models: Properties, Estimation and Testing, *Journal of Economic Surv*ey, **7**, 307-366.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, **31**, 307-327.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2007). Time-Series Analysis: Forecasting and Control. 3rd edition. *Pearson education*, India.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987-1008.

Fan, J. and Yao, Q. (2003). *Nonlinear time series:nonparametric and parametric methods*. Springer, U.S.A.

Taylor, S. J. (1986). Modeling financial time series. Wiley, New York.

CROP YIELD FORECASTING IN INDIA

K N Singh

ICAR-Indian Agricultural Statistics Research Institute New Delhi-110012, India

Abstract

Crop yield is determined by numerous input parameters, it is vital to find important variables and omit the other redundant ones which may decrease the accuracy of predictive models. The machine learning driven feature selection algorithms assist in selecting only those features that are relevant in the predictive algorithms. Instead of a complete set of features, feature subsets give better results for the same algorithm with less computational time. Feature selection has the potential to play an important role in the agriculture domain, with the crop yield depending on multiple factors. Crop yield prediction is a complex phenomenon and has many underlining nonlinear patterns. Such, datasets are difficult to deal with stringent assumptions of the statistical models. Hence, machine learning (ML) techniques which has very few prior assumptions and are data driven provides great deal of flexibility for modelling and forecasting the crop yield. Various researchers have applied different ML techniques for forecasting crop yield and have obtained satisfactory results. We have attempted to build crop yield forecasting model based on features (weather indices) selected by using two very popular machine learning algorithms, i.e., Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest Variable Importance (RFVarImp). Further, we have also applied Random Forest (RF) and Support Vector Regression (SVR) directly to the data sets and have made a comparative analysis among them using appropriate statistical measures.

1. Introduction

Reliable forecast of crop production before the harvest is important for advance planning, formulation and implementation policies dealing with food procurement, its distribution, pricing structure, import and export decisions, and storage and marketing of the agricultural commodities. Weather plays a very important role in crop growth and development. Therefore, model based on weather variables can provide reliable forecast. Weather variables used can be employed for crop production forecast by making appropriate models. Understanding the importance that important variables play for defining a model, we have utilized some important variable section techniques based on machine learning approaches. The least absolute shrinkage and selection operator (LASSO) and Random Forest (RF) are the machine learning approaches whereas classical approach such as stepwise regression methodology has also been applied. At ICAR-IASRI, model by Hendricks and Scholl was modified by expressing effects of weather on yield in successive weeks as a quadratic function of respective correlation coefficients between yield and weather instead of week no. Under this assumption model becomes

$$Y = A_0 + a_0 \sum X_w + a_1 \sum_w r_w X_w + a_2 \sum_{w=1}^{n} r_w^2 X_w + bT + e$$

$$Y = A_0 + \sum_{i=1}^{p}\sum_{j=0}^{2} a_{ij}Z_{ij} + \sum_{i\neq i'=1}^{p}\sum_{j=0}^{2} a_{ii'j}Z_{ii'j} + bT + e$$

where

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw}$$

$$Z_{ii'j} = \sum_{w=1}^{m} r_{ii'w}^{j} X_{iw} X_{i'w}$$

Z's are the weather indices obtained using weather variables

r_{iw} is correlation coefficient of yield with i^{th} weather variable in w^{th} period; $r_{ii'w}$ is correlation coefficient of yield with product of i^{th} and i'^{th} weather variables in w^{th} period; m is period of forecast; p is number of weather variables used and e is random error distributed as $N(0,\sigma^2)$.

2. Predictor variable(s) selection methods:

We have used the following three techniques for most significant variable(s) selection.

2.1 Stepwise regression

Stepwise regression is a modification of the forward selection that all predictor variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a non-significant variable is found, it is removed from the model.

2.2 Lasso (Least Absolute Shrinkage and Selection Operator)

Lasso is a regularization technique. It reduces the number of predictors in a regression model and identifies important predictors. It is able to perform variable selection by shrinking the estimated value of some of the regression coefficient which is less important to exactly equals to zero. Through adopting regularization technique, it minimizes the variance of the estimated regression coefficients and thus makes the estimators more stable.

The LASSO estimate is defined by the solution to the $l_1$optimization problem

$$minimize\left(\frac{\|\boldsymbol{Y}-\boldsymbol{X\beta}\|_2^2}{n}\right) subject\ to \sum_{j=1}^{k}\|\beta\|_1 < t$$

where t is the upper bound for the sum of the coefficients

This optimization problem is equivalent to the parameter estimation that follows:

$$\hat{\beta}(\lambda) = argmin(\|\boldsymbol{Y}-X\boldsymbol{\beta}\|_2^2 + \lambda\|\beta\|_1)$$

where $\|Y-X\beta\|_2^2 = \sum_{i=0}^{n}\boldsymbol{Y}_i - (X\boldsymbol{\beta})_i$, $\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$ $and\ \lambda \geq 0$ is the parameter that controls the strength of the penalty. Larger the value of λ, greater the amount of shrinkage. The relation between λ and the upper bound t is a reverse relationship. As t becomes infinity, the problem becomes an ordinary least squares and λ becomes 0. Vice-versa as t becomes 0, all coefficients shrink to 0 and λ goes to infinity. In this way the features with coefficient equal to zero are excluded from the model.

2.3 Random Forest

Random forests is a popular and very efficient algorithm, based on model aggregation ideas, for both classification and regression problems. It belongs to the family of ensemble methods of machine learning. The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the learning sample L and choosing randomly at each node a subset of explanatory variables X. Explanatory variables are ranked based on random forests score of importance. This technique is useful when the number of true variables is much less than p, and also in the presence of groups of highly correlated predictors. It is based on the principle of minimizing out-of-bag (OOB) error and on the quantification of the variable importance (*VI*) for the obtained forest. The OOB error is defined as

$$errOOB = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

For each tree *t* of the forest, the associated OOB_t sample (data not used in the bootstrap sample used to construct *t*). We have used 2000 trees with 50 runs to build the RF.

Steps to be followed for selecting important variables:

I. Rank the variables by sorting the *VI* in descending order

II. Eliminate the variables of small importance by comparing with the threshold values obtained from CART (Classification and Regression tree) model (only the variables exceeding the threshold is retained (*m*))

III. Compute errOBB rates of RF (averaged over 50 runs) of the nested models starting from the one with only the most important variable and ending with one including all important variables previously kept

IV. Finally, the variables of the model leading to smallest errOOB are selected

3. Bayesian Regression Model

Bayesian inference is one of the most powerful techniques of estimation. This technique of estimation has various advantages over point estimation. Main advantage of this technique of estimation is that it uses the prior information to generate posterior distribution. The posterior distribution contains more information due to incorporation of extra information in the form of prior distribution.

3.1 Steps of Bayesian approach of regression estimation

I. Model specification

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$$

II. Selection of prior distribution

$$\beta \sim N(\hat{\beta}, V_\beta \sigma^2), V_\beta = (X^T X)^{-1}$$

III. Find the likelihood function

The likelihood function is the joint probability density function of normal distribution

IV. Apply Bayes Theorem and generate posterior distribution using Markov Chain Monte Carlo (MCMC) Method

$$\pi(\theta|y) \propto L(Y|\theta)\pi(\theta)$$

V. Find the expectation of the posterior distribution

$$\pi^*(\theta|y) = \frac{f(y|\theta)\,p(\theta)}{\sum_{\theta \in E} f(y|\theta)\,p(\theta)}$$

4. **Time series approaches:**

The yield values can also be modelled using time series models. We have adopted the multivariate approach using weather variables as exogenous variables effecting the yield. The time series modelling can be divided into two approaches:

Univariate Approach: ARIMA (Auto Regressive Integrated Moving Average)

Multivariate Approach: ARIMAX (ARIMA with independent exogenous variables (X)).

ARIMA model is characterised by the notation (*p*,*d*,*q*) where, p ,d and q denote the orders of auto-regression, integration (differencing) and moving average respectively. A stationary ARMA (*p*,*q*) process is defined by the equation :

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} + e_t$$

Where e_t`s are normally distributed with zero mean and a constant variance.

The ARIMAX model is an extended version of the ARIMA model. It includes also other independent exogenous variables (X). Given a time-series process $\{x_t, y_t\}$, where x_t and y_t are real valued random variables, ARIMAX model assumes the form

$$\left(1 - \sum_{s=1}^{p} \alpha_s L^s\right) y_t = \mu + \sum_{s=1}^{q} \beta_s' L^s x_t + \left(1 + \sum_{s=1}^{r} \gamma_s L^s\right) e_t,$$

where $Ls(y_t)=y_{t-s}$

Specification of ARIMAX model

- The first step in building an ARIMAX model consists of identifying a suitable ARIMA model for the endogenous variable Y.
- The ARIMAX model concept requires to test for stationarity of exogenous variable before modeling.
- The transformed variable is added to the ARIMA model in the second step, in which the lag length *r* is also determined.
- Nonlinear least square estimation procedure is employed to estimate the parameters of ARMAX model.

4.1 Model to deal heteroscedastic errors

The GARCH model focuses on capturing the clustering of volatility in yield when the conditional variance at time *t* is modelled as a deterministic function of lagged values of conditional variances and squared yields, given by

$$\varepsilon_t = \xi_t h_t^{1/2}$$

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} b_j h_{t-j}$$

$\xi_t \sim IID(0,1),\ a_0 > 0, a_i \geq 0,\ i = 1,2,\ldots,q.\ \ b_j \geq 0,\ \ j = 1,2,\ldots,p$

The test for conditional heteroscedasticity is the ARCH-Lagrange multiplier (ARCH-LM) test of Engle (1982). This test is equivalent to usual *F*-statistic for testing

$$H_0: a_i = 0, 1, 2, \ldots q$$

in the linear regression

$$\varepsilon_t^2 = a_0 + a_1 \varepsilon_{t-1}^2 + \cdots + a_q \varepsilon_{t-q}^2 + e_t, \qquad t = q+1, \ldots, T$$

$$SSR_0 = \sum_{t=q+1}^{T} (\varepsilon_t^2 - \varpi)^2 \qquad \varpi = \sum_{t=q+1}^{T} \varepsilon_t^2 / T$$

Then, under Ho,

$$F = \frac{(SSR_0 - SSR_1)/q}{SSR_1(T - q - 1)}$$

is asymptotically distributed as chi-squared distribution with q degrees of freedom. The decision rule is to reject

$$H_0 \text{ if } F > \chi_q^2(\alpha)$$

4.2 Weather Indices based Automated Yield Forecasting System (WIAYFS)

For ease of implementation and reaching out to more researchers and users, WIAYFS (Weather Indices based Automated Yield Forecasting System), a webtool has also been developed at ICAR-IASRI. In the webtool stepwise regression model based on weather indices along with other models such as ARIMAX, LASSO regression, Bayesian Regression model and Random Forest technique. The calculation of weather indices from raw weather data process has been automized along with fitting of various models. WIAYFS is being implemented by IMD (India Meteorological Department), New Delhi. The webtool can be accessed at following URL http://wiayfs.icar.gov.in/wiayfs.

5. Conclusions

Machine learning techniques such as support vectors, Elastic net, etc can also be implemented for efficient variable selection.Random Forest and LASSO regression techniques can also be explored and compared with the existing ones. Other non-linear and multivariate models under time series framework can also be explored. These Machine Learning techniques being highly data driven needs to be validated extensively.

Suggested Readings:

Agrawal R, Jain R C and Jha M P. 1986. Models for studying rice crop-weather relationship. Mausam 37(1): 67–70

Gelman A, Carlin J B, Hal S, Dunson D B, Vehtari A and Rubin D B. 2014. Bayesian data analysis. CRC Press, Taylor and Francis Group, New York.

Singh K N, Singh, K K, Kumar S, Panwar S and Gurung, B. 2019. Forecasting crop yield through weather indices through LASSO. Indian Journal of Agricultural Sciences 89 (3): 540–544.

Tannura M A, Irwin S H and Good D L. 2008. A Review of the Literature on Regression Models of Weather, Technology, and Corn and Soybean Yields in the U.S. Marketing and Outlook Research Report 2008-02, Department of Agricultural and Consumer Economics. University of Illinois at Champaign-Urbana.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B 58(1): 267–88.

OVERVIEW OF BIOINFORMATICS AND APPLICATIONS IN AGRICULTURE

Mir Asif Iquebal, Sarika and Dinesh Kumar

ICAR-Indian Agricultural Statistics Research Institute

Introduction

Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyze and interpret biological data. It plays a crucial role in managing, processing, and understanding vast amounts of data generated by modern biological research, such as genomics, proteomics, and metabolomics. In agriculture, bioinformatics has become a transformative tool, enabling advancements in crop improvement, disease resistance, and sustainable farming practices through genomic and computational approaches. Following are the few key applications of bioinformatics in agriculture:

1. **Crop Improvement:** Bioinformatics aids in genome sequencing, gene identification, and trait selection, helping develop high-yield and stress-resistant crops. The various genomic sequencing platforms, along with the bioinformatics analysis help decode plant genomes, identifying genes associated with desirable traits like drought tolerance, disease resistance, and higher yield.
2. **Marker-Assisted Breeding**: By analyzing genetic markers, bioinformatics accelerates the development of improved crop varieties with enhanced traits.
3. **Pest and Disease Management:** By analyzing pathogen genomes, bioinformatics facilitates the identification of disease-resistant genes and development of biopesticides. Genomic data aids in understanding pathogen biology, enabling the development of targeted control strategies.
4. **Soil Microbiome Studies:** Metagenomics helps study beneficial soil microbes that enhance soil fertility and plant growth. Analyzing soil and plant microbiomes helps optimize nutrient uptake and improve soil health, contributing to sustainable agriculture.
5. **Livestock Genomics:** Genomic selection in livestock breeding improves disease resistance, milk production, and meat quality.
6. **Climate Resilience:** Identifying genes linked to drought and heat tolerance supports the development of climate-smart crops.
7. **Data-Driven Precision Farming**: Integrating bioinformatics with agricultural data enhances decision-making for crop management, resource allocation, and yield prediction.

With advances in sequencing technologies and data analytics, bioinformatics is revolutionizing agriculture by promoting precision farming, sustainable resource management, and food security.

Fig. 1: Integration of Omics Knowledge in wider areas of research

The term "Bioinformatics" was first coined by Paulien Hogeweg and Ben Hesper in 1970 who used it to describe the study of informatics processes in biological systems, long before the field became associated with computational biology and genomics. A few definitions of bioinformatics are as follows:

General Definition: Bioinformatics is the use of computational and statistical methods to analyze and interpret biological data, especially large datasets like DNA sequences, protein structures, and gene expressions.

National Center for Biotechnology Information (NCBI) Definition:

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The main goal of bioinformatics is to enable the discovery of new biological insights through the organization and analysis of biological data.

Oxford Dictionary Definition:

Bioinformatics is the science of collecting and analyzing complex biological data such as genetic codes.

National Human Genome Research Institute (NHGRI) Definition:

Bioinformatics is a subdiscipline of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, most often DNA and amino acid sequences. Bioinformatics is a scientific discipline that has emerged in response to accelerating demand for a flexible and intelligent means of storing, managing and querying large and complex biological data sets. The ultimate aim of bioinformatics is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. At the beginning of the genomic revolution, the main concern of bioinformatics was the creation and maintenance of a

database to store biological information such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of an interface whereby researchers could both access existing data as well as submit new or revised data (e.g. to the NCBI, http://www.ncbi.nlm.nih.gov/). More recently, emphasis has shifted towards the analysis of large data sets, particularly those stored in different formats in different databases. Ultimately, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.

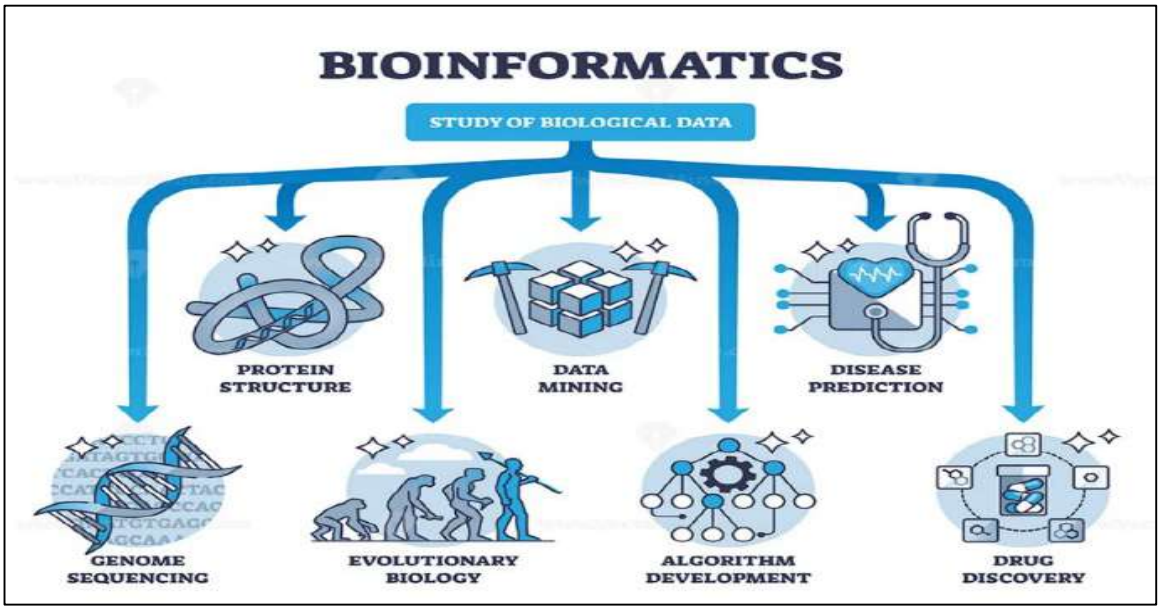


Figure 2. Various facets of Bioinformatics

(Source: https://datascienceforbio.com/computational-biology-vs-bioinformatics/)

Origin & History of Bioinformatics

Over a century ago, bioinformatics history started with an Austrian monk named Gregor Mendel. He is known as the “Father of Genetics". He cross-fertilized different colors of the same species of flowers. He kept careful records of the colors of flowers that he cross-fertilized and the color(s) of flowers they produced. Mendel illustrated that the inheritance of traits could be more easily explained if it was controlled by factors passed down from generation to generation.

After this discovery of Mendel, bioinformatics and genetic record keeping have come a long way. The understanding of genetics has advanced remarkably in the last thirty years. In 1972, Paul Berg made the first recombinant DNA molecule using ligase. In that same year, Stanley Cohen, Annie Chang and Herbert Boyer produced the first recombinant DNA organism. In 1973, two important things happened in the field of genomics:

1. Joseph Sambrook led a team that refined DNA electrophoresis using agarose gel, and
2. Herbert Boyer and Stanely Cohen invented DNA cloning. By 1977, a method for sequencing DNA was discovered and the first genetic engineering company, Genetech was founded.

During 1981, 579 human genes had been mapped and mapping by in situ hybridization had become a standard method. Marvin Carruthers and Leory Hood made a huge leap in bioinformatics when they invented a method for automated DNA sequencing. In 1988, the Human Genome Organization (HUGO) was founded. This is an international organization of scientists involved in Human Genome Project. In 1989, the first complete genome map was published of the bacteria Haemophilus influenza.

The following year, the Human Genome Project was started. In 1991, a total of 1879 human genes had been mapped. In 1993, Genethon, a human genome research center in France produced roduced a physical map of the human genome. Three years later, Genethon published the final version of the Human Genetic Map which concluded the end of the first phase of the Human Genome Project.

Bioinformatics was fuelled by the need to create huge databases, such as GenBank and EMBL and DNA Database of Japan to store and compare the DNA sequence data erupting from the human genome and other genome sequencing projects. Today, bioinformatics embraces protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways of numerous species

Importance

The greatest challenge facing the molecular biology community today is to make sense of the wealth of data that has been produced by the genome sequencing projects. Cells have a central core called nucleus, which is storehouse of an important molecule known as DNA. They are packaged in units known as chromosomes. They are together known as the genome. Genes are specific regions of the genomes (about 1%) spread throughout the genome, sometimes contiguous, many times non-contiguous. RNAs similarly contains informations, their major purpose is to copy information from DNA selectively and to bring it out of the nucleus for its use. Proteins are made of amino acids, which are twenty in count (researchers are debating on increasing this count, as couple of new ones are claimed to be identified).

The gene regions of the DNA in the nucleus of the cell is copied (transcribed) into the RNA and RNA travels to protein production sites and is translated into proteins is the Central Dogma of Molecular Biology. Portions of DNA Sequence are transcribed into RNA. The first step of a cell is to copy a particular portion of its DNA nucleotide sequence (i.e. gene).

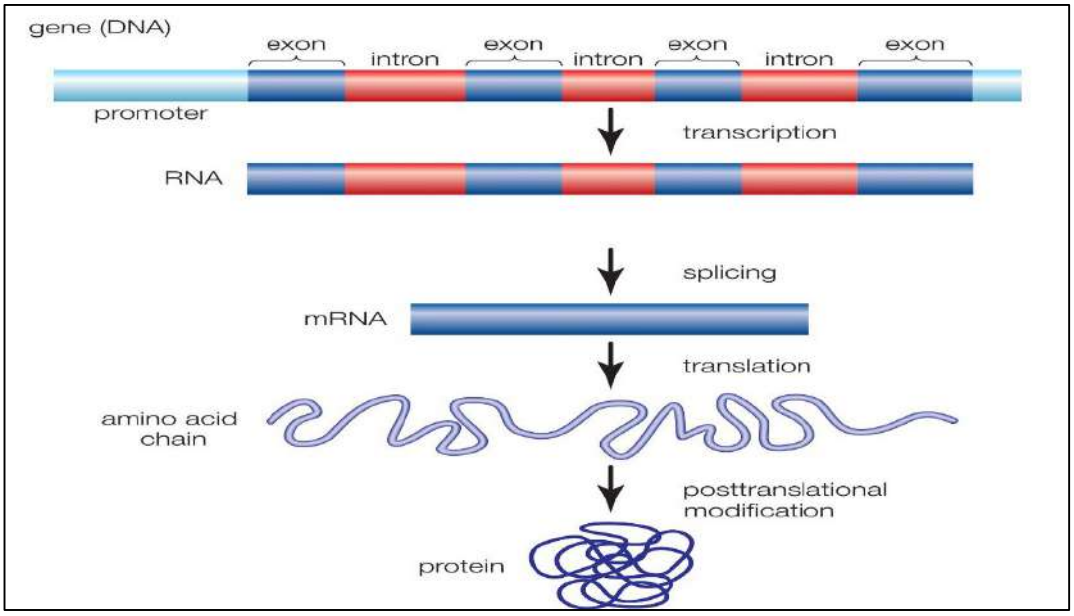


Figure 3. Representation of a gene (*(https://www.britannica.com/science/amino-acid/Nonstandard-amino-acids)*

Biological Database

Sequences and structures are only among the several different types of data required in the practice of the modern molecular biology. Other important data types includes metabolic pathways and molecular interactions, mutations and polymorphism in molecular sequences and structures as well as organelle structures and tissue types, genetic maps, physiochemical data, gene expression profiles, two dimensional DNA chip images of mRNA expression, two dimensional gel electrophoresis images of protein expression data.

Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only proteins. Thus, a biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated (Attwood *et al*., 2002). The main functions of a biological database are to provide the computer readable form of biological data and make the data available to the scientists and researchers spread all over the world.

Current biological databases use all three types of database structures, i.e. flat files, relational, and object oriented. Despite the drawbacks of using flat files in database management, many biological databases still use this format since this system involves minimum amount of database design and the search output can be easily understood by working biologists.

Based on their contents, biological databases can be roughly divided into three categories:

i. primary databases,
ii. secondary databases, and
iii. specialized databases.

Primary databases contain original biological data. They are archives of raw sequence or structural data submitted by the scientific community. GenBank and Protein Data Bank (PDB) are examples of primary databases.

Secondary databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples of secondary databases are SWISS-Prot and Protein Information Resources (PIR).

Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

Global and Local Alignment

Global Alignment

Global alignments, which attempt to align every residue of each sequence, when the size of the sequences are similar or of equal size. A general global alignment technique is based on dynamic programming i.e., Needleman-Wunsch algorithm. This can be easily understood with the following two sequences aligned globally as follows

G A A T T C A G T T A (sequence #1)

G G A T C G A (sequence #2)

In simple dynamic programming principle, we construct a matrix. The matrix will be filled by inserting 0 or 1 where ever there is a mismatch or match. We also penalize the gaps with 0 as a simple case. Following steps are needed for construction of the matrix

i. Initialization
ii. Matrix fill (scoring)
iii. Traceback (alignment)

Local Alignment

Local alignments are more useful for dissimilar sequences that may contains regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment method based on dynamic programming. A local alignment searches for regions of local similarity between two sequences and need not include the entire length of the sequences. This can be done by reading a scoring matrix that contains values for every possible residue or nucleotide match or mismatch. The Smith-Waterman algorithm is a member of the class of algorithms that can calculate the best score and local alignment in the order of m*n steps, where 'm' and 'n' are the lengths of the two sequences. Local alignment methods only report the best matching areas between two sequences while there may be a large number of alternative local alignments which do not score as highly as the best alignment done by this algorithm.

Consider the two DNA sequences to be globally aligned are:

ACACACT (x=7, length of sequence 1)

AGCACAC (y=7, length of sequence 2)

It also follows three steps

i. Initialization
ii. Matrix fill (scoring)
iii. Traceback (alignment)

Genome Assembly

Sequence assembly refers to aligning and merging fragments of DNA sequence in order to construct the original sequence. This is mandatory as DNA sequencing technology cannot read whole genomes in one go, but rather reads small pieces, in random order, of between 20 and 1000 bases, depending on the technology used. Recent advances in sequencing technology made it possible to generate vast amounts of sequence data. The fragments produced by these high-throughput methods are, however, far shorter than in traditional Sanger sequencing.

The first sequence assemblers began to appear in the late 1980s and early 1990s as variants of simpler sequence alignment programs to piece together vast quantities of fragments generated by automated sequencing instruments called DNA sequencers. Algorithms were developed for whole genome shotgun (WGS) fragment assembly, including Atlas,

Arachne, Celera, PCAP, Phrap (www.phrap.org) and Phusion. All these programs rely on the overlap-layout-consensus approach where all the reads are compared to each other in a pair-wise fashion.

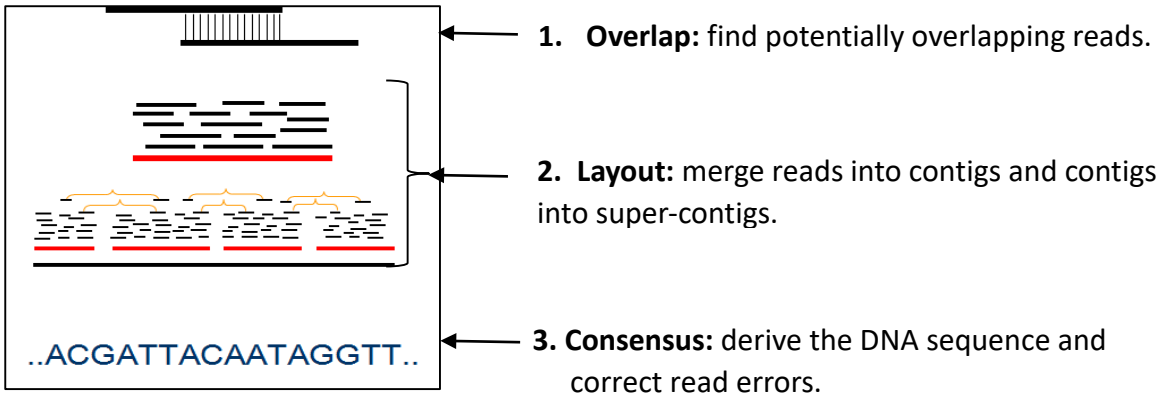


Fig. 4: Steps to assemble a genome

The resulting (draft) genome sequence is produced by combining the information sequenced "**contigs"** and then employing linking information to create **"scaffolds"** (Figure 1.). Scaffolds are positioned along the physical map of the chromosomes creating a "golden path".

Recently, new sequencing methods have emerged. Commercially available technologies include pyrosequencing (454 Sequencing), sequencing by synthesis (Illumina) and sequencing by ligation (SOLiD). The reads produced by these next-generation sequencing technologies are much shorter than traditional Sanger reads. Because of their shorter length, they must be produced in large quantities and at greater coverage depths than the earlier sequencing techniques. Whereas long reads provide long overlaps, to disambiguate repeats from real overlaps, short reads within repeats offer fewer differences to judge from. These issues have led several research teams to design de novo assembly tools specifically for these very short reads.

Types of Sequencers and Data Format

Illumina	:	FASTQ
SoLID/ABI-Life:		FASTA
Roche 454	:	SFF
Ion Torrent	:	SFF or FASTQ

Types of Assembly

There are two type of assembly base on the availability of reference genome:

a) De novo Assembly: Reads are aligned to each other to form a consensus sequences that are called contigs.

b) Reference genome assembly: Here reads are aligned with the available reference genome to form a consensus sequences.

Molecular Markers

Microsatellites are simple sequence tandem repeats (SSTRs). The repeat units are generally di-, tri-, tetra- or penta- nucleotides. For example, a common repeat motif in birds is ACn, where the two nucleotides A and C are repeated in bead-like fashion a variable number of times (n could range from 8 to 50). They tend to occur in non-coding regions of the DNA (this should be fairly obvious for long dinucleotide repeats) although a few human genetic disorders are caused by (trinucleotide) microsatellite regions in coding regions. On each side of the repeat unit are flanking regions that consist of "unordered" DNA. The flanking regions are critical because they allow us to develop locus-specific primers to amplify the microsatellites with PCR (polymerase chain reaction). That is, given a stretch of unordered DNA 30-50 base pairs (bp) long, the probability of finding that particular stretch more than once in the genome becomes vanishingly small (if the four nucleotides occur with equal probability then the probability of a given 50 bp stretch is 0.2550. In contrast, a given repeat unit (say AC19) may occur in thousands of places in the genome. We use this combination of widely occurring repeat units and locus-specific flanking regions as part of our strategy for finding and developing microsatellite primers. The primers for PCR will be sequences from these unique flanking regions. By having a forward and a reverse primer on each side of the microsatellite, we will be able to amplify a fairly short (100 to 500 bp, where bp means base pairs) locus-specific microsatellite region.

Advantages of Microsatellites as Genetic Markers:

- Locus-specific (in contrast to multi-locus markers such as minisatellites or RAPDs)
- Codominant (heterozygotes can be distinguished from homozygotes, in contrast to RAPDs and AFLPs which are "binary, 0/1")
- PCR-based (means we need only tiny amounts of tissue; works on highly degraded or "ancient" DNA)
- Highly polymorphic ("hypervariable") -- provides considerable pattern
- Useful at a range of scales from individual ID to fine-scale phylogenies

What Uses Do Microsatellites Serve in Agriculture?

Microsatellites are useful markers at a wide range of scales of analysis. Until recently, they were the most important tool in mapping genomes -- such as the widely publicized **mapping** of the human genome. They serve a role in biomedical diagnosis as markers for certain disease conditions. That is, certain microsatellite alleles are associated (through genetic linkage) with certain mutations in coding regions of the DNA that can cause a variety of medical disorders. They have also become the primary marker for DNA testing in **forensics (**court) contexts -- both for human and wildlife cases (e.g., Evett and Weir, 1998). The reason for this prevalence as a forensic marker is their high specificity. Match identities for microsatellite profiles can be very high (probability that the evidence from the crime scene is not a match with that of the suspect is < one in many millions in some cases). In a biological/evolutionary context they are useful as markers for **parentage analysis**. They can also be used to address questions concerning **degree of relatedness** of individuals or groups. For captive or endangered species, microsatellites can serve as tools to evaluate inbreeding levels (FIS). From there we can move up to the **genetic structure of subpopulations** and populations (using tools such as F-statistics and genetic distances). They can be used to assess **demographic history** (e.g., to look for evidence of population

bottlenecks), to assess effective population size (Ne) and to assess the magnitude and directionality of **gene flow between populations**. Microsatellites provide data suitable for **phylogeographic studies** that seek to explain the concordant biogeographic and genetic histories of the floras and faunas of large-scale regions. They are also useful for fine-scale phylogenies -- up to the level of closely related species. An overview by Selkoe and Toonen (2006) provides a useful practical guide to the use of microsatellites as **genetic markers**.

STRs/ SSRs plays important role in mapping, trait improvement, variety development, variety identification and product traceability. Traditionally, characterization of varieties is based on phenotypic observation but it is very difficult to distinguish varieties with very similar morphological characteristics and identification of the cultivars accurately is essential for maintaining cultivar integrity and Plant Breeders' Rights.

Limited studies have been reported in variety identification of tomato using STR DNA markers. In one study, out of 20 STR markers, only 11 were able to discriminate 47 varieties (Sardaro *et al*., 2013) and in another study, 12 markers could differentiate 34 varieties (Srivastava *et al*., 2011). Studies based on 6000 SNP markers over 93 varieties have demonstrated that SNP based variety differentiation is also possible (**Viquez-Zamora *et al*., 2013).** However, in such SNP based studies, the genotyping data of "Moneymaker" and "Moneyberg" varieties were completely identical leading to no differentiation at all. To overcome this Iquebal *et al*., 2013 has reported more than 1.4 million markers in tomato. DNA fingerprinting is an appropriate tool to track and trace the tomato supply chain, ensuring not only authenticity and integrity of the products but also the absence of any possible genetic contamination by other species or unwanted components (Marmiroli *et al*., 2003; Marmiroli *et al*., 2009; Agrimonti *et al*., 2011).

Such use of STR in plant variety identification is well reported in many crops like barley varieties (Karakousis *et al*., 2010), S. tuberosum ssp. tuberosum (Kawchuk *et al*., 1991), sugarcane (Manigbas and Villegas, 2004), capsicum (Shirasawa *et al*., 2013), eggplant (Stagel et al., 2008) and identification of Basmati rice from that of non-Basmati rice (Archak *et al*., 2007). Also, the microsatellite STR markers are the method of first choice to complement the DUS (Distinctness, Uniformity and Stability) testing procedure.

Tomato STR database can be a useful tool in MAS programme of tomato improvement (Iquebal *et al*., 2013). Such use of STR in crop improvement is already reported in sorghum (Wang *et al*., 2012), tagging stem rust resistance gene Sr35 in wheat, *Fusarium* head blight resistance in wheat, leaf rust resistance gene Lr35 in wheat and mapping of resistance gene effective against Karnal bunt pathogen of wheat. Wheat improvement programs to enhance leaf rust resistance using STR markers has been attempted. STR markers are also used for introgression programs for trait improvement, for example Soltol QTLs in rice. The location of the Saltol QTL on chromosome 1 and identification of additional QTLs associated with salt tolerance is well identified (Thomson et al., 2010).

Limits to utility of microsatellites: Microsatellite DNA is probably rarely useful for higher-level systematics. That is because the mutation rate is too high. Across highly divergent taxa two problems arise. First, the microsatellite primer sites may not be conserved (that is the primers we use for Species A may not even amplify in Species B). Second, the high mutation rate means that homoplasy becomes much more likely, we can no longer safely assume that two alleles identical in state are identical by descent (from a common, meaning shared not abundant, ancestor). As a concrete example imagine two species, each with an AC19 allele that occurs at high frequency. If the populations diverged long ago it becomes increasingly likely that the way those alleles arose took different pathways (e.g., in one species the AC19 arose from an ancestor that went from AC18 to AC19 to AC20 then back

to AC19; in the other species the ancestral AC18 went to AC19 and stayed there. Any inferences we make about the species relationships based on the AC19 similarity would be misleading). The identity in state does not correspond to the identity by descent that provides (reliable) phylogenetic signal. A further potential drawback of using microsatellites is that we tend to have relatively few loci to work with (4-20). In some situations, that raises the probability of having a bias due to forces such as selection acting on one or more loci that may give a misleading impression relative to the true pattern of change for the genome as a whole.

High Performance Computation at IASRI, New Delhi

The National Agricultural Bioinformatics Grid in ICAR consists of an advanced HPC infrastructure at IASRI, New Delhi and moderate HPC facilities at the domain centres for undertaking research in the field of agricultural bioinformatics. Clusters are collections of computers that are connected together. The special sets of software are used to configure HPC environment. This set up has been named as Advanced Supercomputing Hub for Omics Knowledge in Agriculture (ASHOKA). The importance of HPC is rapidly growing because more and more scientific and technical problems are being studied on the huge data sets which require very high computational power as well. HPC offers environment for biologists, scientists, analysts, engineers and students to utilize the computing resources in making vital decisions, to speed up research and development, by reducing the execution time.

The following HPC infrastructure was set up under NAIP project NABG which are as follows in the form of clusters, network and storage in 2013 (NABG Report). The grid has been established using the following network diagram as in figure 5.

Types of Clusters

a. 256 Nodes Linux Based Cluster with two masters
b. 16 Nodes Windows Based Cluster with one master
c. 16 Nodes GPGPU Based Linux Cluster with one master
d. 16 Nodes Linux based SMP system
e. 16 Nodes Linux Based Cluster at each of the five domains with one master

Types of Networks

a. High bandwidth network with low latency (Q-logic QDR InfiniBand switch)
b. Gigabit network for cluster administration and management
c. ILO3 Management Network

Types of Storage

a. Parallel File System (PFS) for computational purpose
b. Network Attached Storage (NAS) for user Home Directory
c. Archival Storage for back up.

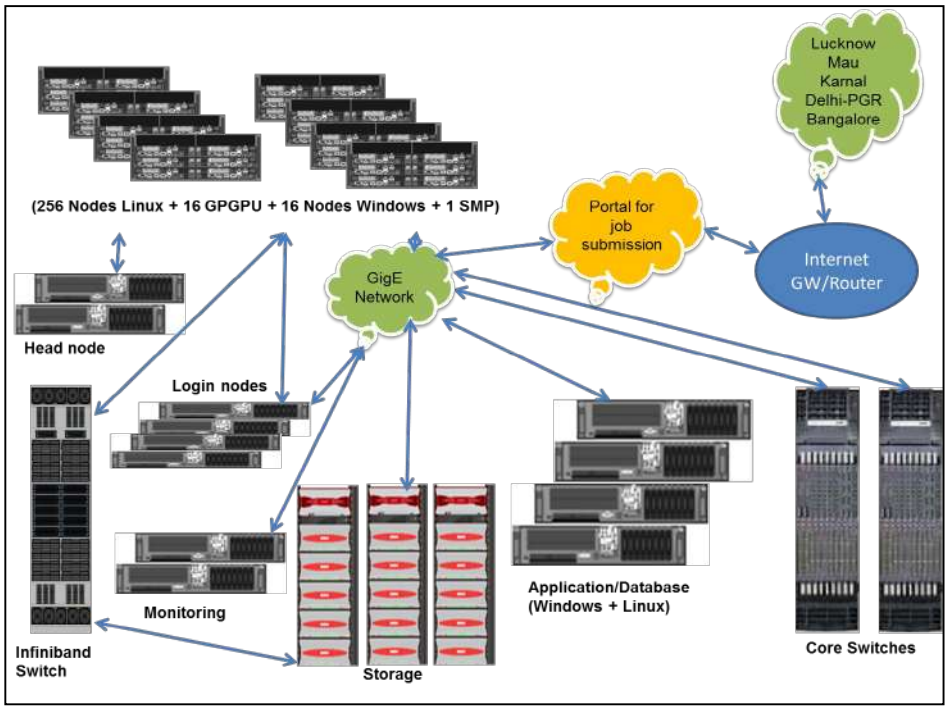


Fig. 5: Network diagram of NABG Grid

Bioinformatics faces several challenges due to the rapid growth of biological data and the need for efficient computational methods. Though wet-lab validation is always recommended owing to the *in silico* analysis. Some key challenges include:

i. **Data Management and Storage:** The explosion of biological data (e.g., genomic sequences, proteomic data) requires vast storage capacity. Efficient databases and retrieval systems are needed to manage, store, and update large datasets.
ii. **Data Integration and Standardization:** Biological data comes from different sources and in various formats, making integration complex. There is a need for standardized formats, protocols, and databases for seamless data exchange.
iii. **Computational Power and Algorithm Development:** Analyzing large-scale biological datasets requires high-performance computing. Developing efficient algorithms for sequence alignment, protein structure prediction, and genome assembly is challenging.
iv. **Accuracy and Noise in Data:** Biological experiments often generate noisy, incomplete, or inconsistent data. Robust computational models and machine learning techniques are needed to improve data accuracy.
v. **Interdisciplinary Skill Gap:** Bioinformatics requires expertise in biology, computer science, mathematics, and statistics. There is a shortage of professionals with expertise across all these disciplines.
vi. **Data Privacy and Security:** Genomic and medical data contain sensitive information, requiring strict security and ethical guidelines. Ensuring data confidentiality while allowing researchers access to datasets is a challenge.
vii. **Interpretation of Biological Meaning:** Computational tools generate vast amounts of data, but deriving meaningful biological insights is difficult. Validation through experimental biology is necessary to confirm computational predictions.

viii. **Scalability and Big Data Challenges:** The increasing size of biological data demands scalable storage, retrieval, and analysis techniques. Cloud computing and distributed computing solutions are being explored to address this issue.

ix. **Cost of Computational Infrastructure:** High-performance computing resources and data storage solutions are expensive. Small research labs and developing countries often struggle with funding and access to bioinformatics tools.

x. **Keeping Up with Rapid Technological Advancements:** New sequencing technologies and data analysis methods are emerging rapidly suggesting the researchers to continuously update their knowledge and tools to stay relevant.

References

1. Agrimonti, C., Vietina, M., Pafundo, S., Marmiroli, N. 2011. The use of food genomics to ensure the traceability of olive oil. *Trends Food Sci. Tech*, 22:237-244.
2. Archak, S., Lakshminarayanareddy, V., Nagaraju J. 2007. High-throughput multiplex microsatellite marker assay for detection and quantification of adulteration in Basmati rice (*Oryza sativa*). *Electrophoresis*, 28:2396-2405.
3. Babiker, E., Ibrahim, A.M.H., Yen, Y., Stein, J. 2009. Identification of a microsatellite marker associated with stem rust resistance gene *Sr35* in wheat. *Aust J. Crop Sci.,* 3:195–200.
4. Becher, S.A., Steinmetz, K., Weising, K., Boury, S., Peltier, D., Renou, J.P., Kahl, G., Wolff, K. 2000. Microsatellites for variety identification in *Pelargonium*. *Theor. Appl. Genet.*, 101:643-651.
5. Marmiroli, N., C. Peano, and E. Maestri. 2003. Advanced PCR techniques in identifying food components. Pp. 3–33 in M. Lees, ed. Food authenticity and traceability. Woodhead Publishing. Cambridge, U.K.
6. Marmiroli, N., E. Maestri, S. Pafundo, and M. Vietina. 2009. Molecular traceability of olive oil: From plant genomics to food genomics. Pp. 157–172 in L. Berti and J. Maury, eds. Advances In Olive Resources. Transworld Research Network, Trivandrum, India.
7. Karakousis, A., Gustafson. J. P., Chalmers. K. J., Barr. A. R., & Langridge, P. (2003). A consensus map of barley integrating SSR, RFLP, and AFLP markers. Aus. J. Agr. Res., 54, 1173-1185.
8. Kawchuk, L. M., Martin, R. R. and Macpherson, J. (1991). Sense and antisense RNA-mediated resistance to potato leafroll virus in Russet Burbank potato plants. Mol Pl-Microbe Interact, 4: 247–253.
9. Manigbas N and Villegas L 2004. Microsatellite markers in hybridity tests to identify true hybrids of sugarcane. Philippine J. Crop Science 29: 23-32.
10. Shirasawa K, Ishii K, Kim C, Ban T, Suzuki M, Ito T, Muranaka T, obayashi M, Nagata N, Isobe S, Tabata S (2013) Development of Capsicum EST–SSR markers for species identification and in silico mapping onto the tomato genome sequence. Mol Breed 31:101–110.
11. Iquebal, M.A., Sarika, Arora, V. et al. First whole genome based microsatellite DNA marker database of tomato for mapping and variety identification. BMC Plant Biol 13, 197 (2013).

OVERVIEW OF CALIBRATION ESTIMATION IN SURVEY SAMPLING

Kaustav Aditya

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Sampling techniques are used in all kinds of surveys all over the world. In many surveys, the objective is usually to obtain some descriptive measures with respect to the characteristics of the entire population under study. Such surveys are very common and are required for generation of data for national planning and socio-economic development. In the context of agriculture, for example, data on crop production, utilization of land and water resources etc. are required for planning purposes. Sampling methods are also used in various censuses. In fact, except for certain basic information required in respect of every individual or area, data on various items are collected on a sampling basis. Sampling methods are used to provide counter checks and speed up tabulation and publication of results. Sampling methods are used extensively in business and industry to increase operational efficiency. They play an important role in problems encountered in market research such as estimating the size of readership of news-magazines and newspapers or finding the reactions of consumers to new products recently introduced in market. Key references are Tchuprow (1923), Neyman (1934), Mahalanobis (1944), Yates (1946), Hansen *et al.* (1953), Cochran (1953), Sukhatme (1954), Kish (1965) and Kish and Frankel (1974).

In sample surveys, auxiliary information on the finite population is often used to increase the precision of estimators of finite population total or mean or distribution function. In the simplest settings, ratio and regression estimators incorporate known finite population parameters of auxiliary variables. The Calibration Approach proposed by Deville and Sarndal (1992) is one of the other techniques widely used for making efficient use of auxiliary information in survey estimation. In this study, they developed estimator for population total, variance of the estimator and the corresponding variance estimator. Here they have considered the cases of uni-variate and multivariate auxiliary information. They have also shown that the calibration estimator can be represented as the generalized regression estimator (GREG) under certain conditions. They have also showed that the GREG estimator was a special case of calibrate estimator when the chosen distance function is the Chi-square distance function. Different distance functions were considered to minimize the distance between the original weights and the new weights. An application of this technique was given in connection after calibration on the known counts of a two way frequency table.

2. Calibration Estimation Technique

Survey statisticians are always concerned with improvement of methods for estimation of the finite population total, mean, proportion and other parameters. The estimators which use auxiliary variables are often more accurate than the standard ones. Calibration is commonly used in survey sampling to include auxiliary information to increase the precision of the estimators of population parameter. A calibration estimator uses calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights while also respecting a set of constraints, the calibration equations. For every distance measure there is a corresponding set of calibrated weights and a calibration estimator (Deville and Särndal (1992)). Definition:

The calibration approach to estimation for finite populations consists of

a. A computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s).
b. The use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units.
c. An objective to obtain nearly design unbiased estimates as long as nonresponse and other non-sampling errors are absent.

To describe calibration some relevant references to earlier literature are,

a. Calibration as a linear weighting method.

Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies. Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horwitz and Thompson (1952). Weighting became widely accepted. Later, poststratification weighting achieved the same status. Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent; the weights depend on the observed sample.

b. Calibration as a systematic way to use auxiliary information

Calibration provides a systematic way to take auxiliary information into account. As Rueda, Martínez, Martínez and Arcos (2007) point out, “in many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation”.

c. Calibration to achieve consistency

Calibration is often described as “a way to get consistent estimates”. (Here “consistent” refers not to “randomization consistent” but to “consistent with known aggregates”.) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. Consistency through calibration has a broader implication than just agreement with known population auxiliary totals. Consistency can, for example, be sought with appropriately estimated totals, arising in the current survey or in other surveys.

There are three major advantages of calibration approach in survey sampling.

I. The calibration approach leads to consistent estimates.
II. It provides an important class of technique for the efficient combination of data sources.
III. Calibration approach has computational advantage to calculate estimates.

The calibration approach focuses on the weights given to the units for the purpose of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The

calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. The population total is estimated by a linear estimator whose weights are as close as possible to some benchmark weights and which at the same time satisfy some calibration constraints with respect to some suitable auxiliary variables.

Consider a finite population $U=\{1,\ldots, k,\ldots, N\}$ consisting of N units. A sample s of size n is drawn without replacement according to a probabilistic sampling plan with inclusion probabilities $\pi_i = p_r(i \in s)$ and $\pi_{ij} = p_r(i \text{ and } j \in s)$ are assumed to be strictly positive and known. The study variable y is observed for each unit in the sample hence is known for all $i \in s,$ and the values $x_1, x_2, ..., x_N$ are known. Let y_i be the value of the variable of interest, y, for the i^{th} population element, with which is also associated an auxiliary variable x_i. For the elements $i \in s,$ observe (y_i , x_i). The population total of auxiliary variable x, $X = \sum_{i=1}^{N} x_i$ is assumed to be accurately known. The objective is to estimate the population total $Y = \sum_{i=1}^{N} y_i$. Deville and Sarndal (1992) used calibration on known population total X to modify the basic sampling design weights.

Let the Horvitz-Thompson estimator of the population total be $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i$,

where $d_i = \frac{1}{\pi_i}$ is the sampling weight, defined as the inverse of the inclusion probability for unit i. An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design. Its variance under the sampling design is given as

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

Now let us suppose that $\{x_i, i = 1, \ldots, N\}$ is available and $X = \sum_{i=1}^{N} x_i$, the population total for x is known. Ideally we would like, $\sum_{i=1}^{n} d_i x_i = X$. But sometimes this is not true. The idea behind calibration estimators is to find weights w_i , $i = 1, \ldots, N$ close to d_i, based on a distance function, such that, $\sum_{i=1}^{n} w_i x_i = X$. We wish to find weights w_i similar to d_i so as to preserve the unbiased property of the HT estimator. Once w_i is found the calibration estimator for $Y = \sum_{i=1}^{N} y_i$ would be $\hat{Y}_c = \sum_{i=1}^{n} w_i y_i$.

Given a sample s, we want to find w_i , $i = 1, \ldots, N$ close to d_i based on a distance function $D(w, d)$ subject to the constraint equation $\sum_{i=1}^{n} w_i x_i = X$. The optimization problem where we want to minimize

$$Q(w_1,...,w_n,\lambda)=\sum_{i=1}^{n}D(w_i,\ d_i)\ -\lambda\left(\sum_{i=1}^{n}w_ix_i-X\right) \qquad \ldots(2.1)$$

using the method of Lagrangian multipliers. There are various distance measures are available, some of them were,

Distance measures	$D(w, d)$
1. Chi-squared distance	$\frac{(w-d)^2}{2dq}$
2. Modified minimum entropy distance	$q^{-1}(w\log(\frac{w}{d})-w-d)$
3. Hellinger distance	$2\left(\sqrt{w}-\sqrt{d}\right)^2\Big/q$
4. minimum entropy distance	$q^{-1}(-d\log(\frac{w}{d})+w-d)$
5. Modified Chi-squared distance	$\frac{(w-d)^2}{2wq}$

Here q is the tuning parameter that can be manipulated to achieve the optimum minimal of the Eq. (2.1). A simple case considered by Deville and Sarndal (1992) is the minimization of chi-square type distance function given by $\sum_{i=1}^{n}\frac{(w_i-d_i)^2}{d_iq_i}$. Where q_i are suitably chosen weights. In most of the situations, the value of $q_i=1$. By minimizing the $\sum_{i=1}^{n}\frac{(w_i-d_i)^2}{d_iq_i}$ subject to constraint equation $\sum_{i=1}^{n}w_ix_i=X$ the weights w_i was obtained $w_i=d_i+\frac{d_iq_ix_i}{\sum_{i=1}^{n}d_iq_ix_i^2}\left(X-\sum_{i=1}^{n}d_ix_i\right)$.

Substitution of the value of w_i in $\hat{Y}_c=\sum_{i=1}^{n}w_iy_i$ gives

$$\hat{Y}_c=\sum_{i=1}^{n}d_iy_i+\frac{\sum_{i=1}^{n}d_iq_ix_iy_i}{\sum_{i=1}^{n}d_iq_ix_i^2}\left(X-\sum_{i=1}^{n}d_ix_i\right)$$

$$=\hat{Y}_{HT}+\hat{B}\left(X-\hat{X}_{HT}\right)$$

Where, $\hat{B}=\frac{\sum_{i=1}^{n}d_iq_ix_iy_i}{\sum_{i=1}^{n}d_iq_ix_i^2}$. Written in Written in this form, we see that $\hat{Y}_c$ is the same as the GREG estimator (Cassel *et al.*, 1976). In fact, the GREG estimator is a special case of the calibration estimator when the chosen distance function is the Chi-square distance

(Deville and Sarndal, 1992). In terms of efficiency, Deville and Sarndal showed that for medium to large samples, the choice of $D(w, d)$ does not make a large impact on the variance of $\hat{Y}_c$. The variance of the calibration estimator was given as,

$$\begin{aligned} V(\hat{Y}_c) &= V\left(\hat{Y}_{HT} + \hat{B}\left(X - \hat{X}_{HT}\right)\right) \\ &= V\left(\hat{Y}_{HT} - \hat{B}\hat{X}_{HT}\right) \\ &= V\left(\sum_{i=1}^{n} d_i (y_i - Bx_i)\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij}\left(d_i (y_i - Bx_i)\right)\left(d_j (y_j - Bx_j)\right) \end{aligned}$$

The estimator of variance of the estimator was given as, $\hat{V}(\hat{Y}_c) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\Delta_{ij}}{\pi_{ij}} (w_i e_i)(w_j e_j)$.

Where $e_i = y_i - \hat{\beta} x_i$ and $\Delta_{ij} = \left(\pi_i \pi_j - \pi_{ij}\right)$. This technique of calibration is called as the lower level calibration approach.

Higher level calibration makes use of known total as well as known variance of auxiliary character, whereas low level calibration utilizes only know total of auxiliary character. In the case of single auxiliary variables, Singh *et al.* (1998) proposed a high level calibration approach for variance estimation. The Yates-Grundy form of variance of the HT estimator is given as $V_{YG} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i \pi_j - \pi_{ij}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$.

…(2.2)

The usual estimator of variance of Eq. (2.2) is

$$\hat{V}_{YG} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 . \qquad \ldots(2.3)$$

The calibrated estimator of variance of Eq.(2.2) is

$$\hat{V}_{ss} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Omega_{ij}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 . \qquad \ldots(2.4)$$

Where Ω_{ij} are the calibrated weights, modified from the design weight $D_{ij} = \left(\pi_i \pi_j - \pi_{ij}\right)\big/\pi_{ij}$, $i \neq j$, given in eq. (2.3) by minimizing a distance measure between Ω_{ij} and D_{ij} subject to the constraint

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Omega_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i \pi_j - \pi_{ij}\right)\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 . \qquad \ldots(2.5)$$

Here, Ω_{ij} are the modified weight attached to the quadratic expression by Yates and

Grundy (1953) from the estimator and are as close as possible in an average sense for a given measure of the D_{ij} w.r.t. the constraint Eq (2.5). For simplicity, Singh *et al.* (1998) considered the two dimensional chi square type distance between Ω_{ij} and D_{ij} and minimized this distance subject to the constraint given in Eq.(2.5). The RHS of the constraint equation is obtained by collecting the information on every unit of the auxiliary character in the population from the past surveys, or census or administrative records etc. the two dimensional chi-square type distance measure is given as

$$D = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(\Omega_{ij} - D_{ij})^2}{D_{ij}Q_{ij}}$$

Now minimizing D w.r.t. to the given constraint equation gives,

$$\Omega_{ij} = D_{ij} + \frac{D_{ij}Q_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2}{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}D_{ij}Q_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^4}\left[\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}D_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2\right]$$

Putting the value of Ω_{ij} in Eq. (2.4) gives us a regression type of estimator of the variance of the lower order calibration estimator.

References

Aditya, K., Sud, U.C., Chandra, H. and Biswas, A., Calibration Based Regression Type Estimator of the Population Total under Two Stage Sampling Design. *Journal of Indian Society of Agriculture Statistics,* 2016, **70**(1), 19-24.

Aditya, K. and Sud, U. C., Higher order Calibration Estimators under Two Stage Sampling. *Statistics and Informatics in Agricultural Research*, Excel India Publication, New Delhi, 2015.

Cassel, C. M., Särndal, C. E. and Wretman, J. H., Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica,* 1976, **63**, 615-620.

Cochran, W. G., *Sampling Techniques*, 3rd Edition. John Wiley & Sons Publication, New York, 1977.

Deville, J. C. and Särndal, C. E., Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 1992, **87**, 376-382.

Estevao, V. M. and Särndal, C. E., A new perspective on calibration estimators. *Joint Statistical Meeting-Section on Survey Research Methods*, 2003, 1346-1356.

STATISTICAL DATA INTEGRATION IN SAMPLE SURVEYS

Rahul Banerjee

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110012

1.0 Introduction:

Probability sampling is widely considered the benchmark in survey statistics when it comes to finite population inference. Essentially, probability samples are chosen using known sampling designs, making them representative of the intended population. The fact that the selection probability is known allows for design-based inference, which adheres to the manner in which the data were collected. For further detailed discussions on this topic, references like Särndal et al. (2003), Cochran (1977), and Fuller (2009) offer valuable insights in their textbooks. Additionally, Kalton (2019) presents an extensive overview of survey sampling research spanning the last 60 years. Indeed, despite the merits of probability sampling, various practical challenges are encountered during the collection and analysis of such data (Baker et al., 2013; Keiding and Louis, 2016). Large-scale survey initiatives consistently confront increasing demands while having limited resources at their disposal. These demands often entail producing estimates for domains with small sample sizes and seeking more timely estimations. However, budget constraints lead to sample size reductions, and the declining response rates raise concerns about non-response bias. These challenges necessitate innovative approaches and techniques to address the complexities involved in maintaining the quality and accuracy of survey data under constrained conditions.

Data integration is an emerging area of research that aims to offer a timely solution to the aforementioned challenges. Its primary objectives are threefold: (1) minimizing the costs associated with surveys, (2) reducing respondent burden, and (3) maximizing the statistical information or, equivalently, the efficiency of survey estimation. More specifically, survey integration involves merging separate probability samples into a unified survey instrument (Bycroft, 2010). On a broader scale, one can also explore the combination of probability samples with non-probability samples.

Non-probability data have become increasingly available in survey statistics for research purposes, presenting exciting opportunities for new scientific discoveries. However, they also introduce additional challenges such as heterogeneity, selection bias, high dimensionality, among others. Over the years, substantial progress has been made in developing theories, methods, and algorithms to tackle these crucial challenges associated with the analysis of non-probability data. This chapter offers a systematic review of data integration techniques, encompassing the combination of probability samples, probability and non-probability samples, as well as probability and big data samples. By delving into this topic, researchers seek to harness the potential of data integration to overcome the limitations posed by traditional survey approaches and make the most of the diverse data sources available for more robust and efficient survey estimations.

2.0 Integration of Probability Samples:

Combining two or more independent survey probability samples is a problem frequently encountered in the practice of survey sampling. For simplicity of exposition, let $U = [U_1, U_2, \ldots, U_N]$ be the index set of N units for the finite population, with N being the known

population size. Let $(x_i', y_i)'$be the realized value of a vector of random variables $(X', Y)'$for unit *i*, where X consists of auxiliary variables and Y is the study variable of interest. The parameter of interest is the finite population mean of Y, i.e., $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Let I_i be the sample indicator, such that $I_i = 1$ indicates the selection of unit i into the sample and $I_i = 0$ otherwise. The probability $\pi_i = P(I_i = 1 | i \in U)$ is called the first-order inclusion probability and is known by the sampling design. The design weight is $d_i = \pi_i^{-1}$. The joint probability $\pi_{ij} = P(I_i I_j = 1 | i, j \in U)$ is called the second-order inclusion probability and is often used for variance estimation of the design-weighted estimator. In particular $\pi_{ii} = \pi_i; \forall i$. The sample size is $n = \sum_{i=1}^{N} I_i$.

The main advantage of probability sampling is to ensure design-based inference. For example, the Horvitz Thompson (HT) estimator of the population mean of *y*, denoted by $\bar{Y}$. It is denoted by:

$$\widehat{\mu_{HT}} = \frac{1}{N}\sum_{i:I_i=1} \frac{y_i}{\pi_i}$$

The design-variance estimator is obtained as:

$$\hat{V} = \frac{n}{N^2}\sum_{i:I_i=1}\sum_{i:I_j=1} \frac{(\pi_{ij} - \pi_i\pi_j)y_i y_j}{\pi_{ij}\pi_i\pi_j}$$

We consider multiple sources of probability data. For multiple datasets, we use the subscript letter to indicate the respective sample; for example, we use $d_{A,i}$ as the design weight of unit *i* in sample A.

Based on the available information from various data sources, intentional missingness is incorporated into each sample design. Table 1 illustrates the combined sample, which displays distinct patterns of missing data: monotone and non-monotone.

Regarding monotone missingness, our framework encompasses two common types of studies. The first type involves a large main dataset, to which additional information on crucial variables is collected for a subset of units. This is achieved through a two-phase sampling design, as described in works by Neyman (1938), Cochran (1977), and Wang et al. (2009). To illustrate, let's consider the U.S. Census of housing and population. The short form gathers 100% of the basic demographic information for all individuals, while the long form obtains about 16% of the sample, including other social, economic, and demographic details. This setup resembles a classical two-phase sampling problem, as discussed by Deming and Stephan (1940), and calibration weighting is employed for demographic variables to match the known population counts from the short form.

Furthermore, there is another approach for dealing with missing data, known as nonnested two-phase sampling. In this setup, we have a smaller, carefully designed validation dataset with extensive covariates, which is linked to a larger main dataset with fewer covariates. An example of nonnested two-phase sampling is evident in the US consumer expenditure survey. Two independent samples, sample A and sample B, are selected from the same finite population. Sample A is a diary survey sample, while sample B is a face-to-face survey sample. In sample A, both auxiliary information X and outcome Y are observed, while in sample B, only the common auxiliary information X is observed. To estimate detailed expenditure and income items, Zieschang (1990) proposed using sample weighting to combine the data from sample A and sample B. Another instance of nonnested two-phase sampling is observed in the Canadian Survey of Employment, Payrolls, and Hours, as

studied by Hidiroglou (2001). Sample A, a small sample from the Statistics Canada Business Register, captures observed study variables Y, such as the number of hours worked by employees and summarized earnings. On the other hand, sample B, a large sample drawn from Canadian Customs and Revenue Agency administrative data, captures the auxiliary variables X. In both examples, the two-phase sampling approach allows researchers to effectively utilize the information available in the different datasets and make more robust inferences by accounting for the missing data.

Table 1: Missingness patterns in the combined samples: "✓" means "is measured"

Monotone missingness				
	d	X	Y	
Sample A	✓	✓	✓	
Sample B	✓	✓		
Non-monotone missingness I				
	d	X	Y_1	Y_2
Sample A	✓	✓	✓	✓
Sample B	✓	✓	✓	
Sample C	✓	✓		✓
Non-monotone missingness II				
	d	X	Y_1	Y_2
Sample A	✓	✓	✓	
Sample B	✓	✓		✓

Note: d is the design weight, where the subscript indicates the sample, X is the vector of auxiliary variables and Y, Y_1 and Y_2 are scalar outcome variables

3.0 Two approaches for probability data integration:

Probability data integration methods can be classified into two main approaches based on the level of information to be combined: the macro approach and the micro approach.

Macro Approach:

In the macro approach, summary information is obtained from multiple data sources. This includes point estimates (e.g., mean or total) and their associated variance estimates. The goal is to combine these summary statistics to obtain a more efficient estimator of the parameter of interest. For example, this approach is commonly used to estimate population means or totals by combining information from different surveys or datasets.

Micro Approach:

In contrast, the micro approach involves creating single synthetic data that incorporate all available information from all data sources. These synthetic datasets are designed to be representative of the underlying population and contain detailed individual-level information. Researchers can then use these synthetic datasets to estimate various types of parameters of interest. This approach allows for more comprehensive and flexible analyses, as it retains the granularity of the original data.

3.1 Macro approach: generalized least-squares (GLS) estimation:

The problem of estimating totals at the population and domain levels by combining data from two independent probability samples has been investigated by several researchers, including Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004), Wu (2004), Ybarra and Lohr (2008), and Merkouris (2010). Merkouris (2004) and Merkouris (2010) offered a comprehensive analysis of survey integration using the generalized method of moments.

Our focus is on the monotone missingness pattern, although similar discussions can be applied to other missingness patterns as well. In each probability sample, we derive different estimators for the means of common items. The Generalized Least Squares (GLS) approach is employed to combine these estimates, resulting in an optimal estimator. This approach effectively accounts for the missing data and allows for more precise and efficient estimation of the means of common items across the various samples. Let, $\widehat{X_A}$ and $\widehat{X_B}$ be unbiased estimators of X from sample A and sample B, respectively. Let, $\widehat{Y_B}$ be an unbiased estimator of y from sample B.

To combine the multiple estimates, we can build a linear model of three estimates with two parameters as follows:

$$\begin{bmatrix} \widehat{X_A} \\ \widehat{X_B} \\ \widehat{Y_B} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

where, $\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$ has mean $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ and dispersion matrix given below:

$$\mathbf{V} = \begin{bmatrix} var(\widehat{X_A}) & cov(\widehat{X_A}, \widehat{X_B}) & cov(\widehat{X_A}, \widehat{Y_B}) \\ . & Var(\widehat{X_B}) & cov(\widehat{Y_B}, \widehat{X_B}) \\ . & . & var(\widehat{Y_B}) \end{bmatrix}$$

and *var(·)* and *cov(·)* are the variance and covariance induced by the sampling probability.

3.2 Micro approach: mass imputation

Mass imputation (also called synthetic data imputation) is a technique of creating imputed values for items not observed in the current survey by incorporating information from other surveys. Breidt et al. (1996) discussed mass imputation for two-phase sampling. Rivers (2007) proposed a mass imputation approach using nearest-neighbor imputation, but the theory is not fully developed. Schenker and Raghunathan (2007) reported several applications of synthetic data imputation, using a model-based method to estimate totals and other parameters associated with variables not observed in a larger survey but observed in a much smaller survey. Legg and Fuller (2009) and Kim and Rao (2012) developed

synthetic imputation approaches to combining two surveys. Chipperfield et al. (2012) discussed composite estimation when one of the surveys is mass imputed. Bethlehem (2016) discussed practical issues in sample matching for mass imputation.

The primary goal is to create a single synthetic dataset of proxy values $\hat{y}_i$ for the unobserved y_i in sample B and then use the proxy data together with the associated design weights of sample A to produce projection estimators of the population mean μ_y. This is particularly useful when sample B is a large-scale survey and item Y is very expensive to measure. The proxy values $\hat{y}_i$ are generated by first fitting a working model relating Y to X, $E(Y|X) = m(X;\beta_0)$ based on the data $\{(x_i, y_i) : i \in A\}$ from sample A. Then, the synthetic values of Y can be created by $\hat{y}_i = m(x_i;\hat{\beta})$ for $i \in B$. Thus, sample A is used as a training sample for predicting Y in sample B. The mass imputation estimator of μ_y is $\hat{\mu}_I = N^{-1}\sum_{i\in \mathrm{B}} d_{B,i}\hat{y}_i$.Kim and Rao (2012) showed that $\hat{\mu}_I$ is asymptotically design-unbiased if $\hat{\beta}$ satisfies:

$$\sum_{i\in A} d_{A,i}\{y_i - m(x_i;\hat{\beta})\} = 0$$

$$\hat{\mu}_I = N^{-1}\sum_{i\in \mathrm{B}} d_{B,i}\hat{y}_i + N^{-1}\sum_{i\in A} d_{A,i}\,(y_i - \hat{y}_i)$$

$$\hat{\mu}_I = N^{-1}\sum_{i\in \mathrm{B}} d_{B,i}m(x_i;\beta_0) + N^{-1}\sum_{i\in A} d_{A,i}\left(y_i - m(x_i;\beta_0)\right) = \widehat{P_B} + \widehat{Q_A}$$

and

$$var(\hat{\mu}_I) = var\left(\widehat{P_B}\right) + var(\widehat{Q_A})$$

The asymptotic unbiasedness holds regardless of whether the regression model is true or not. However, a good regression model will reduce the variance of $\hat{\mu}_I$. For variance estimation, either linearization or replication based sampling (Kim and Rao 2012) can be used.

4.0 Combining probability and Non-Probability Samples:

Baker et al. (2013) have extensively documented the challenges faced in the statistical analysis of non-probability survey samples. Such samples are characterized by unknown selection/inclusion mechanisms, inherent biases, and their inability to accurately represent the target population. To address the issues posed by biased non-probability samples, a popular approach involves assuming the availability of auxiliary variable information from an existing probability survey sample covering the same population. This framework, first introduced by Rivers (2007) and subsequently adopted by other researchers such as Vavreck and Rivers (2008), Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017), Chen et al. (2018), and others, has proven to be effective in dealing with the limitations of non-probability samples. The process of combining the most recent information from a non-probability sample with auxiliary information from a probability sample can be viewed as data integration, which is an emerging area of research in the field of survey sampling (Lohr and Raghunathan 2017). This integration of data from multiple sources helps improve the overall quality of estimations and enhances our ability to draw more reliable inferences in the context of non-probability survey samples.

Data integration for finite population inference shares similarities with the challenge of combining randomized experiments and non-randomized real-world evidence studies in clinical trials. In randomized clinical trials, the treatment assignment mechanism is known,

ensuring that treatment effect evaluation is unconfounded. However, due to restrictive inclusion and exclusion criteria, the trial sample may not fully represent the real-world patient population. On the other hand, real-world evidence studies collected through non-randomized data collection mechanisms are often more representative of the target population. By combining information from both trial and real-world evidence studies, more robust and efficient inference of treatment effects can be achieved for the target patient population. Table 2 presents a parallel comparison of data sources between data integration in survey sampling and treatment effect evaluation.

Table 2 Data integration in survey sampling and biostatistics

Survey sampling	Treatment effect evaluation	Representative of the finite population	Unbiased estimation[a]
Probability sample	Real-world evidence study	✓	
Non-probability sample	Randomized experiment		✓

[a]*In survey sampling, some probability samples may not observe the study variable of interest; for treatment effect evaluation, randomized experiments provide unbiased estimation of treatment effect due to treatment randomization.*

Survey statisticians and biostatisticians have developed various methods for combining information from multiple data sources. In the context of finite population inference, Lohr and Raghunathan (2017) and Rao (2020) provide comprehensive reviews of statistical methods. In biostatistics, meta-analysis has been a longstanding method to synthesize evidence from multiple trial and observational data, accommodating heterogeneity in treatment effects estimated from different sources (Verde and Ohmann, 2015). Existing methods for data integration of a probability sample and a non-probability sample can be categorized into three types:

- **Propensity Score Adjustment (Rosenbaum and Rubin, 1983):** This approach models and estimates the probability of a unit being selected into the non-probability sample, known as the propensity or sampling score, for all units in the non-probability sample. Adjustments, such as propensity score weighting or stratification, are then used to account for selection biases. However, these methods can be biased and highly variable if the propensity score model is mis-specified.
- **Calibration Weighting (Deville and Särndal, 1992; Kott, 2006):** This approach applies calibration weights to align the characteristics of the non-probability sample with those of the probability sample, effectively adjusting for selection biases. This technique calibrates auxiliary information in the non-probability sample with that in the probability sample, so that after calibration, the weighted distribution of the non-probability sample is similar to that of the target population.

The third type is mass imputation, which imputes the missing values for all units in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample constitute a training dataset for developing an imputation model. In the mass imputation, an independent nonprobability sample is used as a training dataset, and imputation is applied to all units in the probability sample.

5.0 Integration of Agricultural Surveys:

Agricultural surveys play a crucial role in understanding and monitoring the agricultural sector in India. These surveys provide valuable information about crop production, land use, agricultural practices, and other relevant factors. The Agricultural Census is conducted by the Ministry of Agriculture and Farmers Welfare once every five years. It collects comprehensive data on the structure of agriculture, including information on operational holdings, land use, irrigation, livestock, machinery, and agricultural practices. The integration of surveys refers to the process of combining data from multiple surveys into a single dataset. This can be done for various reasons, such as improving the accuracy and reliability of the data or increasing the sample size to make more robust statistical inferences.

There are several methods for integrating surveys, including:

- **Data Harmonization:** Harmonization is the process of making survey data comparable across different surveys or time periods by ensuring that the questions, response options, and coding schemes are consistent. This approach is commonly used in cross-national or longitudinal surveys.
- **Weighting:** Survey weighting is a statistical technique that adjusts the results of a survey to make it more representative of the target population. This can involve assigning different weights to different survey respondents based on their demographic characteristics or other factors.
- **Data Fusion:** Data fusion involves combining survey data with other sources of data, such as administrative records or social media data, to create a more comprehensive dataset. This approach can be useful for improving the accuracy and completeness of survey data or for filling in missing data.
- **Meta-Analysis:** Meta-analysis is a statistical technique that combines the results of multiple studies or surveys to provide a more accurate estimate of an effect size. This approach is commonly used in systematic reviews of the literature.

Overall, integrating surveys can be a powerful tool for improving the quality and usefulness of survey data. However, it requires careful planning, analysis, and interpretation to ensure that the results are valid and reliable. The integration of agricultural surveys involves the process of combining and harmonizing data collected from various surveys to provide a comprehensive and unified picture of agricultural activities, production, and related factors. It aims to eliminate duplication, improve data quality, and enhance data usability for analysis and decision-making in the agricultural sector.

Key Steps and Considerations in the Integration of Agricultural Surveys:

There are several significant points that are to be considered before embarking on integration of agricultural surveys the major points have been point wise enumerated below:

1) **Survey Design and Planning:** Proper survey design is crucial to ensure compatibility and harmonization. Surveys should be designed with consistent methodologies, definitions, and data collection tools to facilitate integration.
2) **Standardization:** It is essential to standardize data elements, classifications, and units across surveys to enable meaningful integration. This involves aligning survey questions, response categories, and measurement units to ensure compatibility.

3) **Data Cleaning and Validation:** Before integration, data from individual surveys need to be cleaned, validated, and standardized. This process involves identifying and resolving data inconsistencies, missing values, outliers, and other data quality issues.
4) **Data Harmonization:** Agricultural surveys may cover various aspects, such as crop production, livestock, land use, and agricultural practices. Integrating these diverse data sources requires harmonization, where common variables are identified, and data is transformed or aggregated to a common format.
5) **Statistical Analysis and Integration:** Statistical techniques are employed to combine data from multiple surveys. This may involve weighting the data to account for different sample sizes or developing statistical models to impute missing values or estimate aggregated measures.
6) **Metadata Documentation:** Detailed documentation of survey methodologies, data sources, transformations, and integration procedures is crucial for transparency and reproducibility. Metadata should be recorded to provide information about the integrated dataset's structure, limitations, and data sources.
7) **Data Dissemination:** The integrated dataset should be made accessible to researchers, policymakers, and other stakeholders. This can be achieved through data portals, online platforms, or specialized agricultural databases. Proper data privacy and security measures must be implemented to protect sensitive information.

Benefits of Integrating Agricultural Surveys:

There are several fold benefits of integration of data in agricultural surveys from more precise estimates to greater information content which would eventually help in formulating better agricultural policies. The major benefits are enumerated as follows:

a) **Comprehensive Information:** Integration enables a holistic view of agricultural activities, allowing policymakers and researchers to analyze and understand the sector's dynamics more accurately.
b) **Data Quality Improvement:** By standardizing and harmonizing data, integration helps reduce errors, inconsistencies, and redundancies, resulting in improved data quality.
c) **Enhanced Analysis:** Integrated datasets provide a richer source for in-depth analysis and modeling, enabling better insights into trends, patterns, and relationships within the agricultural sector.
d) **Resource Optimization:** Integration reduces duplication of efforts and resources by leveraging existing survey data, thereby optimizing survey costs and reducing respondent burden.
e) **Policy Development and Decision-making:** Integrated agricultural surveys provide policymakers with a robust evidence base for formulating effective agricultural policies and making informed decisions.

The integration of agricultural surveys is a valuable process that enhances the usability and reliability of agricultural data, facilitating evidence-based decision-making and policy formulation in the agricultural sector.

6.0 Conclusion:

Data integration is an emerging area of research with many potential research topics. Probability sampling remains as the gold standard to obtain a representative sample, but

the measurement of the study variable can be obtained from an independent non-probability sample or big data. In this case, assumptions about the sampling model or the outcome model are required. Most data integration methods are based on the unverifiable assumption that the sampling mechanism for the non-probability sample (or big data) is non-informative (corresponding to the missingness at random in the missing data literature). If the sampling mechanism is informative, imputation techniques can be developed under the strong model assumptions for the sampling mechanism. Like the non-informative sampling case, the informative sampling assumption is unverifiable. In such settings, sensitivity analysis is recommended to assess the robustness of the study conclusions to unverifiable assumptions.

REFERENCES:

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley

Fuller, W. A. (2009). *Sampling statistics*. Hoboken: Wiley.

Särndal, C.E., Swensson, B. & Wretman, J. (2003). *Model assisted survey sampling*. New York: Springer-Verlag.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*, 90–143.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, *87*, S10–S30.

Keiding, N. & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A*, *179*, 319–376.

Bycroft, C. (2010). *Integrated household surveys: A survey vehicles approach*. Wellington: Statistics New Zealand.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101–116.

Deming, W. E. & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*, 427–444.

Wang, W., Scharfstein, D., Tan, Z. & MacKenzie, E. J. (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B*, *71*, 947–969.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, *85*, 986–1001.

Hidiroglou, M. (2001). Double sampling. *Survey Methodology*, *27*, 143–54.

Renssen, R. H. & Nieuwenbroek, N. (1997). Aligning estimates for common variables in two or more sample surveys. *The Journal of the American Statistical Association*, *92*, 368–75.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *The Journal of the American Statistical Association*, *99*, 1131–9.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, *72*, 27–48.

Lohr, S. L. & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*, 293–312.

Rivers, D. (2007). *Sampling for web surveys, ASA proceedings of the section on survey research methods*. Alexandria: American Statistical Association.

Ybarra, L. & Lohr, S. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*, 919–31.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, *34*, 59–77.

Breidt, F. J., McVey, A. & Fuller, W. A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, *49*, 79–90.

Chipperfield, J., Chessman, J. & Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *The Australian and New Zealand Journal of Statistics*, *54*, 223–238.

Verde, P. E. & Ohmann, C. (2015). Combining randomized and non-randomized evidence in clinical research: A review of methods and applications. *Research Synthesis Methods*, *6*, 45–62.

Rao, J.N.K. (2020). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*. https:// doi. org/ 10. 1007/ s13571- 020- 00227-w

Kim, J. K., Park, S., Chen, Y. & Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation, arxiv. org/ abs/ 1812. 10694 .

Kim, J. K. & Rao, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, *99*, 85–100.

Kim, J. K. & Tam, S. (2018). Data integration by combining big data and survey sample data for finite population inference. https:// arxiv. org/ abs/ 2003. 12156

Rosenbaum, P.R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*, 133–142.

Deville, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382

MODEL-BASED AND MODEL-ASSISTED APPROACHES IN SURVEY SAMPLING

Rahul Banerjee, Ankur Biswas, Kaustav Aditya and Tauqueer Ahmad

ICAR-Indian Agricultural Statistics Research Institute,

Library Avenue, New Delhi-110 012

1.0 Introduction:

Survey sampling is a crucial field in statistics that focuses on gathering information from a subset of a population to estimate characteristics of the entire population. Traditional survey sampling relies on the design-based approach, where inference is based on the sampling design. This approach uses the stochastic structure induced by the sampling design. The randomness is incorporated via the selection of population units according to a sampling design. However, when estimating the total of a real-valued variable for a population, the census is the only sampling design that guarantees a uniformly best unbiased estimator. For any other sampling design, a uniformly best unbiased estimator doesn't exist. Even among linear unbiased estimators, only a very limited type of sampling design (unicluster) allows for a best estimator; non-unicluster designs do not. However, for any sampling design, a "complete class" of unbiased estimators can be identified, where estimators within this class are superior to those outside it. Additionally, a "minimal sufficient statistic" can be easily derived from the raw survey data, suggesting that any useful estimator should be a function of this statistic. Despite these theoretical advancements, practically defining and applying exclusive classes of estimators remains challenging. To address the practical difficulties in finding optimal estimators, the concept of a "super-population" and its modeling has emerged as a promising approach. This approach is referred to as the Model based approach of survey sampling. In contrast, the model-based approach assumes that the population is a random realization of a data-generating stochastic process. This approach incorporates randomness through distributional assumptions on this process and may not always rely on random sampling.

2.0 Super Population Modelling:

Estimating population totals for an arbitrary vector of real values $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]$ is challenging when seeking optimal unbiased estimators. A proposed solution involves treating $\mathbf{Y}$ as a random vector with a probability distribution. The population associated with this probability distribution is termed the "super-population," distinct from the finite survey population U={1, 2,…,i,…,N}. Crucially, the exact form of this distribution doesn't need to be specified. A general class of probability distributions with finite low-order moments referred to as a "modelled population" is sufficient for deriving optimal estimation procedures. This approach of considering a class of hypothetical super-populations is known as super-population modelling, and such a class of probability distributions constitutes a "super-population model."

2.1 Model Based Approach:

2.1.1 Concept

In the model-based approach, the population is considered to be generated by an underlying probabilistic model. The sample is treated as a realization from this model, and inference is based on the assumed stochastic process.

2.1.2. Key Assumptions

- A superpopulation model describes the relationship between the response variable and auxiliary information.
- The sample is treated as a realization from the assumed model.
- Estimators are derived using model parameters rather than design probabilities.

Model-based inferences for survey sampling are built on the concept of a *superpopulation model*, was first introduced by Cochran (1939). Suppose that the finite population values $\{\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]$ can be viewed as a random sample from a superpopulation model, ξ. Under the model, the y_i's are generally viewed as random variables with probability distributions specified by the assumed model.

Model I: The common mean model for the survey population:

$$y_i = \mu + \epsilon_i; i = 1,2, \dots, N$$

where $\varepsilon_1, \dots, \varepsilon_N$ are independent and identically distributed error terms with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma^2$. Here E_ξ and V_ξ refer to expectation and variance under the model, ξ. The μ and σ^2 are superpopulation parameters, which are conceptually different from the finite population parameters μ_y and σ_y^2.

Model II: The simple linear regression model without an intercept:

$$y_i = \beta x_i + \epsilon_i; i = 1,2, \dots, N$$

where $\varepsilon_1, \dots, \varepsilon_N$ are independent error terms with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i\sigma^2$. For any i, the v_i is a known constant often depending on x_i. The β and σ^2 are superpopulation parameters.

Model III: The simple linear regression model with an intercept:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; i = 1,2, \dots, N$$

where $\varepsilon_1, \dots, \varepsilon_N$ are independent error terms with $E_\xi(\varepsilon_i)=0$ and $V_\xi(\varepsilon_i)=v_i\sigma^2$. The superpopulation parameters in this case are β_0, β_1 and σ^2.

Within the framework of a super-population model, the finite population mean μ_y is treated as a random variable because it's defined in terms of the random values $y_1, y_2, \dots, y_N$. Estimating μ_y under this model becomes a prediction problem. We assume the super-

population model ξ, which is assumed to apply to the entire survey population, also holds for any sample S drawn from that population, irrespective of the sampling method. When this condition is met, the probability sampling design is considered "ignorable" for the model-based prediction approach.

Let $\widehat{\mu_y}$ be an estimator of μ_y, computed based on the survey sample data and the available auxiliary information. We term $\widehat{\mu_y}$ a model-unbiased prediction estimator of μ_y if:

$$E_\xi\left(\widehat{\mu_y} - \mu_y\right) = 0$$

Clearly, constructing a model-based prediction estimator relies on both the chosen super-population model and the available auxiliary information. The unbiasedness of such an estimator must be assessed within the context of the assumed model.

3.0 Model-Assisted Estimation Methods:

While model-based prediction using accurate models can be highly efficient, misspecification can lead to disastrous results. In contrast, design-based survey sampling inferences make no model assumptions, and the sampling design is chosen by the surveyor based on the specific population. Large-sample confidence intervals in design-based inference are often asymptotically valid regardless of the population's characteristics.

The advantages of modeling can be incorporated into design-based inferences through a model-assisted approach. This approach uses a plausible model to guide estimator construction, but evaluates the estimator under both model-based and design-based frameworks. An estimator is considered model-assisted if: It is a model-unbiased (or approximately model-unbiased) prediction estimator under the assumed model.

It is approximately design-unbiased (or design-consistent) under the probability sampling design, regardless of the model.

Design-consistency, a stronger condition than approximate design-unbiasedness, requires the estimator to converge in probability to the parameter of interest under the sampling design. This typically implies that the design-based variance approaches zero as the sample size increases, assuming the parameter being estimated is of a constant order.

Model-assisted estimators generally require that certain auxiliary population information is available. For the rest of this section, we assume that the sample **S** is selected by SRSWOR and the sample data are given by $\{(y_i, x_i), i \in S\}$, where x is a single auxiliary variable. In addition, the population mean μ_x is available as the known auxiliary information.

3.1. Advantages of Model-Assisted Approach

- **Robust to model misspecification**: Even if the model is incorrect, estimates remain approximately unbiased.
- **Improved efficiency**: Utilizes auxiliary information without being solely reliant on model assumptions.

- **Retains design-based properties**, making it preferable in official statistics.

3.2. Limitations of Model-Assisted Approach

- Efficiency gain depends on the strength of the auxiliary variable.
- May require additional computations for variance estimation.

4.0 Estimators in Model-Assisted Approach:

Two commonly used estimators under model-assisted survey sampling are:

- **Generalized Ratio Estimator**
- **Generalized Regression (GREG) Estimator**

4.1 Generalized Ratio Estimator:

The ratio estimator is an early model-assisted method that leverages an auxiliary variable X correlated with the variable of interest Y. It assumes a linear model:

$$y_i = \beta x_i + \epsilon_i; i = 1,2, \dots, N, E(\epsilon_i) = 0$$

where β is an unknown proportionality constant. The generalized ratio estimator extends this concept by allowing for probability sampling weights and multiple auxiliary variables.

4.1.1 Estimation Procedure:

Estimating β: $\hat{\beta} = \frac{\sum_{i \epsilon s} w_i Y_i}{\sum_{i \epsilon s} w_i X_i}$

where:

- $w_i = 1/\pi_i$ are the design weights,
- π_i is the inclusion probability for unit iii,
- s is the sample,
- X_i is the auxiliary variable.

4.1.2 Estimator for Population Total Y:

$$\widehat{Y_{GR}} = \hat{\beta} X$$

where $X = \sum_{i=1}^{N} X_i$ is the known population total of the auxiliary variable.

Properties:

- Unbiased if the ratio model holds

- Asymptotically design-unbiased even if the model is not perfect.
- Improves efficiency compared to simple expansion estimators.

4.2 The Generalized Regression Estimator:

Suppose that auxiliary information is available on k covariates $x_1, x_2, \dots, x_k$. Let $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$ be the values of $(y, x_1, x_2, \cdots, x_k)$ attached to unit i. Consider the following superpopulation model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i; i = 1,2, \dots, N$$

Where, the ε_i's are independent with $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = v_i\sigma^2$. The v_i's are known constants depending on $x_i = (x_{i1}, x_{i2}, \cdots, x_{ik})'$.

The most popular model-assisted estimator of μy is the generalized regression (GREG) estimator (Cassel et al. 1976; Särndal 1980). A large body of literature on model-assisted inferences has been developed under the linear model (5.11). The book by Särndal et al. (1992) provides comprehensive coverage of generalized regression estimation under linear regression models with additional references on the topic. Fuller (2002) contains discussions on practical and theoretical aspects of regression estimation for survey samples.

4.2.1 GREG Under SRSWOR:

Suppose that the survey sample **S** of size n is selected by SRSWOR. Let $\{(y_i, \boldsymbol{x_i}); i \in \boldsymbol{S}\}$ be the survey sample data. Let, $\bar{y} = n^{-1}\sum_{i\epsilon S} y_i$ and , $\bar{x} = n^{-1}\sum_{i\epsilon S} x_i$ be the sample means. Let $\widehat{\mathbf{B}} = (\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_k})'$ be the ordinary least squares estimators of β:

$$\widehat{\mathbf{B}} = (\sum_{i\epsilon S} \mathbf{z_i} \mathbf{z_i}')^{-1} \left(\sum_{i\epsilon S} \mathbf{z_i} y_i\right)$$

Under the mean model, we have $E_\xi(\widehat{\mathbf{B}}) = \beta$ by the properties of the least squares estimators. Let,

$$\widehat{\mu_{yGREG}} = \widehat{\mathbf{B}}'\mu_z$$

where the subscript GREG indicates "Generalized Regression". It can be shown that

$$E_\xi(\widehat{\mu_{yGREG}} - \mu_y) = 0$$

the GREG estimator is model-unbiased prediction estimator of μ_y under the mean model.

4.2.2 Properties

- **Design-consistent**, ensuring unbiasedness under repeated sampling.
- **More flexible than the ratio estimator**, allowing for multiple auxiliary variables.

- **Reduces variance**, leading to more precise estimates.

4.2.3 Special Case: Post-Stratified Estimator

If auxiliary variables indicate strata, the GREG estimator reduces to a **post-stratified estimator**, ensuring improved estimation within strata.

5.0 Comparison of Generalized Ratio and GREG Estimators:

Feature	Generalized Ratio Estimator	GREG Estimator
Model Assumption	Linear relationship: $Y_i = \beta X_i + \epsilon_i$	Multiple regression: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \epsilon_i$
Number of Auxiliary Variables	Single X	Multiple $X_1, X_2, \ldots$
Estimator Formula	$\hat{Y}_{\text{GR}} = \hat{\beta} X$	$\hat{Y}_{\text{GREG}} = \sum w_i Y_i + (X - X_s' w)\hat{\beta}$
Efficiency	Moderate	High
Robustness	Limited	High

6.0 Comparison of Model-Based and Model-Assisted Approaches:

Feature	**Model-Based Approach**	**Model-Assisted Approach**
Basis of Inference	Assumed superpopulation model	Design-based principles with model adjustments
Sensitivity to Model	High	Low
Efficiency	High (if model is correct)	Moderate to high
Design Consistency	No	Yes

Applicability	Small area estimation, predictive modeling	Official statistics, large-scale surveys

7.0 Applications and Case Studies:

- **Agricultural Surveys:** Model-assisted estimators like GREG are used in crop yield estimation with remote sensing data (Battese, Harter, & Fuller, 1988).
- **Economic Surveys:** The model-based approach is used in poverty mapping and small area estimation (Rao & Molina, 2015).
- **Energy Audits:** Model-assisted estimation helps in carbon footprint assessments (Sarndal, Swensson, & Wretman, 2003).

8.0 Conclusion:

Both model-based and model-assisted approaches play a crucial role in modern survey sampling, each with its own strengths and limitations. The model-based approach relies entirely on an assumed superpopulation model, making it highly efficient when the model is correctly specified but vulnerable to bias if the assumptions are incorrect. It is particularly useful in small area estimation and predictive modeling, where auxiliary data is abundant, and probability sampling may be infeasible. On the other hand, the model-assisted approach blends design-based principles with model-driven efficiency, ensuring that estimators remain design-consistent while utilizing auxiliary information to reduce variance. The generalized ratio estimator offers a simple improvement over basic expansion estimators when a strong linear relationship exists between the study variable and a single auxiliary variable. Meanwhile, the generalized regression (GREG) estimator extends this concept to multiple auxiliary variables, providing even greater precision and adaptability. These model-assisted methods are widely applied in large-scale agricultural, economic, and environmental surveys, where maintaining design-based robustness while improving efficiency is crucial. Ultimately, the choice between model-based and model-assisted approaches depends on the survey's objectives, data availability, and the acceptable level of reliance on modeling assumptions. As survey methodologies continue to evolve, these approaches will remain fundamental in enhancing the accuracy and reliability of survey estimates.

References:

1. Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association, 83*(401), 28-36.

2. Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation.* Wiley.

3. Sarndal, C. E., Swensson, B., & Wretman, J. (2003). *Model-Assisted Survey Sampling.* Springer.

4. Cassel, C. M., Sarndal, C. E., & Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika, 63*(3), 615-620.

5. Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley.

DEVELOPMENT OF INDICES WITH SURVEY DATA AND APPLICATIONS

Deepak Singh

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Indicators are helpful in recognizing patterns and attracting consideration regarding specific issues. A composite indicator is framed when singular indicators are assembled into a single index on the premise of a basic model. Composite indicators/index are much similar to mathematical or computational models which should ideally measure multi-dimensional concepts, which can't be caught by a single indicator alone. Therefore, in most of the situations, composite index are based on simple or weighted average method which does not consider the effect of multicollinearity among the indicator variables that are used for index construction. Principal component (PC) based index accounts for the effect of multicollinearity among the indicator variables through the eigen values and eigen vectors derived from the variance-covariance matrix using maximum likelihood (ML)/ordinary least squares (OLS) methods of estimation. However, these methods of estimation of variance covariance matrix are based on the assumption that sample elements, on which the indicator variables are measured, are independent and identically distributed. This assumption of independence holds good if the data are collected through simple random sampling with replacement. However, it does not hold good for others sampling schemes. Now a day, most of the survey designs are complex in nature involving stratification, unequal probabilities of selection, clustering, multi-stages, multi-phases and auxiliary information. In case of large scale surveys, stratified multistage sampling design is widely used. Here, also the units in a stratum are relatively homogenous which violates the assumption of independence of sample elements. Any deviation from independence assumption leads to erroneous estimation of variance covariance matrix which in turn leads to erroneous estimation of eigenvalues and eigenvectors, and thereby resulting in poor PC based index. Therefore, in case of survey data there is a need to develop PC based index using survey weights and auxiliary information which excludes the effect of multicollinearity among the indicator variables as well as accounts for the effect of complex survey designs through which the sample data is collected.

2. Methodology

In the principal component analysis, the principal components are the linear functions of social indicators maintaining orthogonality between each other (Singh, D. (2009)). The principal components are derived from variance-covariance matrix $\mathbf{\Sigma}_{yy.}$ In the development of index the survey weights and prior auxiliary information is incorporated in the variance covariance matrix $\mathbf{\Sigma}_{yy}$ (function of population variance-covariance matrix of survey variables $\mathbf{\Sigma}_{yy}$) from which the indices are developed using PCA approach.

Let s be a probabilistic sample of size n drawn from a finite population $U = (1, 2, \ldots, i, \ldots, N)$ having l subpopulations/blocks/states such that the h^{th} subpopulation has N_h units such that $\sum_{h=1}^{l} N_h = N$ *(h= 1, 2… l)*. Let s_h is the probabilistic subsample of size n_h from h^{th} sub-population such that $\sum_{h=1}^{l} n_h = n$ and d_{ih} denotes the survey weight associated with i^{th} unit of the sample present in h^{th} subpopulation. The data of the sampling units is selected from all the subpopulations/blocks/states among which the indices/ranking of indices is needed to be developed.

let $\mathbf{y} = \left(y_{1,} y_{2,} ..., y_p\right)'$ and $\mathbf{x} = \left(x_{1,} x_{2,} ..., x_q\right)'$ be the p and q set of standardised indicators and auxiliary variables respectively for a finite population $U= (1, 2,...,i,...,N)$. Let $\mathbf{x}$ is assumed to be known for each unit of population, $\boldsymbol{\Sigma}_{xx}$ (variance covariance matrix of auxiliary variables) is positive definite $\forall\ i \in s,\ h = 1,...l,\ v = 1,..., p,\ w = 1,..., q$. Here, parameter of interest is variance-covariance matrix of indicator variables $\mathbf{y}$ i,e.

Estimators of $\boldsymbol{\Sigma}_{yy}$

1) OLS estimator

$$\hat{\Sigma}_{yy} = \boldsymbol{V}_{\boldsymbol{yys}} = (n-1)^{-1} \sum_{i=1}^{n} \left(\boldsymbol{y}_{ih} - \bar{\boldsymbol{y}}_{sh}\right)\left(\boldsymbol{y}_{ih} - \bar{\boldsymbol{y}}_{sh}\right)' \tag{1}$$

where, $\bar{\mathbf{y}}_S = \sum_{i=1}^{n} \mathbf{y}_{ih} / n$

2) Survey weighted estimator (Smith & Holmes, 1989)

$$\hat{\Sigma}_{yyw} = \mathbf{V}^*_{\mathbf{yys}} = \sum_{i=1}^{n} d_{ih} \mathbf{y}_{ih} \mathbf{y}'_{ih} - \bar{\mathbf{y}}^*_{sh} \bar{\mathbf{y}}^{*\prime}_{sh} \Big/ \sum_{i=1}^{n} d_{ih} \tag{2}$$

where, $\bar{\mathbf{y}}^*_{sh} = \sum_{i=1}^{n} d_{ih} \mathbf{y}_{ih}$

3) Un-weighted regression estimator (Smith & Holmes, 1989)

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yyr}} = \mathbf{V}_{\mathbf{yys}} + \mathbf{b}_{\boldsymbol{yx}} \left(\boldsymbol{\Sigma}_{\boldsymbol{xx}} - \boldsymbol{V}_{\boldsymbol{xxs}}\right) \mathbf{b}'_{\boldsymbol{yx}} \tag{3}$$

where,

$$\mathbf{b}_{yx} = \mathbf{V}_{xys} \mathbf{V}^{-1}_{xxs},$$

$$\mathbf{V}_{xxs} = (n-1)^{-1} \sum_{i=1}^{n} \left(\mathbf{x}_{ih} - \bar{\mathbf{x}}_{sh}\right)\left(\mathbf{x}_{ih} - \bar{\mathbf{x}}_{sh}\right)',$$

$$\mathbf{V}_{xys} = (n-1)^{-1} \sum_{i=1}^{n} \left(\mathbf{x}_{ih} - \bar{\mathbf{x}}_{sh}\right)\left(\mathbf{y}_{ih} - \bar{\mathbf{y}}_{sh}\right)'$$

4) Survey weighted regression estimator (Smith & Holmes, 1989)

$$\hat{\Sigma}^*_{yywr} = \mathbf{V}^*_{\mathbf{yys}} + \mathbf{b}^*_{yx} \left(\Sigma_{xx} - \mathbf{V}^*_{xxs}\right) \mathbf{b}^{*\prime}_{yx} \tag{4}$$

where,

$$\mathbf{b}^*_{yx} = \mathbf{V}^*_{xys} \mathbf{V}^{*-1}_{xxs},\ \mathbf{V}^*_{\mathbf{xys}} = \sum_{i=1}^{n} d_{ih} \mathbf{x}_{ih} \mathbf{y}_{ih}{}' - \bar{\mathbf{x}}^*_{sh} \bar{\mathbf{y}}^*_{sh} \Big/ \sum_{i=1}^{n} d_{ih},$$

$$\mathbf{V}^*_{\mathbf{xxs}} = \sum_{i=1}^{n} d_{ih} \mathbf{x}_{ih} \mathbf{x}_{ih}{}' - \bar{\mathbf{x}}^*_{sh} \bar{\mathbf{x}}^{*\prime}_{sh} \Big/ \sum_{i=1}^{n} d_{ih}$$

Skinner et al. (1986) and Smith & Holmes (1989) discussed the Survey weighted estimator $\hat{\Sigma}_{yyw}$ and Un-weighted regression estimator $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yyr}}$ while all the estimators mentioned above of variance covariance matrix have been mentioned by Smith & Holmes (1989).

PCA method for index development for disaggregated data

For the development of PCA based indices the principal components are needed to be extracted for which the variance covariance matrix $\underset{\sim}{\Sigma}$ is the target parameter. For normal,

survey weighted, un-weighted regression, survey weighted regression PCA based indices the $\hat{\Sigma}_{yy}$, $\hat{\Sigma}_{yyw}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yyr}}$ and $\hat{\Sigma}^{*}_{yywr}$ estimators of $\underset{\sim}{\Sigma}$ are used respectively. For the extraction of PCA based indices from estimators of $\underset{\sim}{\Sigma}$, let us assume that the $\underset{\sim}{\Sigma}$ is a real positive definite matrix having distinct eigen values $\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_p$. The eigenvectors $\gamma_1, \gamma_2, \gamma_3 \ldots\ldots \gamma_p$ to the corresponding eigenvalues are defined by

$$\underset{\sim}{\Sigma}\gamma_j = \lambda_j \gamma_j, \quad j=1,\ldots,p$$

Such that

$$\gamma_j'\gamma_k = 1, \quad j \neq p$$
$$= 0 \quad \text{otherwise},$$

The principal components (P vector) is the linear combination of standardised indicators

$$\underline{P} = \gamma' Y$$

The composite index or sub-indices for each sampled unit $\left(C_{ih}\right) \quad \forall\, i \in s,\ h \in l$ is constructed by using the obtained eigen values of variables and principal components as

$$C_{ih} = \frac{\left[\sum_{j=1}^{p} \lambda_j PC_{ihj}\right]}{\sum_{j=1}^{p} \lambda_j} \tag{5}$$

where λ_js are eigen values and PC_{ihj}'s are principal component scores for $\forall\, i \in s,\ h \in l, j = 1,\ldots, p.$

The average of $C_{ih}'s$ within h^{th} subpopulation gives the composite index value for h^{th} subpopulation as

$$C_h = \sum_{i=1}^{n_h} C_{ih} \Big/ n_h \tag{6}$$

(i) Rescaling of Composite Index

The composite index values $\left(C_h\right)$ of subpopulations are rescaled by using the following formula as

$$CI_h = \frac{C_h - \min(C_h)}{\max(C_h) - \min(C_h)} \tag{7}$$

The ranking of l subpopulations is done based on the composite index values $\left(CI_h\right)$. The subpopulation having CI_h value equal to one is ranked as first while the rank is last for the subpopulation having value equal to zero.

In our study, we have proposed the indices based on weighted estimator, un-weighted regression estimator and survey weighted regression estimator. The survey weighted principal component index is proposed when the data is collected through complex survey designs in which the inclusion probability is not same for each unit. The second index is proposed when the auxiliary information is available under which un-weighted regression estimator can be utilised to develop index while the third index is proposed under complex survey designs and the auxiliary information is also available. All the 3 developed indices were compared to normal PCA index which uses ordinary least squares estimator $\hat{\Sigma}_{yy}$.

2.3 Steps for the development of Survey weighted estimator, Un-weighted regression estimator and Survey-weighted regression estimator based indices

Step 1: Data standardization for making data unit free.

$$\left(Y_i - \overline{Y}\right) \Big/ S_i = Z_i$$

Step 2: Calculation of Survey weighted, regression e and Survey-weighted regression estimators of correlation/Covariance matrix.

Step 3. Extraction of principal components from the above developed estimators of correlation/Covariance matrix

***Step 4*:** The composite index or sub-indices for each sampled unit is constructed by using the obtained eigen values of variables and principal component scores as given below:

$$C_{hi} = \frac{\left[\sum_{j=1}^{p} \lambda_j PC_{hij}\right]}{\sum_{j=1}^{p} \lambda_j}$$

where λ_j's are eigen values and the PC_{hij}'s are principal component scores corresponding to the sample unit $\forall\ h= 1, 2,\ldots, l;\ i = 1,2, \ldots, n_h; j =1, 2,\ldots,p$.

***Step 5*:** Calculating the composite index value for each subpopulation by averaging the C_{hi} values of sampled units within the subpopulation i.e. $C_h = \sum_{i=1}^{n_h} C_{hi} \Big/ n_h$.

***Step 6*:** The composite index for each subpopulation were rescaled by using given formula.

$$CI_h = \frac{C_h - \min(C_h)}{\max(C_h) - \min(C_h)}$$

Step 7: The ranking of *l* subpopulations is done based on the composite index values $\left(CI_h\right)$. All the composite index values $\left(CI_h\right)$ lie between 0 and 1 where one value denotes highest rank and zero denotes the lowest rank.

3. Measures of evaluation

Developed indices were evaluated on the percentage Relative Root Mean Squared Error (%RRMSE) of any estimator of the population parameter θ as given b

$$RRMSE(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^{S}\left[\sum_{i=1}^{k}\left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)^2\right]}{S}} * 100 \qquad (8)$$

where, $\hat{\theta}_i$ are the index value (estimator) of population parameter θ_i of i^{th} district in population (state in our case) and S denotes the number of sample replications (i.e., S = 5000).

4. Simulation study on Artificially generated data

For the simulation study, an artificial population from multivariate normal distribution $\mathbf{X} = \left(\mathbf{Y}', Z'\right)'$ satisfying the linearity and homoscedasticity assumptions was generated having mean vector $\mathbf{u}_x$ and $\sum_{xx}$ as variance covariance matrix . The vector of *Y* variables comprised the twelve variables and the single design variable *Z* which is the sum of all the twelve variables. The mean vector $\mathbf{u}_x$ and variance covariance $\sum_{xx}$ matrix were chosen to be those estimated for the all India on the twelve set of variables from NSS 68th round household consumption expenditure data i.e. Cereals (Z1), Pulses & pulse products (Z2),

Milk & milk products (Z3), Salt & sugar (Z4), Edible oil (Z5), Egg, fish & meat (Z6), Vegetables (Z7), Fruits (fresh) (Z8), Spices (Z9), Beverages (Z10), Served processed food (Z11) and Packaged processed food (Z12). A finite population of one lakh units were generated independently from the above mentioned thirteen dimensional multivariate normal distribution of *X.* Based on the ordered z-values of the design variable *Z*, the population was stratified into five equal strata, each strata having equal units. The simulation was done using five thousand number of sample replications i.e., S = 5000.

Table 1. Mean of variables considered for population data generation

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
807.93	213.31	721.86	127.64	65.13	166.69	120.73	70.09	58.70	49.10	97.85	56.63

Table 2. Variance covariance matrix of variables considered for population generation

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
Y1	263931	34252	76232	19551	11007	26131	19112	6828	6776	6280	3673	7081
Y2	34252	22964	41505	9415	4393	4090	4918	3003	2995	2189	1594	2829
Y3	76232	41505	620633	54028	14295	8103	18088	16563	8402	11975	10450	15495
Y4	19551	9415	54028	14862	3476	2290	3434	1884	1883	1560	896	1835
Y5	11007	4393	14295	3476	3891	3192	2433	973	1508	982	555	1601
Y6	26131	4090	8103	2290	3192	28266	4843	3285	2466	2187	3389	2588
Y7	19112	4918	18088	3434	2433	4843	6531	1726	1520	1361	924	1772
Y8	6828	3003	16563	1884	973	3285	1726	4607	887	1316	1763	1731
Y9	6776	2995	8402	1883	1508	2466	1520	887	1852	717	733	890
Y10	6280	2189	11975	1560	982	2187	1361	1316	717	2733	1582	1184
Y11	3673	1594	10450	896	555	3389	924	1763	733	1582	101389	2261
Y12	7081	2829	15495	1835	1601	2588	1772	1731	890	1184	2261	6203

Samples of size 2000 were selected from this stratified population with the following sample sizes in each strata.

Table 3: Stratum sample sizes

Strata	1	2	3	4	5
Stratum Sample sizes	600	300	200	300	600

The table 4 represents the performance of indices obtained from simulation study (i) that has been carried out. It presents the Relative Root Mean Squared Error (% RRMSE) of all the proposed indices methods for comparison among indices.
From the table it is clear that the survey weighted regression estimator based index method, which utilises the auxiliary variable as well as survey weights available through survey design, performs best followed by regression estimator based index method, survey weighted estimator and least by unweighted estimator PCA based index method which concludes that the proposed indices perform better in comparison to existing PCA based index method which utilises unweighted variance covariance matrix.

Table 4. % RRMSE of all the proposed indices

S.no.	Unweighted UNWTD-PCA	Survey Weighted SW-PCA	Unweighted regression estimator UNWTD-REG- PCA	survey weighted regression estimator SW- REG- PCA
1	22.84	19.46	19.20	19.03

It is clear that the survey weighted PCA based index method, the regression estimator based index which utilises the auxiliary variable and survey weighted regression estimator which utilises both the auxiliary variable and survey weights are performing better in comparison to traditional PCA based index method. The survey weighted regression estimator based index method, which utilises the auxiliary variable as well as survey weights available through survey design, performs best among all the indices.

References:

Athreya, V. B., Rukmani, R., Bhavani, R. V., Anuradha, G., , Gopinath, R. Velan S. A. (2010). Report on the state of food in security in urban india. Published by M. S. Swaminathan Research Foundation. ISBN: 978-81-88355-21-1.

Human Development Report 2016. (hdr.undp.org/sites/default/files/2016_human_development_report. pdf).

Klein, L. R. and Ozmucur, S. (2002/2003) The estimation of China's economic growth. Journal of Economic and Social Measurements, 62(8): 187-202.

Kumar, M. (2008). A study on development indices and their sensitivity analysis. M.Sc. Thesis, ICAR-IARI.

Narain, P., Rai, S.C. and Sarup, S. (1991). Statistical evaluation on development on socio-economic front. *Jour. Ind. Soc. Ag. Statistics.,* **43(3)**, 329-345.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2004). Estimation of socio-economic development in hilly States. *Jour. Ind. Soc. Ag. Statistics.,* **58(1)**, 126-135.

Narain, P., Sharma, S.D., Rai, S.C. and Bhatia, V.K. (2005). Dimension of socio-economic development in Jammu and Kashmir. *Jour. Ind. Soc. Ag. Statistics.,* **59(3)**, 243-250.

Narain, P., Sharma, S.D., Rai, S.C., and Bhatia, V.K. (2003). Evaluation of economic development at micro level in Karnatka. *Jour. Ind. Soc. Ag. Statistics.,* **56(1)**, 52-63.

Rai, A., Sharma, S.D., Sahoo, P.M, and Malhotra, P.K. (2008). Development of Livelihood Index for Different Agro-Climatic Zones of India. *Agricultural Economics Research Review*, **21**, 173-182.

Sharpe, A. (1999) A Survey of Indicators of Economic and Social Well-being., Canada.: Center for Study of Living Standard.

Singh, D. (2009). Principal component analysis in modelling lactation milk yield in Hariana cows. M.Sc. Thesis, CCSHAU, Hisar.

Skinner, C. J., Holmes, D. J., & Smith, T. M. F. (1986). The Effect of Sample Design on Principal Component Analysis. *Journal of the American Statistical Association*, *81*(395), 789–798. https://doi.org/10.2307/2289011

OVERVIEW OF REMOTE SENSING AND APPLICATIONS IN AGRICULTURAL SURVEYS

Prachi Misra Sahoo

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. A DEFINITION OF REMOTE SENSING

The transport of information from an object to a receiver (observer) by means of radiation transmitted through the atmosphere. The interaction between the radiation and the object of interest conveys information required on the nature of the object (eg. reflection coefficient, emittance, roughness).

Examples

(i) The reflection of sunlight from vegetation will give information on the reflection coefficient of the object and its spectral variation, and thus on the nature of the object (green trees, etc.).

(ii) Microwave radiation transmitted from a radar system and scattered from a rain cloud in the back direction to a receiver will give information on the raindrop size and intensity.

1.1 PASSIVE AND ACTIVE SENSING

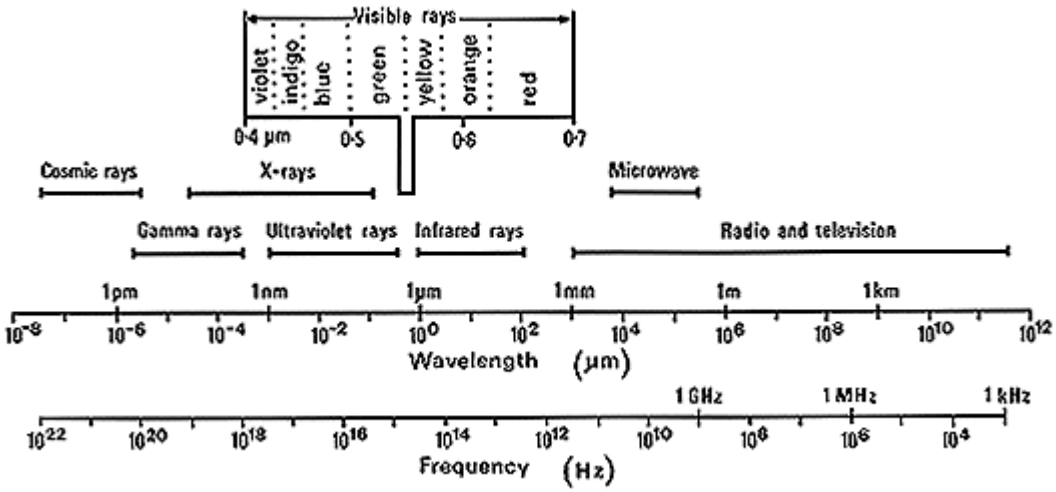
The first example above is an example of passive remote sensing, where the reflected radiation observed originates from a natural source - the sun.
The second example is an example of active remote sensing, where the scattered radiation originates from a specially designed active radar system.

1.2 ELECTROMAGNETIC RADIATION

Radiation can be observed either as a wave motion, or as single discrete packets of energy, photons. Normally, one is dealing with a large number of photons arriving in a short time, and the radiation can be treated physically as a wave motion. However, in the visible and ultraviolet regions, very weak sources are typified by the detection of single photons. The wave theory of radiation has been developed extensively. It impacts on remote sensing in the way that radiation is reflected at a surface and transmitted, absorbed and scattered in a medium.

1.3 THE ELECTROMAGNETIC SPECTRUM

Electromagnetic radiation covers a very large range of wavelengths. In Remote Sensing we are concerned with radiation from the ultraviolet (UV), which has wavelengths of from 0.3 to 0.4 um (10^-6 m) to radar wavelengths in the region of 10 cm (10^-l m) (see Figure below). Thus the phenomena observed in the various wavelength regions differ considerably.



Range of electromagnetic wavelengths and the transmission through the atmosphere.

1.4 DATA RESOLUTIONS

Resolution refers to the intensity or rate of sampling, and extent refers to the overall coverage of a data set. Extent can be seen as relating to the largest feature, or range of features, which can be observed, while resolution relates to the smallest. For a feature to be distinguishable in the data, the resolution and extent of the measurement dimensions of the data set need to be appropriate to the measurable properties of the feature. For a feature to be separable from other features, these measurements must also be able to discriminate between the differences in reflectance from the features

1.5 SPECTRAL

As indicated in the preceding sections, different materials respond in different, and often distinctive, ways to EM radiation. This means that a specific spectral response curve, or spectral signature, can be determined for each material type. Basic categories of matter (such as specific minerals) can be identified on the basis of their spectral signatures alone, but may require that the spectra be sufficiently detailed in terms of wavelength intervals and covers a wide spectral range. Composite categories of matter (such as soil which contains several different minerals) however, may not be uniquely identifiable on the basis of spectral data alone.

1.6 SPATIAL

Spatial resolution defines the level of spatial detail depicted in an image. This may be described as a measure of the smallness of objects on the ground that may be distinguished as separate entities in the image, with the smallest object necessarily being larger than a single pixel. In this sense, spatial resolution is directly related to image pixel size. In terms of photographic data, an image pixel may be compared to grain size while spatial resolution is more closely related to photographic scale.
In practical terms, the 'detectability' of an object in an image involves consideration of spectral contrast as well as spatial resolution. Feature shape is also relevant to visual discrimination in an image with long thin features such as roads showing up more readily than smaller symmetric ones. Pixel size is usually a function of the platform and sensor, while the detectability may change from place to place and time to time.

1.7 RADIOMETRIC

Radiometric resolution in remotely sensed data is defined as the amount of energy required to increase a pixel value by one quantisation level or 'count'. The radiometric extent is the dynamic range or the maximum number of quantisation levels that may be recorded by a particular sensing system. Most remotely sensed imagery is recorded with quantisation levels in the range 0&shyp;255, that is, the minimum 'detectable' radiation level is recorded as 0 while the 'maximum' radiation is recorded as 255. This range is also referred to as 8 bit resolution since all values in the range may be represented by 8 bits (binary digits) in a computer. Radiometric resolution in digital imagery is comparable to the number of tones in a photographic image &shyp; both measures being related to image contrast.

1.8 TEMPORAL

The temporal resolution of remotely sensed data refers to the repeat cycle or interval between acquisition of successive imagery. This cycle is fixed for spacecraft platforms by their orbital characteristics, but is quite flexible for aircraft platforms. Satellites offer repetitive coverage at reduced cost but the rigid overpass times can frequently coincide with cloud cover or poor weather. This can be a significant problem when field work needs to coincide with image acquisition. While aircraft data are necessarily more expensive than satellite imagery, these data offer the advantage of user-defined flight timing, which can be modified if necessary to suit local weather conditions. The off-nadir viewing capability of the SPOT&shyp;HRV provides some flexibility to the usual repeat cycle of satellite imagery by imaging areas outside of the nadir orbital path. This feature allows daily coverage of selected regions for short periods and has obvious value for monitoring dynamic events such as flood or fire.

1.9 DIGITAL IMAGE PROCESSING

The roots of remote sensing reach back into ground and aerial photography. But modern remote sensing really took off as two major technologies evolved more or less simultaneously: 1) the development of sophisticated electro-optical sensors that operate from air and space platforms and 2) the digitizing of data that were then in the right formats for processing and analysis by versatile computer-based programs. Today, analysts of remote sensing data spend much of their time at computer stations, but nevertheless still also use actual imagery (in photo form) that has been computer-processed. Now it can be seen that the individual bands and color composites that have introduced in the previous lectures and it is interesting to investigate the power of computer-based processing procedures in highlighting and extracting information about scene content, that is, the recognition, appearance, and identification of materials, objects, features, and classes (these general terms all refer to the specific spatial and spectral entities in a scene).

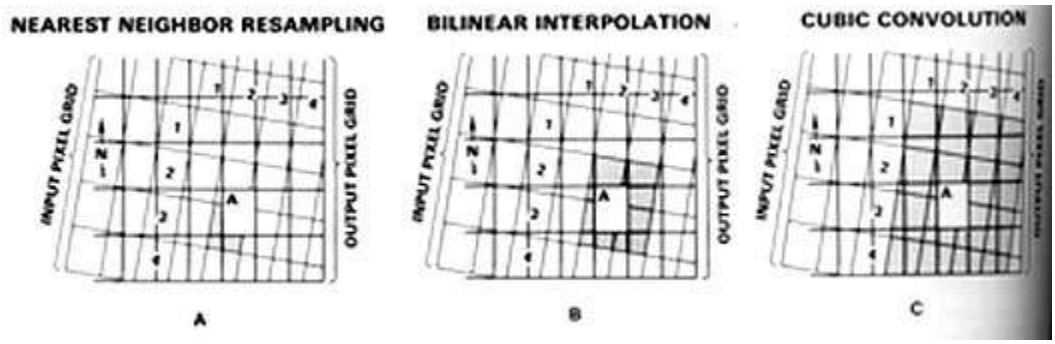
Processing procedures fall into three broad categories: *Image Restoration (Preprocessing); Image Enhancement; and Classification and Information Extraction*. Apart from preprocessing the techniques of contrast stretching, density slicing, and spatial filtering will be discussed. Under Information Extraction, ratioing and principal components analysis have elements of Enhancement but lead to images that can be interpreted directly for recognition and identification of classes and features. Also included in the third category but treated outside this lecture is Change Detection and Pattern recognition.

The data in satellite remote sensing is in the form of *Digital Number* or DN. It is said that the radiances, such as reflectance and emittances, which vary through a continuous range of values are digitized onboard the spacecraft after initially being measured by the sensor(s) in use. Ground instrument data can also be digitized at the time of collection. Or, imagery obtained by conventional photography is capable of digitization. A DN is simply one of a set of numbers based on powers of 2, such as 2^6 or 64. The range of radiances, which instrument-wise, can be, for example, recorded as varying voltages if the sensor signal is one which is, say, the conversion of photons counted at a specific wavelength or wavelength intervals. The lower and upper limits of the sensor's response capability form the end members of the DN range selected. The voltages are divided into equal whole number units based on the digitizing range selected. Thus, a IRS band can have its voltage values - the maximum and minimum that can be measured - subdivided into 2^8 or 256 equal units. These are arbitrarily set at 0 for the lowest value, so the range is then 0 to 255.

1.10 PREPROCESSING

Preprocessing is an important and diverse set of image preparation programs that act to offset problems with the band data and recalculate DN values that minimize these problems. Among the programs that optimize these values are atmospheric correction (affecting the DNs of surface materials because of radiance from the atmosphere itself, involving attenuation and scattering); sun illumination geometry; surface-induced geometric distortions; spacecraft velocity and attitude variations (roll, pitch, and yaw); effects of Earth rotation, elevation, curvature (including skew effects), abnormalities of instrument performance (irregularities of detector response and scan mode such as variations in mirror oscillations); loss of specific scan lines (requires destriping), and others. Once performed on the raw data, these adjustments require appropriate radiometric and geometric corrections.

Resampling is one approach commonly used to produce better estimates of the DN values for individual pixels. An estimate of the new brightness value (as a DN) that is closer to the B condition is made by some mathematical re-sampling technique. Three sampling algorithms are commonly used:



In the Nearest Neighbor technique, the transformed pixel takes the value of the closest pixel in the pre-shifted array. In the Bilinear Interpolation approach, the average of the DNs for the 4 pixels surrounding the transformed output pixel is used. The Cubic Convolution technique averages the 16 closest input pixels; this usually leads to the sharpest image.

2 FALSE COLOR COMPOSITE

The first example of a color composite, made by combining (either photographically or with a computer-processing program) any three bands of images with some choice of color filters, usually blue, green, and red. The customary false color composite made by projecting a green band image through a blue filter, a red band through green, and the photographic infrared image through a red filter.

2.1 TRUE COLOR VIEW

By projecting IRS Bands 1, 2, and 3 through blue, green, and red filters respectively, a quasi-true color image of a scene can be generated.

2.2 OTHER COLOR COMBINATIONS

Other combinations of bands and color filters (or computer assignments) produce not only colorful new renditions but in some instances bring out or call attention to individual scene features that, although usually present in more subtle expressions in the more conventional combinations, now are easier to spot and interpret.

2.3 CONTRAST STRETCHING:

Almost without exception, the image will be significantly improved if one or more of the functions called Enhancement are applied. Most common of these is contrast stretching. This systematically expands the range of DN values to the full limits determined by byte size in the digital data. *For IRS this is determined by the eight-bit mode or 0 to 255 DNs.* Examples of types of stretches and the resulting images are shown. Density slicing is also examined. The contrast stretching, which involves altering the distribution and range of DN values, is usually the first and commonly a vital step applied to image enhancement. Both casual viewers and experts normally conclude from direct observation that modifying the range of light and dark tones (gray levels) in a photo or a computer display is often the single most informative and revealing operation performed on the scene. When carried out in a photo darkroom during negative and printing, the process involves shifting the gamma (slope) or film transfer function of the plot of density versus exposure (H-D curve). This is done by changing one or more variables in the photographic process, such as, the type of recording film, paper contrast, developer conditions, etc. Frequently the result is a sharper, more pleasing picture, but certain information may be lost through trade-offs, because gray levels are "overdriven" into states that are too light or too dark.

2.4 SPATIAL FILTERING

Just as contrast stretching strives to broaden the image expression of differences in spectral reflectance by manipulating DN values, so spatial filtering is concerned with expanding contrasts locally in the spatial domain. Thus, if in the real world there are boundaries between features on either side of which reflectance (or emissions) are quite different (notable as sharp or abrupt changes in DN value), these boundaries can be emphasized by any one of several computer algorithms (or analog optical filters). The resulting images often are quite distinctive in appearance. Linear features, in particular, such as geologic faults can be made to stand out. The type of filter used, high- or low-pass, depends on the spatial frequency distribution of DN values and on what the user wishes to accentuate.

Another processing procedure falling into the enhancement category that often divulges valuable information of a different nature is *spatial filtering*. Although less commonly

performed, this technique explores the distribution of pixels of varying brightness over an image and, especially detects and sharpens boundary discontinuities. These changes in scene illumination, which are typically gradual rather than abrupt, produce a relation that we express quantitatively as "spatial frequencies". The spatial frequency is defined as the number of cycles of change in image DN values per unit distance (e.g., 10 cycles/mm) along a particular direction in the image. An image with only one spatial frequency consists of equally spaced stripes (raster lines). For instance, a blank TV screen with the set turned on has horizontal stripes. This situation corresponds to zero frequency in the horizontal direction and a high spatial frequency in the vertical.

3 PRINCIPAL COMPONENTS ANALYSIS

There is a tendency for multi-band data sets/images to be somewhat redundant wherever bands are adjacent to each other in the (multi-)spectral range. Thus, such bands are said to be correlated (relatively small variations in DNs for some features). A statistically based program, called Principal Components Analysis, decorrelates the data by transforming DN distributions around sets of new multi-spaced axes. The underlying basis of PCA is described in this section. Color composites made from images representing individual components often show information not evident in other enhancement products. Canonical Analysis and Decorrelation Stretching are also mentioned.

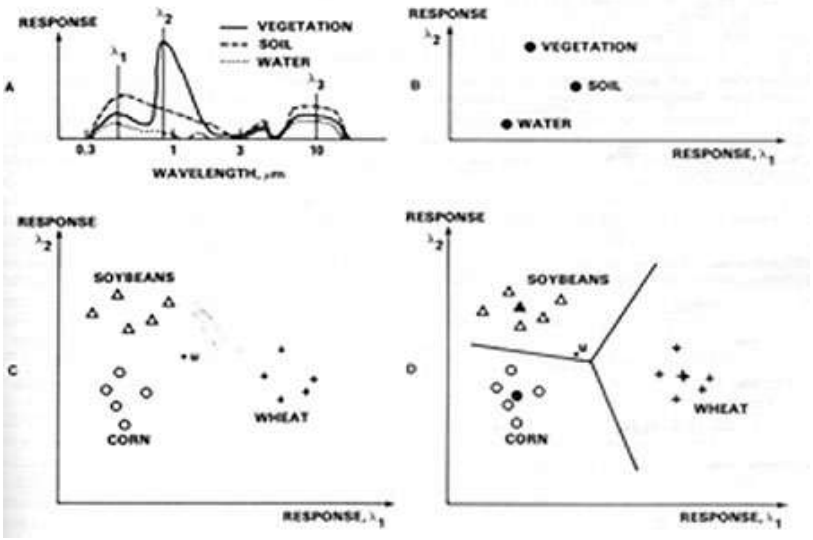
4 RATIOING

Ratioing is an enhancement process in which the DN value of one band is divided by that of any other band in the sensor array. If both values are similar, the resulting quotient is a number close to 1. If the numerator number is low and denominator high, the quotient approaches zero. If this is reversed (high numerator; low denominator) the number is well above 1. These new numbers can be stretched or expanded to produce images with considerable contrast variation in a black and white rendition. Certain features or materials can produce distinctive gray tones in certain ratios. Three band ratio images can be combined as color composites, which highlight certain features in distinctive colors. Ratio images also reduce or eliminate the effects of shadowing.

5 CLASSIFICATION

There are two of the common methods for identifying and classifying features in images: *Unsupervised* and *Supervised Classification*. Closely related to Classification is the approach called *Pattern Recognition*.

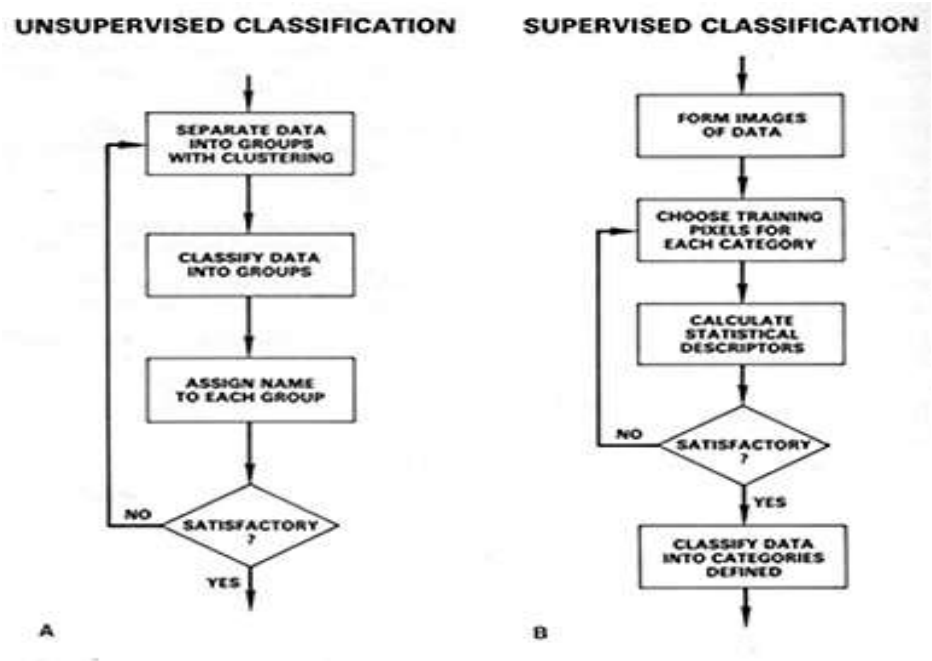
Before starting, it is well to review several basic principles, with the aid of this diagram:



In the upper left are plotted spectral signatures for three general classes: Vegetation; Soil; Water. The relative spectral responses (reflectance in this spectral interval), in terms of some unit, e.g., reflected energy in appropriate units or percent (as a ratio of reflected to incident radiation, times 100), have been sampled at three wavelengths. (The response values are normally converted [either at the time of acquisition on the ground or aircraft or spacecraft] to a digital format, the DNs or Digital Numbers cited before, commonly subdivided into units from 0 to 255 [2^8]).

Two methods of classification are commonly used: Unsupervised and Supervised. The logic or steps involved can be grasped from these flow diagrams.

In *unsupervised classification* any individual pixel is compared to each discrete cluster to see which one it is closest to. A map of all pixels in the image, classified, as to which cluster each pixel is most likely to belong, is produced (in black and white or more commonly in colors assigned to each cluster. This then must be interpreted by the user as to what the color patterns may mean in terms of classes, etc. that are actually present in the real world scene; this requires some knowledge of the scene's feature/class/material content from general experience or personal familiarity with the area imaged. In *supervised classification* the interpreter knows beforehand what classes, etc. are present and where each is in



or more locations within the scene. These are located on the image, areas containing examples of the class are circumscribed (making them training sites), and the statistical analysis is performed on the multiband data for each such class. Instead of clusters then, one has class groupings with appropriate discriminant functions that distinguish each (it is possible that more than one class will have similar spectral values but unlikely when more than 3 bands are used because different classes/materials seldom have similar responses over a wide range of wavelengths). All pixels in the image lying outside training sites are then compared with the class discriminants, with each being assigned to the class it is closest to - this makes a map of established classes (with a few pixels usually remaining unknown) which can be reasonably accurate (but some classes present may not have been set up; or some pixels are misclassified.

5.1 SUPERVISED CLASSIFICATION

Supervised classification is much more accurate for mapping classes, but depends heavily on the cognition and skills of the image specialist. The strategy is simple: the specialist must recognize conventional classes (real and familiar) or meaningful (but somewhat artificial) classes in a scene from prior knowledge, such as, personal experience with the region, by experience with thematic maps, or by on-site visits. This familiarity allows the specialist to choose and set up discrete classes (thus supervising the selection) and the, assign them category names. The specialists also locate training sites on the image to identify the classes. **Training Sites** are areas representing each known land cover category that appear fairly homogeneous on the image (as determined by similarity in tone or color within shapes delineating the category). Specialists locate and circumscribe them with polygonal boundaries drawn (using the computer mouse) on the image display. For each class thus outlined, mean values and variances of the DNs for each band used to classify them are calculated from all the pixels enclosed in the site. More than one polygon can be established for any class. When DNs are plotted as a function of the band sequence (increasing with wavelength), the result is a **spectral signature** or spectral response curve for that class. In reality the spectral signature is for all of the materials

within the site that interact with the incoming radiation. Classification now proceeds by statistical processing in which every pixel is compared with the various signatures and assigned to the class whose signature comes closest. A few pixels in a scene do not match and remain unclassified, because these may belong to a class not recognized or defined).

Many of the classes in general are almost self-evident ocean water, waves, beach, marsh, shadows. In practice, we could further sequester several such classes. For example, we might distinguish between ocean and bay waters, but their gross similarities in spectral properties would probably make separation difficult. Other classes that are likely variants of one another, such as, slopes that faced the morning sun as IRS flew over versus slopes that face away, might be warranted. Some classes are broad-based, representing two or more related surface materials that might be separable at high resolution but are inexactly expressed in the IRS image. In this category we can include trees, forests, and heavily vegetated areas (the golf course or cultivated farm fields).

5.2 MINIMUM DISTANCE CLASSIFICATION

One of the simplest supervised classifiers is the parallelopiped method. But on we employ a (usually) somewhat better approach (in terms of greater accuracy) known as the Minimum Distance classifier. This sets up clusters in multidimensional space, each defining a distinct (named) class. Any pixel is then assigned to that class if it is closest to (shortest vector distance).

We initiate our exemplification of Supervised Classification by producing one using the Minimum_Distance routine. The software program acts on DNs in multidimensional band space to organize the pixels into the classes we choose. Each unknown pixel is then placed in the class *closest* to the mean vector in this band space We can elect to combine classes to have either color themes (similar colors for related classes) and/or to set apart spatially adjacent classes by using disparate colors

5.3 MAXIMUM LIKELIHOOD CLASSIFICATION

The most powerful classifier in common use is that of Maximum Likelihood. Based on statistics (mean; variance/covariance), a (Bayesian) Probability Function is calculated from the inputs for classes established from training sites. Each pixel is then judged as to the class to which it most probably belongs. This is done with the IRS data, using three reflected radiation bands. The result is a pair of quite believable classification maps whose patterns (the classes) seem to closely depict reality but keep in mind that several classes are not normal components of the actual ground scene, e.g., shadows.

In many instances the most useful image processing output is a classified scene. This is because you are entering a partnership with the processing program to add information from the real world into the image you are viewing, in a systematic way, in which you try to associate names of real features or objects with the spectral/spatial patterns evident in individual bands, color composites, or PCI images. The most of the software are capable of producing both unsupervised and supervised classifications.

REFERENCES

1. Lillesand, T.M. and Kiefer.R .W (1987): Remote Sensing and Image Interpretaion. Jhon Wiley & Sons, New York.

2. Sabins, F.F..Jr (1987): Remote Sensing – Principles and Interpretations, W.H.Freeman and Company, New York.

OVERVIEW OF GEOGRAPHIC INFORMATION SYSTEM AND APPLICATION IN AGRICULTURAL SURVEYS

Prachi Misra Sahoo

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

Geographic Information System (GIS) is a computer based information system used to digitally represent and analyse the geographic features present on the Earth surface and the events that takes place on it. The meaning to represent digitally is to convert analog into a digital form. "Every object present on the Earth can be geo-referenced", is the fundamental key of associating any database to GIS. Here, term 'database' is a collection of information about things and their relationship to each other, and 'geo-referencing' refers to the location of a layer or coverage in space defined by the co-ordinate referencing system. Evolution of GIS has transformed and revolutionized the ways in which planners, engineers, managers etc. conduct the database management and analysis.

2. Defining GIS

A GIS is a system of hardware, software, data, people, organizations, and institutional arrangements for collecting, storing, analyzing and disseminating information about areas of the earth. It is also defined as an information system designed to work with data referenced by spatial / geographical coordinates. In other words, GIS is both a database system with specific capabilities for spatially referenced data as well as a set of operations for working with the data. A Geographic Information System is a computer based system which is used to digitally reproduce and analyse the feature present on earth surface and the events that take place on it. In the light of the fact that almost 70% of the data has geographical reference as it's denominator, it becomes imperative to underline the importance of a system which can represent the given data geographically. The three perspectives of GIS are:

2.1 GIS as a Toolbox

GIS as a toolbox: if so then what kind of tools? This is classification based on functional tasks of GIS like

1. Tools for automating *spatial data* (data capture via digitizing, scanning, remote sensing, satellite geo-position system)
2. For storing spatial data (data bases and data structures)
3. For spatial data management/retrieval
4. For analysis (overlay, buffering, proximity, network functions, spatial statistics)
5. For display of spatial data and analysis results

2.2 GIS as an Information System

As Definition of GIS indicates GIS as a specialized information system stresses "spatially distributed features (points, lines, areas), activities (physical and human-invoked), and events (time).

2.3 GIS as an approach to Geographic Information Science

1. research on GIS (algorithms, analytical methods, visualization tools, user interfaces, human-computer-human interaction)

2. research with GIS: GIS as a tool used by many substantive disciplines in their own ways (anthropology, archeology, forestry, geology, engineering, business and management sciences)

3. History of GIS

Work on GIS began in 1950s, but first GIS software came only in late 1970s from the lab of the ESRI. Canada was the pioneer in the development of GIS as a result of innovations dating back to early 1960s. Much of the credit for the early development of GIS goes to Roger Tomilson. The events which took place in the development of GIS in chronological order are:

1. Early 1950s: thematic map overlay (superimposition of maps drafted at the same scale)

2. 1959: use of transparent blacked-out overlays to find suitable locations (overhead)

3. 1960s: early computer mapping packages SURFACE II, SYMAP; early spatial data banks (CIA's World Data Bank)

4. 1960s: first GIS (Canada Geographic Information System, Minnesota Land Management System)

5. 1970s: advances in algorithms and data structures to handle spatial data, Harvard Lab for Computer Graphics (first modern GIS software)

6. 1980s: introduction of PC, advances in hardware, mature mainframe GIS software

7. 1990s: desktop GIS, Internet-based GIS services, proliferation of GIS applications

8. 2000 and beyond??? omnipresent GIS, wireless, networked, every-day applications everywhere.

4. Components of GIS

GIS constitutes of five key components: Hardware, Software, Data, People, Method

4.1 Hardware

It consists of the computer system on which the GIS software will run. The choice of hardware system range from 300MHz Personal Computers to Super Computers having capability in Tera FLOPS. The computer forms the backbone of the GIS hardware, which gets it's input through the Scanner or a digitizer board. Scanner converts a picture into a digital image for further processing. The output of scanner can be stored in many formats e.g. TIFF, BMP, JPG etc. A digitizer board is flat board used for vectorisation of a given map objects. Printers and plotters are the most common output devices for a GIS hardware setup.

4.2 Software

GIS software provides the functions and tools needed to store, analyze, and display geographic information. GIS softwares in use are MapInfo, ARC GIS, AutoCAD Map, etc. The software available can be said to be application specific. When the low cost GIS work is to be carried out desktop MapInfo is the suitable option. It is easy to use and supports many GIS feature. If the user intends to carry out extensive analysis on GIS, ARC/Info is the preferred option. For the people using AutoCAD and willing to step into GIS, AutoCAD Map is a good option. The software for a geographical information system may be split into five functional groups as mentioned below:

1. Data input and verification
2. Data storage and database management
3. Data transformation
4. Data output and presentation
5. Interaction with the user

4.3 Data

Geographic data and related tabular data can be collected in-house or purchased from a commercial data provider. The digital map forms the basic data input for GIS. Tabular data related to the map objects can also be attached to the digital data. A GIS will integrate spatial data with other data resources and can even use a DBMS, used by most organization to maintain their data, to manage spatial data.

4.4 People

GIS uses range from technical specialists who design and maintain the system to those who use it to help them perform their everyday work. The people who use GIS can be broadly classified into two classes. The CAD/GIS operator, whose work is to vectorise the map objects. The use of this vectorised data to perform query, analysis or any other work is the responsibility of a GIS engineer/user.

4.5 Method

And above all a successful GIS operates according to a well-designed plan and business rules, which are the models and operating practices unique to each organization. There are various techniques used for map creation and further usage for any project. The map creation can either be automated raster to vector creator or it can be manually vectorised using the scanned images. The source of these digital maps can be either map prepared by any survey agency or satellite imagery.

5. Data in GIS

GIS stores information about the world as a collection of thematic layers that can be used together. A layer can be anything that contains similar features such as customers, buildings, streets, lakes, or postal codes. This data contains either an explicit geographic reference, such as a latitude and longitude coordinate, or an implicit reference such as an address, postal code, census tract name, forest stand identifier, or road name. There are two components: spatial data that show where the feature is; and attribute data that provide information about the feature. These are linked by the software. The fact that there are both spatial and attribute data allows the database to be exploited in more ways than a conventional database allows, as GIS provides all the functionality of the DBMS and adds spatial functionality.

5.1 Spatial Data

Spatial data is spatially referenced data that act as a model of reality. Spatial data represent the geographical location of features for example points, lines, area etc. Spatial data typically include various kinds of maps, ground survey data and remotely sensed imagery and can be represented by points, lines or polygons.

5.2 Attribute Data

Attribute data refers to various types of administrative records, census, field sample records and collection of historical records. Attributes are either the qualitative characteristics of the spatial data or are descriptive information about the geographical location. Attributes are stored in the form of tables, where each column of the table describes one attribute and each row of the table corresponds to a feature.

6. The Nature of Geographical Data

1. **Geographical position** (spatial location) of a **spatial object** is presented by 2-, 3- or 4-dimensional coordinates in a geographical reference system (e.g. Latitude and Longitude).

2. **Attributes** are descriptive information about specified spatial objects. They often have no direct information about the spatial location but can be linked to spatial objects they describe. Therefore, it is often to call attributes "*non-spatial*" or "*aspatial*" information.

3. **Spatial relationship** specifies inter-relationship between spatial objects (e.g. direction of object B in relation to object A, distance between object A and B, whether object A is enclosed by object B, etc.).

4. **Time** records the time stamp of data acquisition, specifies life of the data, and identifies the locational and attribute change of spatial objects.

7. Methods of Data Input in GIS

7.1 Conversion of Existing Data

After receiving existing spatial data from government or private sources, GIS users often need to convert the data to a format compatible with the GIS package. Existing data cannot be easily converted because of a large variety of GIS packages and data formats in use. The choice of a conversion method basically depends upon the specificity of the data format. Some data formats are proprietary and require special translators for conversion of data from one GIS package to another. Such a conversion method can be described as direct translation. On the other hand, some data formats are neutral or public. In that case a GIS package needs to have translators that can work with the neutral format for importing and exporting data.

7.2 Creating New Data

You can create new data by using primary data sources, such as satellite or GPS (Global Positioning System) data, or secondary data sources, such as paper maps. The process of converting paper maps to digital data is called digitizing, and it can be done using digitizer or scanner technology.

7.2.1 Digitizing

Digitizing using a digitizing tablet is also called manual digitizing. The digitizing tablet has a built-in electronic mesh and can sense the position of the cursor and transmit it to the computer. The units of measurement on digitized coverages are in inches. Digitizing begins with a set of tics, which are used later for converting the coverage to real-world coordinates. Two considerations for manual digitizing are: point versus stream mode, and resolution and accuracy.

7.2.2 Scanning

A scanner converts a paper map to a scanned file, which contains raster data with values of 1 and 0. Pixels with the value 0 represent lines scanned from the paper map, and pixels with the value 1 represent non-inked areas. The scanned file is then converted to a coverage through tracing. GIS packages such as ARC/INFO have algorithms that enable users to perform semi-automatic tracing or manual tracing.

7.2.3 On-screen Digitizing

On-screen digitizing is an alternative to manual digitizing and scanning for limited digitizing work such as editing or updating an existing coverage. On-screen digitizing is manual digitizing on the computer monitor using a data source such as a DOQ as the background. This is an efficient method for digitizing, for example, new trails or roads that are not on an existing coverage but are on a new DOQ. Likewise, the method can be used for editing a vegetation coverage based on new information from a new DOQ that shows recent clear-cuts or burned areas. Obviously, the major shortcoming of this method is the resolution of the computer monitor, which is much coarser than a digitizing table or a scanner.

8. Uses of GIS

There are three basic categories of use that GIS can be put to:

1. as a spatially referenced database;
2. as a visualisation tool; and
3. as an analytic tool.

A spatially referenced database allows us to ask questions such as 'what is at this location?', 'where are these features found?', and 'what is near this feature?'. It also allows us to integrate data from a variety of disparate sources. For example to study the dataset on hospitals we might also want to use census data on the population of the areas surrounding each hospital. Census data are published for districts that can be represented in the GIS using polygons as spatial data. As we have the coordinates of the hospitals and the coordinates of the district boundaries we can bring this data together to find out which district each hospital lay in, and then compare the attribute data of the hospitals with the attribute data from the census. We may also want to add other sorts of data to this: for example data on rivers represented by lines; or wells represented by points to give information about water quality. In this way information from many different sources can be brought together and interrelated through the use of location. This ability to integrate is one of the key advantages of GIS.

Once a GIS database has been created, mapping the data it contains is possible almost from the outset. This allows the researcher a completely new ability to explore spatial patterns in the data right from the start of the analysis process. As the maps are on-screen they can be zoomed in on and panned around. Shading schemes and classification methods can be changed, and data added or removed at will. This means that rather than being a product of finished research, the map now becomes an integral part of the research process. New ways of mapping data are also made possible, such as animations, fly-throughs of virtual landscapes, and so on. It is also worth noting the visualization in GIS is not simply about mapping: other forms of output such as graphs and tables are equally valid ways of visualising data from GIS.

9. Questions GIS can answers

Till now GIS has been described in two ways (i) Through formal definitions, and (ii) Through technology's ability to carry out spatial operations, linking data sets together. However there is another way to describe GIS by listing the type of questions the technology can (or should be able to) answer. Location, Condition, Trends, patterns, Modelling, Aspatial questions, Spatial questions. There are five type of questions that a sophisticated GIS can answer:

Location What is at…….?
Condition Where is it…….?
Trends What has changed since……..?
Patterns What spatial patterns exists……..?
Modelling What if……..?
Aspatial Questions
Spatial Questions

10. Application Areas

GIS is now used extensively in government, business, and research for a wide range of applications including environmental resource analysis, landuse planning, locational analysis, tax appraisal, utility and infrastructure planning, real estate analysis, marketing and demographic analysis, habitat studies, and archaeological analysis.

One of the first major areas of application has been in **natural resources management**, including management of

1. wildlife habitat
2. wild and scenic rivers
3. recreation resources
4. floodplains
5. wetlands
6. agricultural lands
7. aquifers
8. forests

One of the largest areas of application has been in **facilities management**. Uses for GIS in this area includes

1. locating underground pipes and cables,
2. balancing loads in electrical networks,
3. planning facility maintenance,
4. tracking energy use.

Local, state, and federal governments have found GIS particularly useful in **land management**. GIS has been commonly applied in areas like

1. zoning and subdivision planning,
2. land acquisition,
3. environmental impact policy,

4. water quality management,
5. maintenance of ownership.

More recent and innovative uses of GIS have used information based on **street-networks**. GIS has been found to be particularly useful in

1. address matching,
2. location analysis or site selection,
3. development of evacuation plans.

The range of applications for GIS is growing as systems become more efficient, more common, and less expensive.

11. Basic Data Models in GIS

The data model represents a set of guidelines to convert the real world (called entity) to the digitally and logically represented spatial objects consisting of the attributes and geometry. The attributes are managed by thematic or semantic structure while the geometry is represented by geometric-topological structure. There are two major types of geometric data model; vector and raster model as shown in Fig. 1.

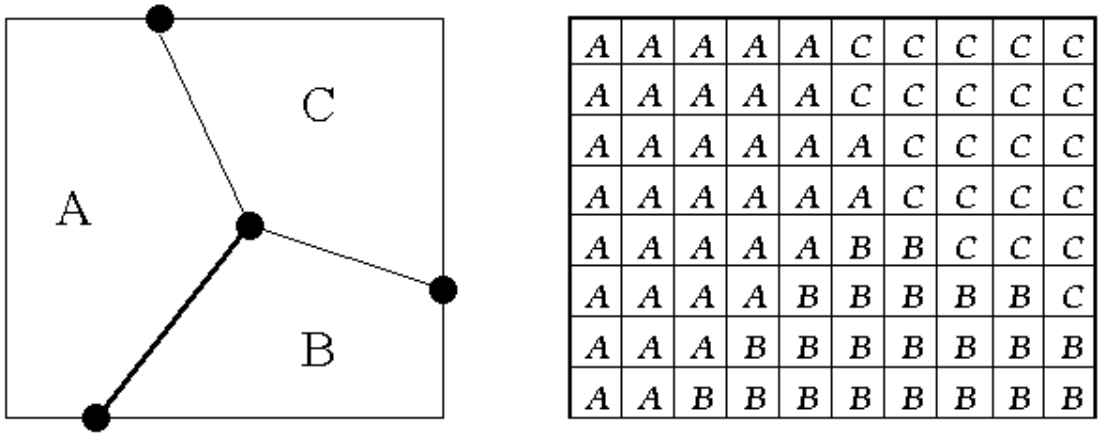


Figure 1. Vector and Raster data models

11.1 Vector Data Model

Vector model uses discrete points, lines and/or areas corresponding to discrete objects with name or code number of attributes. Given a map, you can tell how map features are like and how the map features are related to one another spatially.

11.1.1 Geometry of Vector Data Model

The vector data model consists of three types of geometric objects: point, line, and area. A point may represent a gravel pit, a line may represent a stream, and an area may represent a vegetated area.

A point has 0 dimension. A point feature occupies a location and is separate from other features (Figure 2).

A line is one-dimensional and has the property of length. A line feature is made of points: a beginning point, an end point, and a series of points marking the shape of the line, which may be a smooth curve or a connection of straight-line segments. Smooth curves are typically generated or fitted by mathematical equations, such as cubic polynomial equations. Straight-line segments may represent human-made features or approximations of curves in data entry. Points that mark the shape of a line feature but are not nodes are called vertices. Line features may intersect or join with other lines and may form a network (Figure 3).

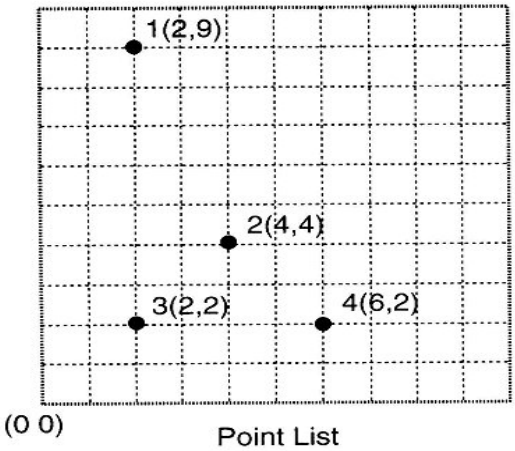
An area is two-dimensional and has the properties of area and boundary. The boundary of an area feature separates the interior area from the exterior area. Area features may be isolated or connected. An isolated area feature typically has a node serving as both the beginning and end node. Area features may be surrounded by other areas and form holes within them. Area features may overlap one another and create overlapped areas. For example, the fired areas from previous forest fires may overlap each other (Figure 4).

Vector data representation using point, line, area, and volume is not always straightforward because it may depend on map scale and, occasionally, criteria established by government mapping agencies. A city on a 1:1,000,000-scale map is represented as a point, but the same city is shown as an area on a 1:24,000-scale map. A stream is shown as a single line near its headwaters but as an area along its lower reaches. In this case, the width of the stream determines how it should be represented on a map.

11.1.2 Topology of Vector Data Model

Topology expresses explicitly the spatial relationships between geometric objects. The vector data model in ARC/INFO supports three basic topological concepts:

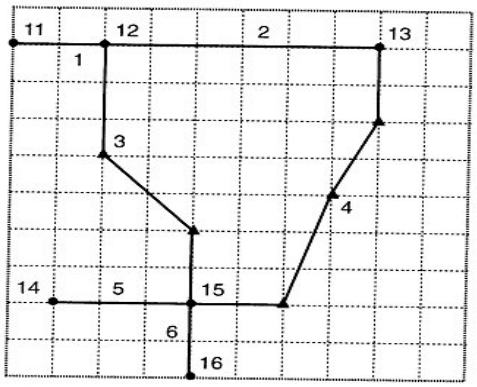
1. Connectivity: Arcs connect to each other at nodes
2. Area definition: An area is defined by a series of connected arcs
3. Contiguity: Arcs have directions and left and right polygons



Point List

ID	x,y
1	2,9
2	4,4
3	2,2
4	6,2

Figure 2. Points with x-, y-coordinates



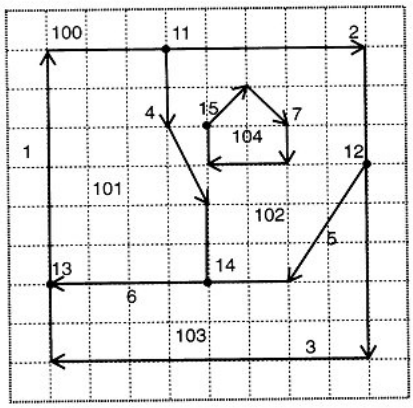
Arc-node list

Arc#	Fnode	Tnode
1	11	12
2	12	13
3	12	15
4	13	15
5	15	14
6	15	16

Arc-coordinate list

Arc#	x,y Coordinates
1	(0,9) (2,9)
2	(2,9) (8,9)
3	(2,9) (2,6) (4,4) (4,2)
4	(8,9) (8,7) (7,5) (6,2) (4,2)
5	(4,2) (1,2)
6	(4,2) (4,0)

Figure 3. The data structure of a line data model



Left/Right list

Arc#	Lpoly	Rpoly
1	100	101
2	100	102
3	100	103
4	102	101
5	103	102
6	103	101
7	102	104

Polygon/Arc list

Polygon #	Arc#
101	1,4,6
102	4,2,5,0,7
103	6,5,3
104	7

Figure 4. The data structure of an area data model

11.1.3 Advantages and Disadvantages of Vector Data Models

The advantages of the vector data model are:

1. Good representation of entity data models. Compact data structure.
2. Topology can be described explicitly – therefore good for network analysis.
3. Coordinate transformation and rubber sheeting is easy.
4. Accurate graphic representation at all scales.
5. Retrieval, updating and generalization of graphics and attributes are possible.

The disadvantages of the vector data model are:

1. Complex data structures
2. Combining several polygon networks by intersection and overlay is difficult and requires considerable computer power.

3. Display and plotting may be time consuming and expensive, particularly for high-quality drawing, colouring, and shading.

4. Spatial analysis within basic units such as polygons is impossible without extra data because they are considered to be internally homogeneous.

5. Simulation modeling of process of spatial interaction over paths not defined by explicit topology is more difficult than with raster structures because each spatial entity has a different shape and form.

11.2 Raster Format

Raster model uses regularly spaced grid cells in specific sequence. An element of the grid cell is called a pixel (picture cell). The conventional sequence is row by row from the left to the right and then line by line from the top to bottom. Every location is given in two dimensional image coordinates; pixel number and line number, which contains a single value of attributes.

11.2.1 Geometry of Raster Data

The geometry of raster data is given by point, line and area objects as follows (see Figure 5)

a. Point Objects: A point is given by point ID, coordinates (i, j) and the attributes

b. Line Objects: A line is given by line ID, series of coordinates forming the line, and the attributes

c. Area Objects: An area segment is given by area ID, a group of coordinates forming the area and the attributes. Area objects in raster model are typically given by "Run Length" that rearranges the raster into the sequence of length (or number of pixels) of each class as shown in Figure 5.

The topology of raster model is rather simple as compared with the vector model as shown in Figure 5. The topology of line objects is given by a sequence of pixels forming the line segments. The topology of an area object is usually given by "Run Length" structure which includes Start line no., (start pixel no., number of pixels), second line no., (start pixel no., number of pixels).

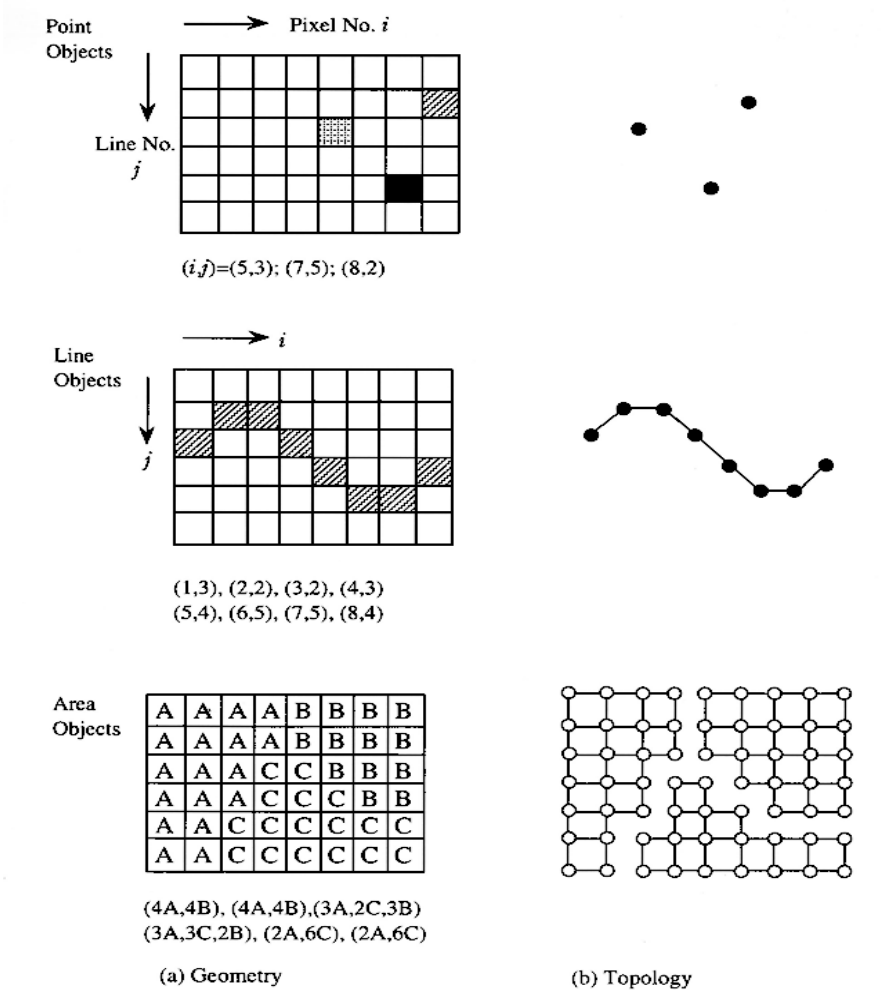


Figure 5. Geometry and Topology of Raster Data

11.2.2 Topology of Raster Data

One of the weak points in raster model is the difficulty in network and spatial analysis as compared with vector model. For example, though a line is easily identified as a group of pixels which form the line, the sequence of connecting pixels as a chain would be a little difficult in tracing. In case of polygons in raster model, each polygon is easily identified but the boundary and the node (when at least more than three polygons intersect) should be traced or detected.

a. Flow Directions

A line with directions can be represented by four directions called as the Rook's move in the chess game or eight directions called as the Queen's move, as shown in Figure 6 (a), (b), (c).

Figure 6 (c) shows an example of flow directions in the Queen's move. Water flow, links of a network, roads etc. can be represented by the flow directions (or called Freeman chain code).

b. Boundary

Boundary is defined as 2 x 2 pixel window that has two different classes as shown in Figure 7 (a). If a window is traced in the direction shown in Figure 7 (a), the boundary can be identified.

c. Node

A node in polygon model can be defined as a 2 x 2 window that has more than three different classes as shown in Figure 7 (b). Figure 7 (c) and (d) show an example of identification of pixels on boundary and node.

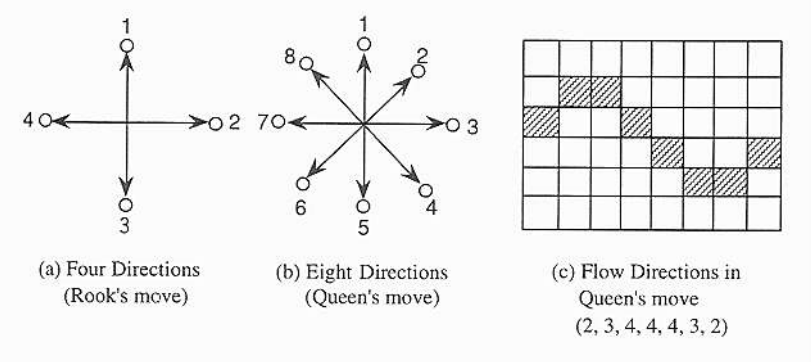


Figure 6. Flow Directions

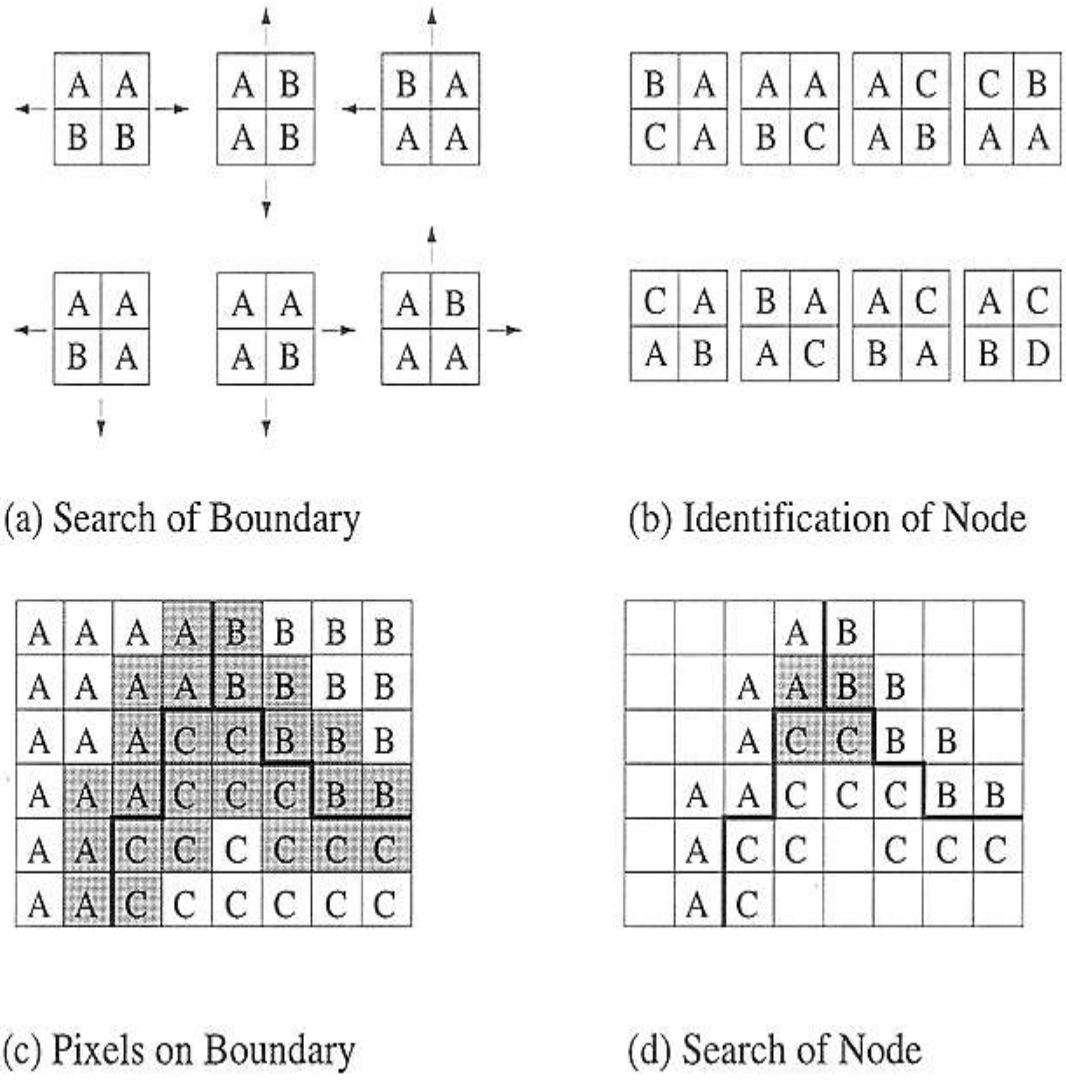


Figure 7. Identification of Boundary and Node

11.2.3 Advantages and Disadvantages of Raster Data Models

The advantages of the raster data model are:

1. Simple data structures.
2. Location-specific manipulation of attribute data is easy.
3. Many kinds of spatial analysis and filtering may be used.
4. Mathematical modeling is easy because all spatial entities have a simple, regular shape.

5. The technology is cheap.

6. Many forms of data are available.

The disadvantages of the raster data model are:

1. Large data volumes.

2. Using large grid cells to reduce data volumes reduces spatial resolution, result in loss of information and an inability to recognize phenomenological defined structures.

3. Crude raster maps are inelegant though graphic elegance is becoming much less of a problem today. Coordinate transformations are difficult and time consuming unless special algorithms and hardware are used and even then may result in loss of information or distortion of grid cell shape.

11.3 Quadtree Data Model

Traditionally, the raster model is based on dividing the real world into equal-sized rectangular cells. However, in many cases, it can be more practical to use a model with varying cell size. Larger cells (lower resolution) may be used to represent larger homogeneous areas, and smaller cells (higher resolution) may be used for more finely detailed areas. This approach, known as the quad-tree representation, is a refinement of the block code method. In representing a given areas, the aggregate amount of data involved is proportional to the square of the resolution (into cells). Because the quad-tree model is a very practical concept, it is preferable for the storage of both small and large volumes of data.

The quad-tree paradigm divides a geographical area into square cells of sizes varying from relatively large to that of the smallest cell of the raster. Usually, the squares are then quartered into four smaller squares. The quartering may be continued to a suitable level until a square is found to be so homogeneous that it no longer needs to be divided, and the data on it can be stored as a unit. A larger square may therefore comprise several raster cells having the same values. However, homogeneous areas that are not square or do not coincide with the pattern of squares employed may be further divided into homogeneous squares. The structure of the quad-tree resembles an inverted tree, whose leaves are pointers to the attributes of homogeneous squares and whose branch forks are pointers to smaller squares – hence the name quad-tree (Figure 8).

11.3.1 Advantages and Disadvantages of Quadtree Data Models

The advantages of the quad-tree model are:

1. Rapid data manipulation, because homogeneous areas are not divided into the smallest cells used.

2. Rapid search, because larger homogeneous areas are located higher up in the point structure

3. Compact storage, because homogeneous squares are stored as units.

4. Efficient storage structure for certain operations, including searching for neighboring squares or for a square containing a specific point.

5. The disadvantages of the quad-tree model are:

6. Establishing the structure requires considerable processing time.

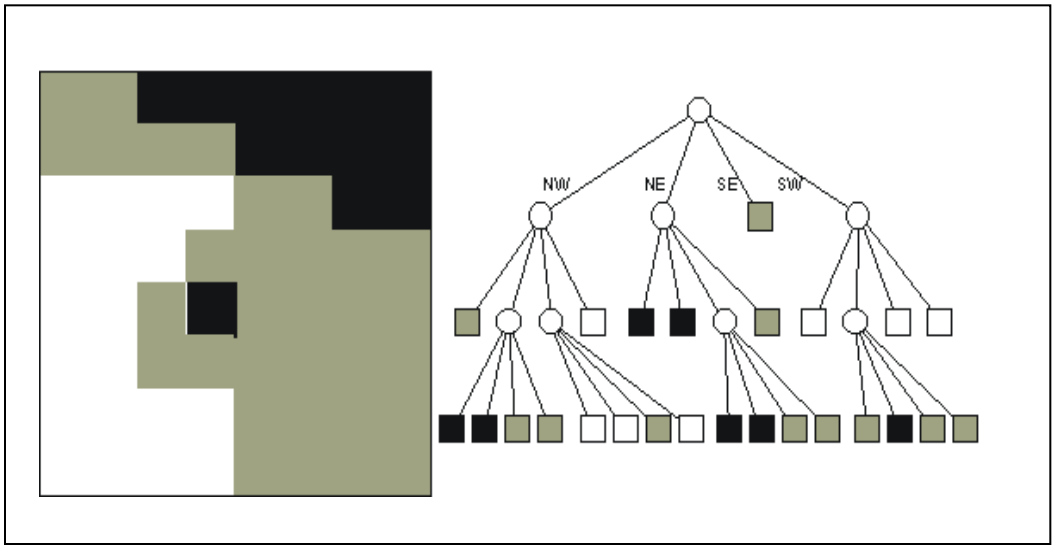


Figure 8. Quadtree data model

7. Protracted processing may prolong alterations and updating

8. Data entered must be relatively homogeneous

9. Complex data may require more storage capacity than ordinary raster storage.

12. Advanced Data Models in GIS

In GIS, continuous surface such as terrain surface, meteorological observation (rain fall, temperature, pressure etc.) population density and so on should be modeled. As sampling points are observed at discrete interval, a surface model to present the three dimensional shape; z = f (x, y) should be built to allow the interpolation of value at arbitrary points of interest. Usually the following four types of sampling point structure are modeled into DEM.

12.1 Grid Model

A systematic grid, or raster, of spot heights at fixed mutual spaces is often used to describe terrain. Elevation is assumed constant within each cell of the grid; that is, the area represented by each cell is shown as a flat area in the model. Thus, small cells detail terrain more accurately than large cells. The size of cells is constant in a model, so areas with a

greater variation of terrain may be described less accurately than those with less variation. The grid model is most suitable for describing random variations in the terrain, whereas the systematic linear structures can easily disappear or be deformed. A possible solution is to store the data as individual points and generate grids of varying density as required. It is debatable whether the grid model represents samples on a grid and can therefore be called a point model, or represents an average across raster cells. In the United States the former seems to be the most usual. Elevation values are stored in a matrix, and the contiguity between points is thus expressed through the column and line numbers.

Different interpolation techniques are used to generate an elevation grid from source data such as points, contour lines, and break lines. In interpolation of elevation values for the cells, it is usual to assume that points located at a distance. The averages of the elevations of those closed to grid points, within a given circle or square, can be assigned to the grid points with inverse weighting in proportion to the intervening distances involved. More advanced statistical methods can replace this kind of simple weighting in order to obtain a best possible model of the terrain based on available data. When the data relate to profiles or contours, grid point elevations are interpolated, in the same way, from the elevations at the intersections of the original data lines and the lines of the grid.

12.2 TIN Model

An area model is an array of triangular areas with their corners stationed at selected points of most importance, for which the elevations are known. The inclination of the terrain is assumed to be constant within each triangle. The areas of the triangles may vary, with the smallest representing those areas in which the terrain varies most. The resulting model is called the triangulated irregular network(TIN)

In so far as possible, small equilateral triangles are preferable. To construct a TIN, as measured points are built and the model thus represents lines of fracture, single points, and random variations in the terrain. The points are established by triangulation and in such a way that no other points are located within each triangle's converted circle. In the TIN model, the x-y-z coordinates of all points, as well as the triangle attributes of inclination and direction, are stored. The triangles are stored in a topological vector data storage structure comprising polygons and nodes, thereby preserving the triangle's contiguity, which eases the calculation of z values for new points.

12.3 Contour lines

Interpolation based on proportional distance between adjacent contours is used. TIN is also used.

12.4 Profile

Profiles are observed perpendicular to an alignment or a curve such as high ways. In case the alignment is a straight line, grid points will be interpolated. In case the alignment is a curve, TIN will be generated. Figure 9 shows different types of DEMs.

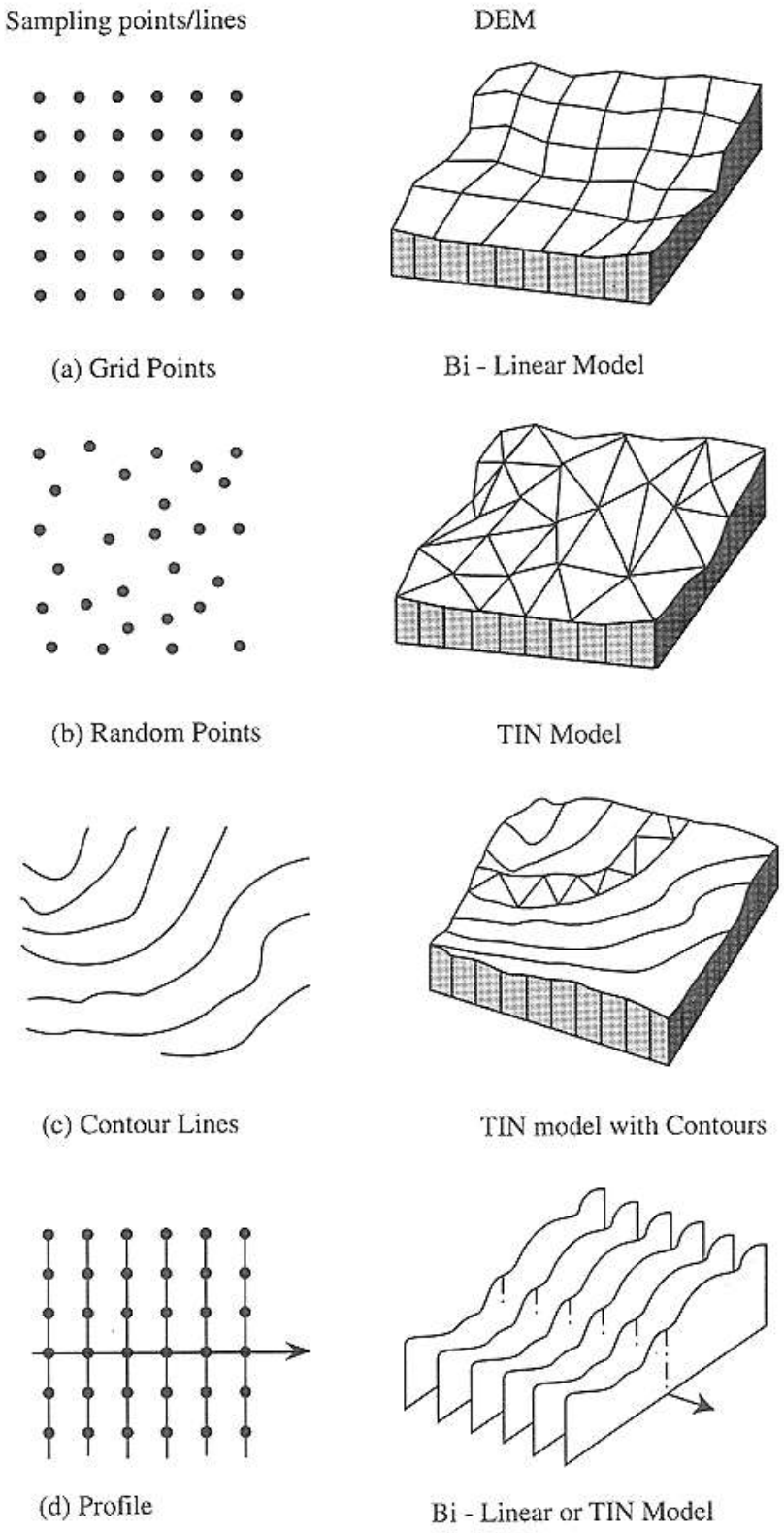


Figure 9. Different types of DEMs

References

Antenucci, J.C., Brown, K., Croswell, P. L., Kevany, M. J., and Archer, H.N. (1991). "Introduction," "Evolution of the Technology," and "Applications." Chaps. 1-3 in Geographic Information Systems: A Guide to the Technology. New York: Van Nostrand Reinhold.

Bernhardsen, T. (2002) Geographic Information Systems: An Introduction. John Wiley & Sons, Inc.

Buckley, D. J. The GIS Primer: An Introduction to Geographic Information System. http://www.innovativegis.com/basis/primer/primer.html.

Buckley, D. J. The GIS Primer: An Introduction to Geographic Information System. http://www.innovativegis.com/basis/primer/primer.html.

Burrough, P.A. (1986). Principles of Geographic Information Systems for Land Resources Assessment. Oxford: Oxford University Press.

Burrough, P.A. and McDonnell R.A. (1986) Principles of Geographic Information Systems. Oxford: Oxford University Press.

Davis, B. E. (2001) GIS: A Visual Approach. Onward Press.

Geographic Information System: Primer, Geospatial Training and Analysis Cooperative http://geology.isu.edu/geostac/Field_Exercise/gisprimer/Frameset.html.

Geographic Information System: Primer, Geospatial Training and Analysis Cooperative http://geology.isu.edu/geostac/Field_Exercise/gisprimer/Frameset.html.

Huxhold, E. (1991). Information in the Organization. An Introduction to Urban Geographic Information Systems. New York: Oxford University Press.

Star and John. (1990). Geographic Information Systems: An Introduction. Englewood Cliffs, NJ: Prentice- Hall.

Tobler, W.R. (1959). Automation and Cartography. *Geographical Review 49,* 526-534.

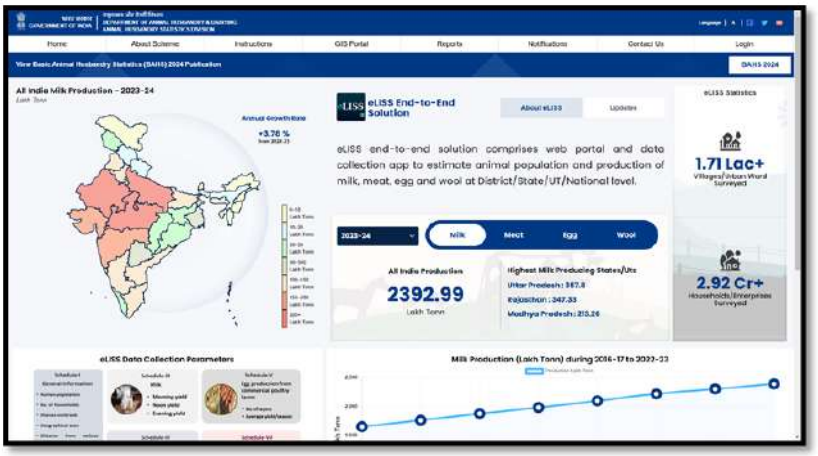
GENERATION OF LIVESTOCK STATISTICS IN INDIA USING eLISS PORTAL AND APP

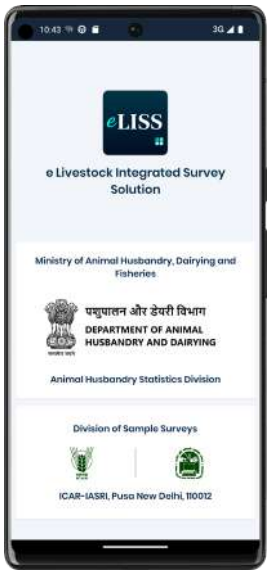
Prachi Misra Sahoo

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1.0 Introduction

The efficient collection and management of livestock data is a cornerstone for informed decision-making and policy formulation in India's agricultural sector. Given the country's vast geographical expanse and diversity in agricultural practices, accurate and timely data on key livestock products pertaining to milk, meat, eggs, and wool are indispensable. This data provide the foundation for crafting evidence-based strategies, driving investments, and ensuring the sustainable growth of the livestock industry. However, conventional methods of data collection have often proven inadequate in meeting these demands, plagued by inefficiencies, delays, and inaccuracies. These limitations underscore the need for a comprehensive, technology-driven solution capable of addressing the complexities of livestock data management in a country as diverse and expansive as India.





The eLISS (e Livestock Integrated Sample Survey) was conceptualized and developed as a robust solution to address these challenges. The system consists of the eLISS web portal and the eLISS Data Collection App, which together create an integrated framework for the end-to-end management of livestock data. This solution enables seamless operations across all stages of data management, including three-stage sample selection, real-time sample allocation, field data collection, two-tier data supervision, and system-based data monitoring. It also supports data analysis and generates estimates of number of animals and production at the National/State/UT/District levels. The developed solution is tested and validated and has been implemented across all 36 States and Union Territories, ensuring nationwide applicability and efficiency.

The implementation process involved dividing the country into six geographical zones to manage training and adoption effectively. A national-level training session was conducted, followed by zonal trainings in both online and physical mode, ensuring that state and district officials were well-equipped to use the system. Comprehensive training materials, including operational manuals, tutorials, and FAQs, were made available to users, streamlining the on boarding process. During implementation, the system was continuously improved through user feedback, resulting in the release of multiple versions of the app and multiple updates to the web portal. The recent launch of eLISS V2.0 introduced three new modules: a State/UT-level feature on the app, a real-time supervision module, and an automated query resolution system, further enhancing the platform's functionality.

The advantages of eLISS are evident in its scale and efficiency. It has facilitated data collection from over 2.92 crore households and enterprises, spanning more than 1.72 lac villages,
654 slaughterhouses, and 33,000 commercial poultry farms. With over 28,000 active enumerators, 9,200 supervisors, and 742 district nodal officers, the system ensures extensive coverage and accuracy. The platform is accessible through the web portal and the mobile app, making it user-friendly and widely accessible. The automated features for data supervision and analysis reduce human errors and enable real-time monitoring, while the modular nature of the system ensures adaptability to new requirements.

The success of eLISS was demonstrated during the ISS survey for 2023-2024, where the system was used in all districts across three seasons Summer, Rainy, and Winter. More than 99% of the collected data was synced with the server, highlighting the system's reliability and effectiveness. By revolutionizing the way livestock data is managed, eLISS provides a dependable framework for decision-making, ensuring that policies and strategies in the livestock sector are based on accurate and comprehensive data. This end-to-end solution marks a significant step forward in harnessing technology to transform livestock data management in India.

The entire methodology followed to develop this End-to-End system is described below

1.1 Coverage of Survey

The Survey is conducted in the entire rural and urban areas of all States/UTs and all districts. The survey is conducted in the selected sampled villages/urban wards enlisting all households/ enterprises and butcher shops. The survey also covers all the Commercial Poultry Farms and Slaughter Houses present in the district.

1.2 Period of Survey

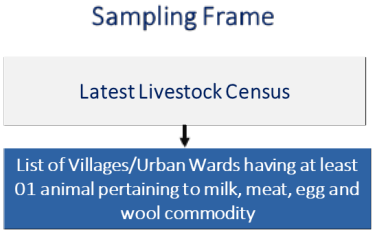
The Survey is conducted from March to February every year. The entire period of one year is divided into three Seasons of 4 months each. These seasons are:

Name of the Season	Period of collection of data
Summer Season	1st March to 30th June (122 days)
Rainy Season	1st July to 31st October (123 days)
Winter Season	1st November to 28th or 29th February(120 days or 121 days in a leap year)

2.0 Preparation of sampling frame

The lists of villages/urban wards as per the latest Livestock Census constitutes the sampling frame for selection of first stage unit i.e. villages/urban wards. In addition to this, sampling frame for commercial poultry farms and slaughter houses is prepared by District Nodal Officers (DNOs) at district level. The details are mentioned below.

2.1 Sampling frame for First Stage Unit i.e. villages/urban wards



Sampling frame consists of all villages/urban wards having at least 01 animal obtained from the latest livestock census data. Every year this sampling frame needs to be updated for all States/UTs. This updation includes, change in names of villages/urban wards/districts, formation of new districts in States/UTs, movement of villages/urban wards from one district to another and removal of uninhabited villages/urban wards.

To facilitate this updation of sampling frame, AHS/States/UTs can view the current sampling frame available on the eLISS web portal under the *Sampling Frame Module*. AHS/States/UTs can also download this sampling frame which includes district name, block name, block code, village/urban ward name and village/urban ward census code in excel format.

2.2. Preparation of sampling frame for commercial poultry farms and slaughter houses using eLISS web portal

An additional module was developed for preparation of sampling frame for commercial poultry farms and slaughter houses on the eLISS web portal. This module is currently live and all DNOs have prepared the complete frame for commercial poultry farms and slaughter houses. Further, DNOs are can add new commercial poultry farms and slaughter houses in each season to update this frame.

DNOs also have option to mark a commercial poultry farm or slaughter house as operational/non-operational for a season. Further, DNOs can mark commercial farm operational/non-operational for schedule-V and schedule-VIII separately. In case there is no slaughter house present in a district then DNOs are instructed to add atleast two butcher shops covering all slaughtering animals of their district.

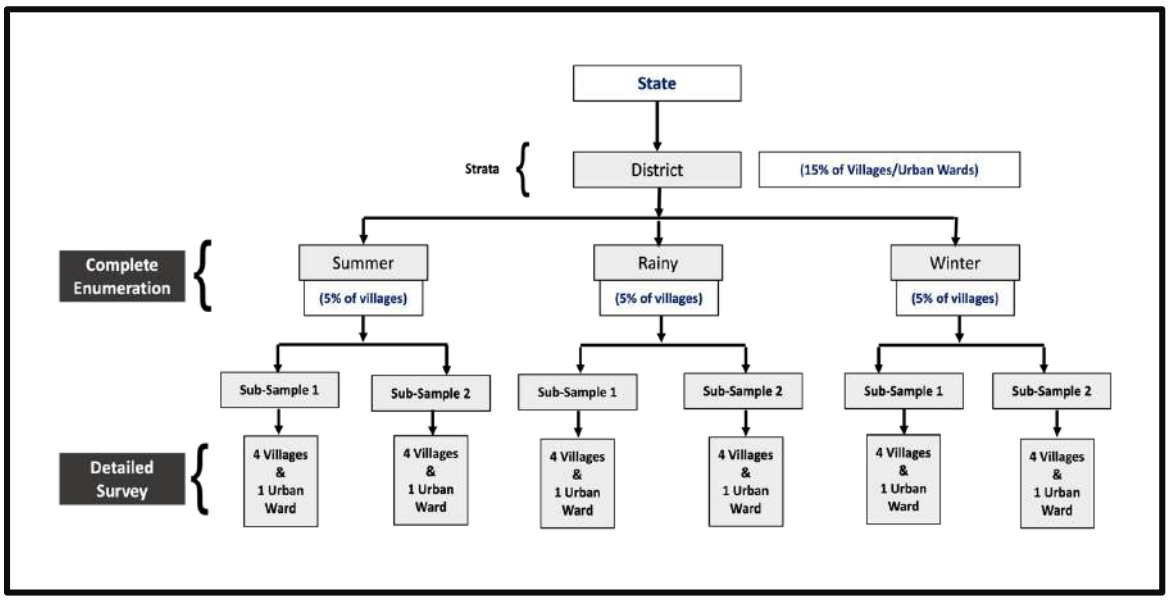
This sampling frame of commercial poultry farms and slaughter houses is being used for field data collection in every season.

3.0 Sampling design and Sample size

The sampling design followed for selection of sample is a stratified three stage design with district as stratum. The first stage units are villages or urban wards, second stage units are households and third and ultimate stage units are animals which are explained in the subsequent sections.

The State Department of Animal Husbandry is the nodal agency for implementation of ISS scheme. The survey under ISS scheme is conducted in two stages - Complete Enumeration and Detailed Survey. The estimated livestock population is computed based on complete enumeration and the average yield is estimated based on detailed survey.

The sampling design along with sample size for each stage is explained below.

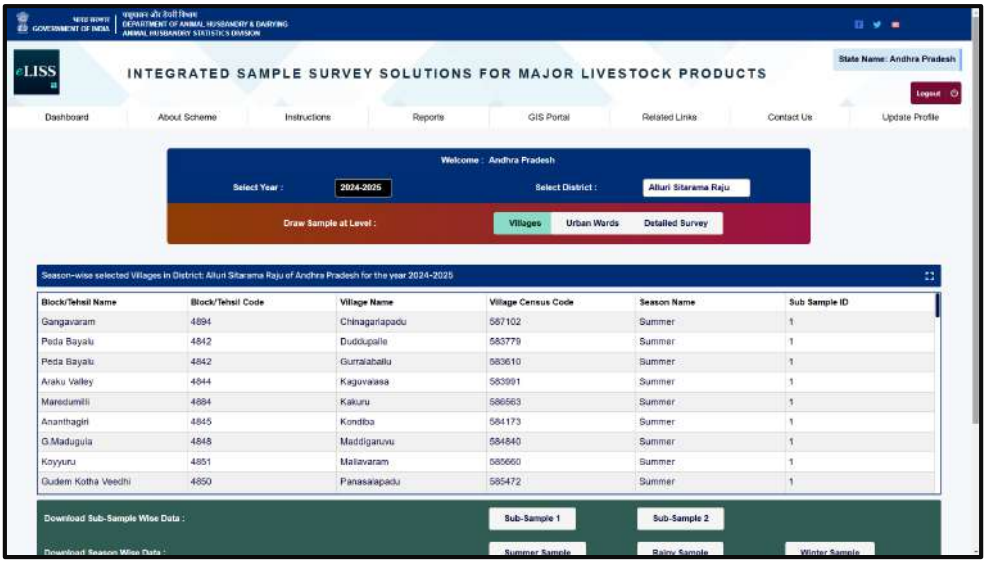


3.1 Selection of First stage sampling unit

The first stage units i.e. villages/urban wards are selected using the sample selection module of eLISS web portal. States/UTs are authorized for selecting the first stage unit for all of their districts.

Sample selection module follows ISS sampling design i.e. the sample of villages/urban wards is selected using simple random sampling without replacement (SRSWOR). On a single click of a button using eLISS web portal, the sample of villages/urban wards for each district is selected both for complete enumeration and detailed survey for all three season of a year.

Once State/UT selects the first stage unit sample for a district, then district can view the selected sample on eLISS web portal. Further, this module allows State/UT officials to download the selected sample in excel based file for both complete enumeration and detailed survey.



3.1.1 Substitution of first stage sample i.e. villages/urban wards and grant permission for allocation using eLISS web portal

In case, the selected village/urban ward is uninhabited or inaccessible, the system provides authority to State/UT officials to substitute the village/urban by clicking the re-draw village/urban ward option on the eLISS web portal under the survey status page. This re-draw option will substitute the uninhabited or inaccessible village/urban ward and replace that village/urban ward with a new randomly selected village/urban ward. Presently, two re-draw chances are provided to state/UT officials for every district in each season. The re-draw chances can be further increased later, only on request of State/UT officials. For substituting the selected village/urban ward, a period of 10 days is provided in the start of each season. During this period, DNOs needs to check the selected sample and report to the State/UT, if substitution of any villages/urban ward is required. State/UT then re-draw the village/urban ward and provide new village/urban ward. After substitution of villages/urban wards, the States/UTs needs to give permission to the DNOs for allocation of selected villages/urban wards to the enumerators. In case DNOs do not require substitution of any village/urban ward then they should request State/UT officials for granting permission of allocation immediately.

It is to be noted here that once permission for allocation is granted for a district, no substitution of villages/urban wards is possible at State/UT level.

After the 10 days windows, system will automatically grants permission for allocation of villages/urban wards to all the DNOs.

3.2 Selection of Second and Third stage sampling units

The list of households/enterprises of the selected village/urban ward for detailed study serves as the sampling frame for second and third stage sample. Second and third stage sample will be selected using the eLISS data collection app after completion of schedule-II. eLISS data collection app follows ISS methodology for selecting the sample. The sampling design and sample size of second and third stage sample is as explained and shown below:

1) Milk Commodity

- **Second stage sample consists of** 02 households each for exotic, crossbred, indigenous, non-descript cattle and indigenous & non-descript buffaloes. Total

12 households are selected for every season. In case selected households are not having in milch goats, two additional Households having goats have to be selected.

- **Third stage** sample consists of two in-milk animals from each household of Exotic, Crossbred, Indigenous, non-descript cattle and Indigenous, non-descript Buffaloes i.e. 24 animals are to be selected. All in-milk goats in selected households will be considered.

2) Egg Commodity

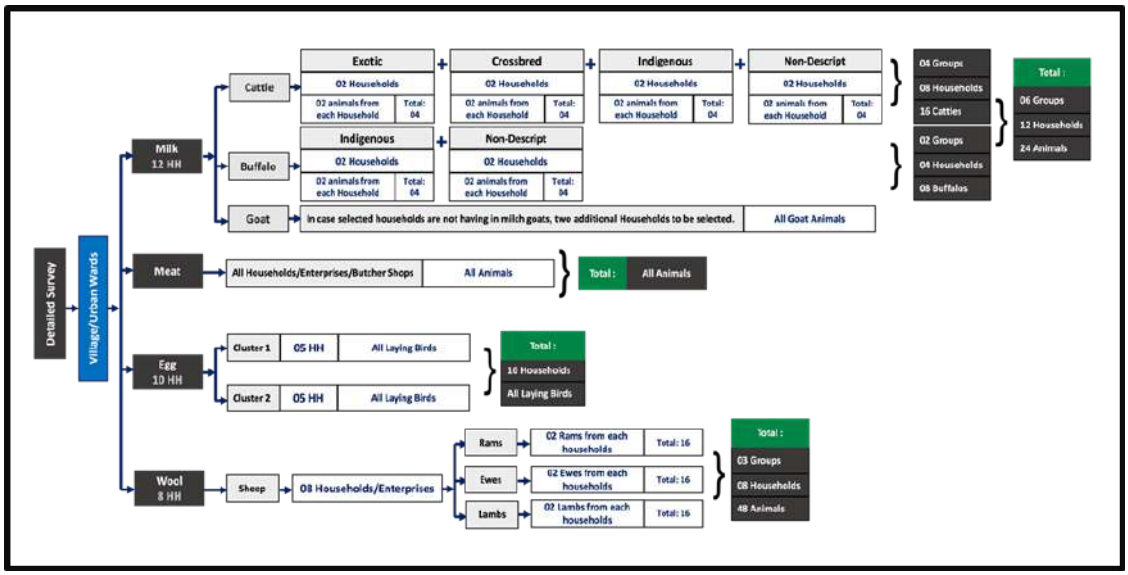
- **Second stage sample consists of** total 10 households fixed for every season.
- **Third stage** sample consists of all laying birds.

3) Meat Commodity

- **Second stage sample consists of** all the households/butcher shops.
- **Third stage** sample consists all animals/poultry slaughtered in each households/ butcher shops.

4) Wool Commodity

- **Second stage sample consists of** 08 households fixed for every season
- **Third stage** sample consists of two rams/wethers, two ewes, two lambs {Total 48 sheep} or as available in the selected sample.



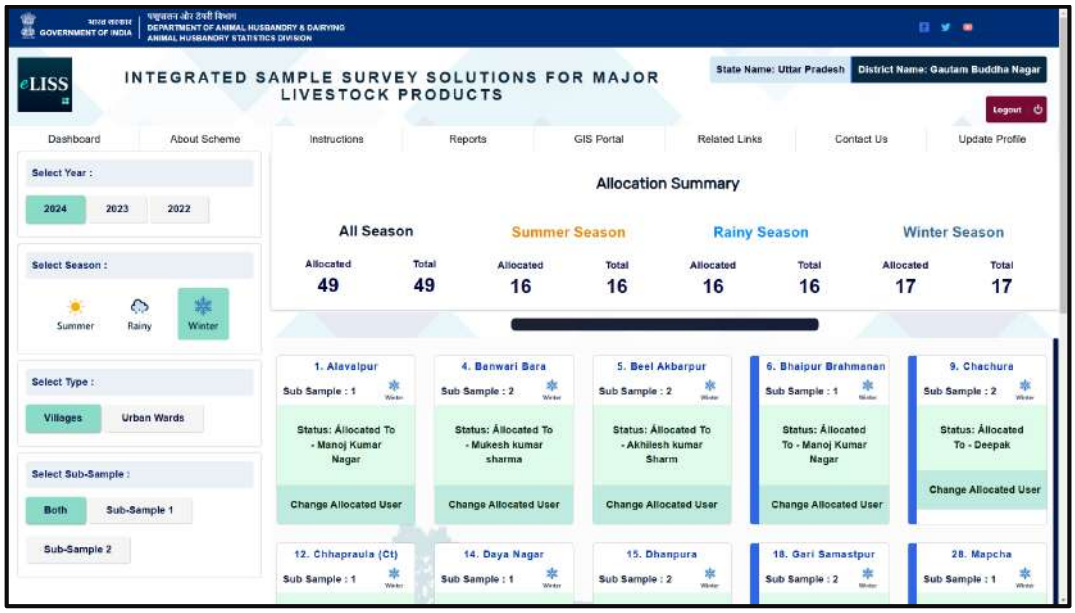
4.0 Allocation of villages/urban wards using eLISS web Portal

The selected sample of villages/urban wards for both complete enumeration and detailed survey are to be allocated to enumerators and supervisors.

Sample allocation module in the eLISS web portal is used for allocating the selected villages/urban wards among the registered and active enumerators and supervisors. DNOs are authorized to allocate the villages/urban wards which is a two-step process - first village/urban wards are allocated to the enumerators and then supervisors are allocated to the enumerators.

For performing the first step of allocation, DNOs have to visit sample allocation page from their dashboard to allocate the selected villages/urban wards to the enumerators. DNOs can allocate any number of village/urban ward to a single enumerator. A village/urban ward can be allocated to only one enumerator at a time. For performing the second step, DNOs have to select the supervisor for an enumerator from the dashboard page. DNOs can allocate any number of enumerators to a single supervisor but an enumerator can be allocated to only one supervisor at a time.

On allocating a supervisor to an enumerator, all the villages/urban wards allocated to that enumerator will be available to that supervisor. DNOs can change the supervisor for an enumerator at any point of time.



4.1 Re-Allocation for villages/urban wards

If any village/urban ward needs to be allocated to any other enumerator

In case the enumerator for any villages/urban ward needs to be changed, eLISS web portal allows DNOs to re-allocate any village/urban ward to a new enumerator during the ongoing survey. This re-allocation of villages/urban wards will change the authority of data collection from one enumerator to another. In addition to this, all the collected data of last enumerator will be transferred to the new enumerator.

For performing re-allocation, both the enumerators need to sync the data and logout from the app for avoiding any type of data loss. After re-allocation enumerators can login and continue the survey.

DNOs should carefully check the date and time of last sync and logout status of both the enumerators while performing re-allocation activity.

5.0 Monitoring of survey

eLISS End-to-End solution has a unique capability of real time monitoring of survey. AHS/State/UT/District can monitor the ongoing survey in respect of completion percentage of schedules, villages/urban wards, districts, states/UTs and field level survey status at various stages. The monitoring of survey is done at three different levels (i) Monitoring the completion of survey using eLISS Web Portal at AHS/State/UT level (ii) Monitoring the completion of survey using eLISS data collection app at State/UT level (iii) Monitoring the field level schedule-wise survey status of selected villages/urban wards at AHS/State/UT/ District level.

5.1 Monitoring the completion of survey using eLISS Web Portal at AHS/State/UT level

eLISS web portal allows AHS/States/UTs to monitor the completion of the survey at National level, separately for each State/UT and separately for each District within the States/UTs from the dashboard. The various stages at which the survey can be monitored are explained in the table below:

S No.	Stages	Schedule
1.	Allocation percentage for villages/urban wards selected for complete enumeration	Schedule-I and II
2.	Allocation percentage for villages/urban wards selected for detailed survey	Schedule-III, IV and VI
3.	Survey completion percentage for villages/urban wards selected for complete enumeration	Schedule-I and II
4.	Survey completion percentage for villages/urban wards selected for detailed survey	Schedule-III, IV and VI
5.	Survey completion percentage for commercial poultry farms which are operational for egg production	Schedule-V
6.	Survey completion percentage for commercial poultry farms which are operational for meat production	Schedule-VIII
7.	Survey completion percentage for slaughter houses	Schedule-VII

Note:

(1.) The completion percentage for complete enumeration includes completion of both schedule-I and schedule-II for a village/urban ward.

(2.) The completion percentage for detailed survey includes completion of schedule-III, schedule-IV and schedule-VI for a village/urban ward for a round.

5.2 Monitoring the completion of survey using eLISS data collection app at State/UT level

eLISS data collection app allows States/UTs to monitor the completion of the survey separately for their State/UT and for their respective Districts from the dashboard. The various stages at which the survey can be monitored are explained in the table below:

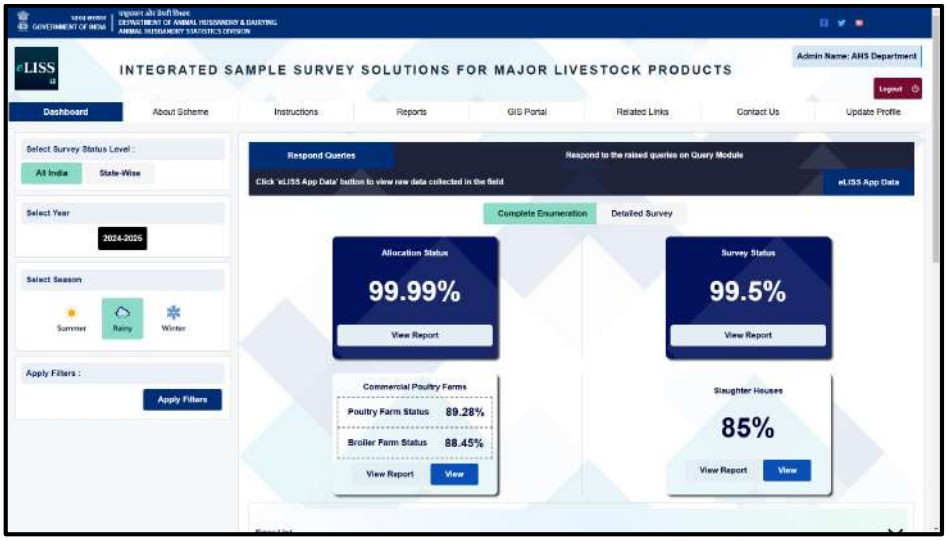
S No.	Stages	Schedule
1.	Allocation percentage for villages/urban wards selected for complete enumeration	Schedule-I and II
2.	Allocation percentage for villages/urban wards selected for detailed survey	Schedule-III, IV and VI
3.	Survey completion percentage for villages/urban wards selected for complete enumeration	Schedule-I and II
4.	Survey completion percentage for villages/urban wards selected for detailed survey	Schedule-III, IV and VI
5.	Survey completion percentage for commercial poultry farms which are operational for egg production	Schedule-V
6.	Survey completion percentage for commercial poultry farms which are operational for meat production	Schedule-VIII
7.	Survey completion percentage for slaughter houses	Schedule-VII

Note:

(1.) The completion percentage for complete enumeration includes completion of both schedule-I and schedule-II for a village/urban ward.

(2.) The completion percentage for detailed survey includes completion of schedule-III, schedule-IV and schedule-VI for a village/urban ward for a round.

In addition to this, State/UT officials can monitor the schedule-wise completion status for all eight ISS schedule for their respective districts. Further, eLISS data collection app provides option to State/UT official to contact DNOs on call and mail for timely completion of survey.



5.3 Monitoring the field level schedule-wise survey status of selected villages/urban wards at AHS/State/UT/ District level

eLISS web portal allows schedule-wise survey monitoring of villages/urban wards under the survey status web page which can be accessed from the dashboard of

AHS/State/UT/District officials. This module provides survey status at 10 stages for a village/urban ward for every schedule. These 10 stages consist of all the major milestones of the field survey and are shown in the table below.

Stage 1:	Enumerator not allocated	**Stage 6:**	Supervisor not allocated
Stage 2:	Survey not started	**Stage 7:**	Supervisor working
Stage 3:	Enumerator Working	**Stage 8:**	Supervisor completed
Stage 4:	Second stage sample not selected	**Stage 9:**	District working
Stage 5:	Enumerator completed	**Stage 10:**	Survey Completed

The status is updated in real time once the specific activities are performed by the users in field.

6.0 Field data collection using eLISS data collection app

The field data pertaining to milk, meat, egg and wool is collected using eight (08) schedules. Schedule-I and Schedule-II are used for generating the estimates for number of animals. Schedule-III, Schedule-IV and Schedule-VI are used for collecting the data on milk, egg and wool respectively and generating the estimates for these commodities. Schedule-V is used for collecting data and generating estimates for commercial egg production. Schedule-VII is used for collecting data and generating estimates for meat production from slaughter houses. Schedule-VIII is used for collecting data and generating estimates for meat production from commercial poultry farms.

The block-wise details of each schedules are explained below:

6.1 Schedule-I: General information

Schedule-I records general information of the villages/urban wards. It consists of two blocks which are listed below:

- **Block 1 -** General information about village selected for complete enumeration
- **Block 2 -** Particulars of the selected village

6.2 Schedule-II: Complete enumeration

Schedule-II records the data on number of livestock animals present in the household enterprises and non-household enterprises/institution by completely enumerating the selected villages/urban wards. It consists of three blocks which are listed below:

- **Block 1 -** General information about selected village /urban ward
- **Block 2 -** Household/Butchers Summary Detail which includes Number of livestock animals, Number of Animals Slaughtered during last 4 months
- **Block 3 -** Summary Detail of Milch Animals and Poultry in the sample

6.3 Schedule-III: Details of milk production at household/enterprise Level

Schedule-III records the data on milk yield produced, utilization of dung and average feed consumption in the selected households/enterprises in the selected villages/urban wards. It consists of six blocks which are listed below:

- **Block 1** - General information about selected village/urban ward
- **Block 2** – Selection of Household/Household Enterprises for detailed enquiry
- **Block 3** - Identification of selected H.H./Enterprise in the sub-sample of villages/ Urban Ward
- **Block 4** - Milk yield of individual animal on the day of visit
- **Block 5** - Details of average daily feed consumption during last 30 days including Details of utilization of Cow, Buffaloes and Goat milk produced on the previous day (kg.), Summary of total utilization milk from Cow, Buffaloes and Goats
- **Block 6** Utilization of dung Cattle, Buffaloes & Goats collected on the previous day

6.4 Schedule-IV: Details of egg production at household/enterprise Level

Schedule-IV records the data on Production, Purchase and Disposal of eggs in the selected households/enterprises in the selected villages/urban wards. It consists of two blocks which are listed below:

- **Block 1**: General information about selected village/urban ward
- **Block 2:** Production, Purchase and Disposal of eggs during the last 7days

6.5 Schedule-VI: Details of wool production at household/enterprise Level

Schedule-VI records the data on Disposal of sheep, Utilization of dung and Wool yield in the selected households/enterprises in the selected villages/urban wards. It consists of four blocks which are listed below:

- **Block 1:** General information about selected village/urban ward
- **Block 2:** Disposal of sheep during the last season
- **Block-3**: Utilization of dung
- **Block 4**: Wool yield of selected sheep

6.6 Schedule-V: Details of commercial poultry farms for egg production

Schedule-V records the data on yield rate of eggs laid by poultry birds from the Commercial Poultry Farms (Farms having 1000 poultry birds and more and Government Farms irrespective of their size). It consists of two blocks which are listed below:

- **Block 1:** General Information:
- **Block 2:** Details of poultry birds and its yield rate

6.7 Schedule-VII: Details of slaughter houses for meat production

Schedule-VII records the data on number of animals slaughtered and yield rate of slaughtered animals from the registered slaughter houses for last four months. It consists of three blocks which are listed below:

- **Block 1:** General Information about recognized slaughter houses
- **Block 2:** Details of number animal slaughtered on 1st day of month of the season
- **Block 3:** Details of yield rate of animals slaughtered on 1st day of month of the season

6.8 Schedule-VIII: Details of commercial poultry farms for meat production

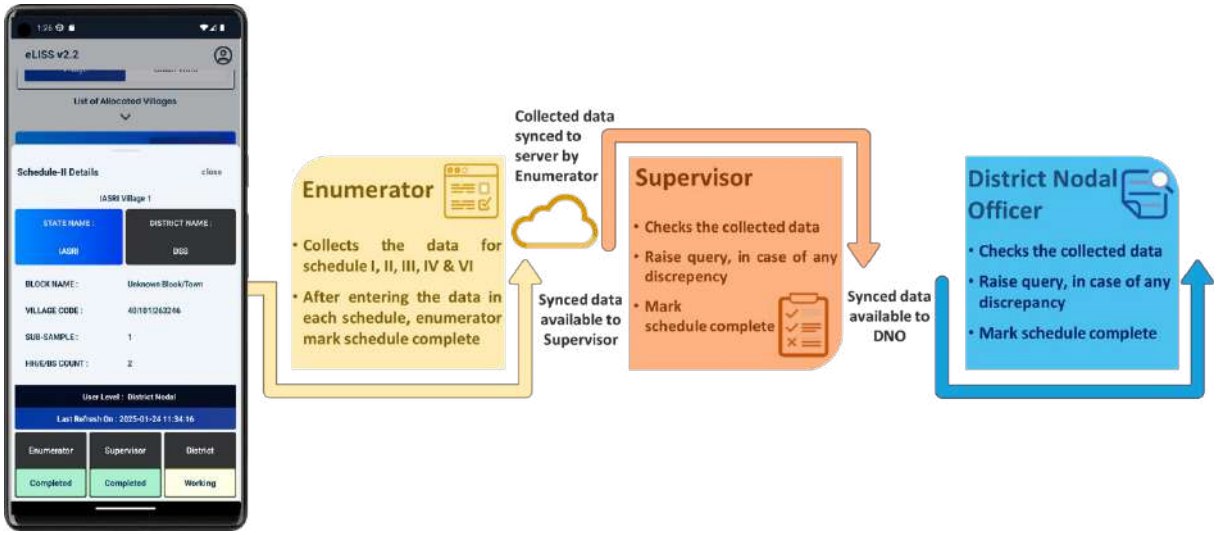
Schedule-VIII records the data on yield rate of meat from broilers and layers from the Commercial Poultry Farms (Farms having 1000 poultry birds and more and Government Farms irrespective of their size). It consists of two blocks which are listed below:

- **Block 1:** General information about commercial poultry farms:
- **Block 2:** Yield rate from broilers and layers on Commercial Poultry Farms

7.0 Supervision of collected data using eLISS data collection app

7.1 3-tier Data Supervision

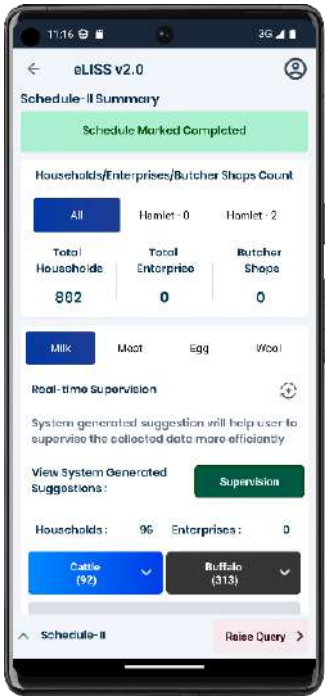
Ensuring high-quality data requires meticulous supervision at the field level, a task facilitated by the eLISS end-to-end solution, which employs a 3-tier data supervision process. Initially, data is collected through ISS schedules by enumerators using the eLISS data collection app. Upon syncing the collected data and marking the schedule as complete, it becomes accessible to supervisors. Supervisors meticulously review each schedule for every allocated village or urban ward, marking them complete for further processing. If any discrepancies are found, supervisors can raise queries, detailing the issue in 100 words and contacting enumerators via the app for clarification. Enumerators then have the opportunity to rectify the data and respond to queries. Supervisors may either raise further queries or mark the schedule as complete based on the enumerator's response. Once supervisors finalize schedules, the data is then reviewed by District Nodal Officers (DNOs) using a similar process. DNOs have the authority to mark schedules as complete, finalizing the data for the respective village or urban ward, with no further modifications permitted.

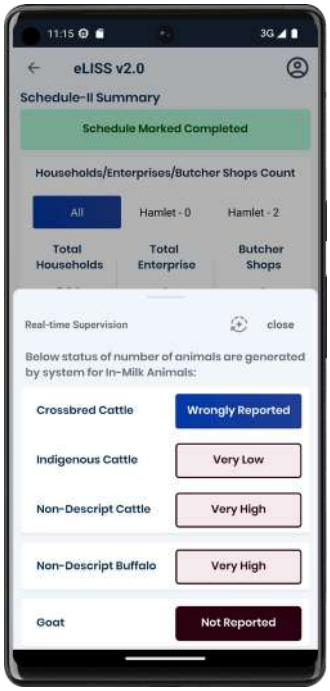


7.2 Real-Time System Based Supervision

In addition to 3-tier data supervision, a real-time system based supervision model (RTSBS V1.1) has been developed for the supervision of data collected to generate suggestions for Supervisors/DNOs based on the data input by the enumerators pertaining to the number of animals to be entered in Schedule II for all four commodities viz. milk, meat, egg and wool.

These suggestions are generated on the eLISS data collection app for supervisors and DNOs to supervise the collected data more efficiently once the enumerators mark the schedule complete. The suggestions are pertaining to the number of animals if entered are high, very high, low or very low. It also happens that a particular group of animals is present in the census but is missing in the survey or a particular group of animal is available in survey but is not reported in census, in such cases also the Supervisors/DNO will get suggestions to recheck.





S. No.	Generated Suggestion/ Comment	Details
1.	**Not Reported**	A particular group of animal is present in the census but not reported in the survey for a village/urban ward.
2.	**Wrongly Reported**	A particular group of animal is reported in the survey but not available in census for a village/urban ward.
3.	**Low**	If percentage of a particular group of animal reported in the survey is low as compared to the census.
4.	**Very Low**	If percentage of a particular group of animal reported in the survey is very low as compared to the census.
5.	**High**	If percentage of a particular group of animal reported in the survey is high as compared to the census.
6.	**Very High**	If percentage of a particular group of animal reported in the survey is very high as compared to the census.

8.0 Data cleaning/scrutiny

In order to improve the quality of data collected many validation checks have been applied in each schedule while recording the data. The details of which are given below:

8.1 Validation Checks

S No.	Types of Validation checks	Details
1.	Checks for Maintaining Consistency	For maintaining the consistency in data, many validation checks are implemented in each schedule which restrict user to enter same type of values in fields such as numeric, text, alphanumeric and decimal. For e.g. Yield values in schedule-III must be in decimal format. This type of validation checks maintains consistency over data for all States/UTs.
2.	Checks for Data Quality	For maintaining the data quality, many validation checks are implemented in each schedule which restrict user to record outliers/wrong data. For e.g. Value of number of sheep sheared captured in block-IV of schedule-VI cannot be more than the number of sheep present value captured in block-I of schedule-VI.
3.	Checks for Data Integrity	For maintaining the data integrity, many validation checks are implemented while syncing the data with the server. As user is allowed to sync the data any number of times, data integrity

		checks are implemented at API level for controlling duplicity of data and inconsistency of data at device and server level.

8.1.1 Some Major Validation Checks

Schedule-I

- If human population is more than 04 times of the number of households mentioned in schedule-I then a warning message will appear confirming for the same.

Schedule-II

- Schedule-II can be canvassed only after filling data for Schedule-I, as Hamlet information is required in-prior to listing of households/enterprises.
- In case no Hamlet groups are formed then at-least 80% of the number of households mentioned in the Schedule-I must be surveyed in Schedule II.
- Second Stage sample i.e. Households/Enterprises can be selected once Schedule II is marked as complete.

Schedule-III

- In Schedule-III, Block-III, if the day's total yield value (sum of morning, noon and evening yield values) for a group of animal (Exotic/Crossbred etc.) is more than the acceptable range then a warning message will appear for confirming the same.

8.2 Imputation

Imputation in Schedule-II: Complete Enumeration

If number of animal at sub-sample level is not available at census for a group (Exotic, Crossbred, Duck, Fowl, Ram, Ewe etc.) then to compute the ratio, survey value is imputed as census value which makes the ratio as '1'.

Imputation in Schedule-III: Milk, Schedule-IV: Egg and Schedule-VI: Wool:

If number of animals for a group (Exotic, Crossbred, Duck, Fowl, Ram, Ewe etc.) are present at district level (rural/urban sub-sample wise) and yield for that group (Exotic, crossbred etc.) is not available, then following cases arises where yield value is imputed as mentioned below:

Case 1:

Yield value for Rural, Sub-Sample–I is not available	Yield value for Rural, Sub-Sample–II is available	Impute Yield value of Rural, Sub-Sample–II
Yield value for Rural, Sub-Sample–I is available	Yield value for Rural, Sub-Sample–II is not available	Impute Yield value of Rural, Sub-Sample–I

Case 2:

Yield value for Urban , Sub-Sample–I is not available	Yield value for Urban , Sub-Sample–II is available	Impute Yield value of Urban, Sub-Sample–II
Yield value for Urban , Sub-Sample–I is available	Yield value for Urban , Sub-Sample–II is not available	Impute Yield Valueof Urban, Sub-Sample–I

Case 3:

Yield value for Rural, Sub-Sample–I is not available & Yield value for Rural, Sub-Sample–II is not available	Mean of Rural, Sub-Sample-I of all the Districts is imputed to obtain Yield value of Rural Sub-Sample- I.
	Mean of Rural, Sub-Sample-II of all the Districts is imputed to obtain Yield value of Rural Sub-Sample- II.

Case 4:

Yield value for Urban, Sub-Sample–I is not available & Yield value for Urban, Sub-Sample–II is not available	Mean of Urban, Sub-Sample-I of all the Districts is imputed to obtain Yield value of Urban Sub-Sample- I.
	Mean of Urban , Sub-Sample-II of all the Districts is imputed to obtain Yield value of Urban Sub-Sample- II.

9.0 Implementation of End-to-End Solution

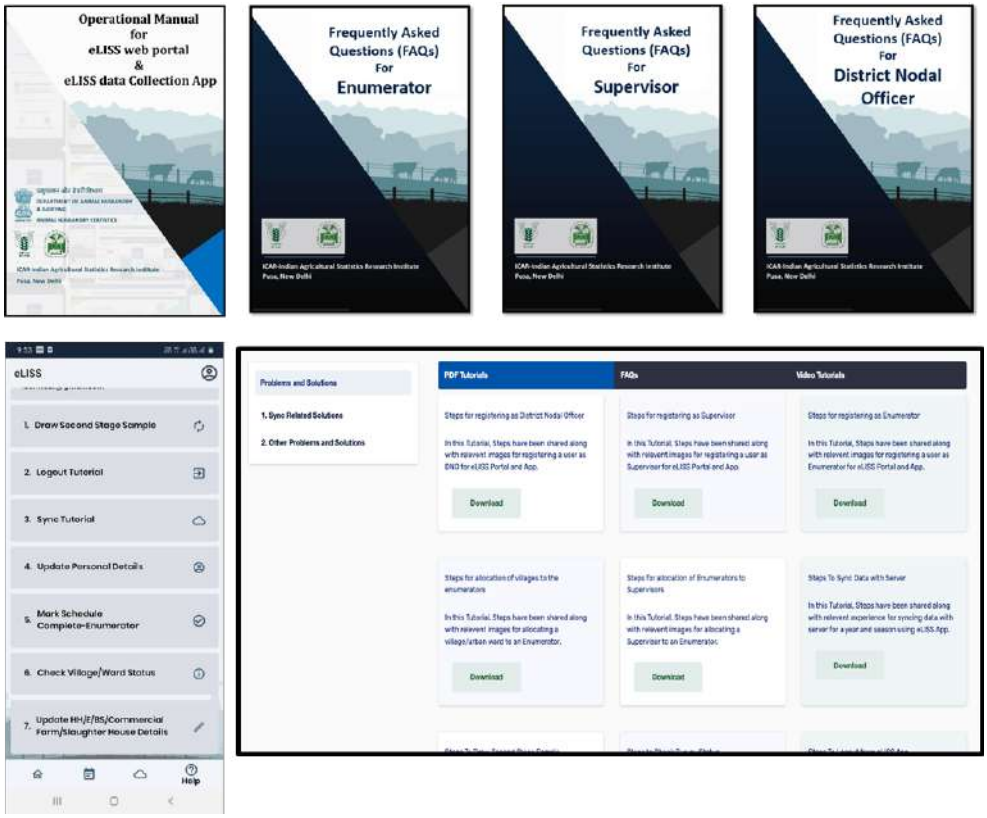
9.1 Pilot Studies

A pilot study is crucial in assessing the functionality, efficiency, and reliability of a system before full-scale implementation. The eLISS web portal and Data Collection App underwent pilot testing in two phases. In 2020, the modified eLISS Portal and App were tested in **three states**, evaluating all design elements, database connectivity, API performance, and data integrity. The collected data was synchronized with the server and analysed using a dedicated data analysis module. In 2022, another round of pilot testing was conducted with eLISS Data Collection App V1.4 across **10 states, covering 30 districts (three districts per state)**. This phase focused on validating the updates made to the app, ensuring smooth data synchronization, and generating reports shared with AHS and the states involved. The 10 states that participated in the 2022 pilot testing were **Andhra Pradesh, Chhattisgarh, Gujarat, Jammu & Kashmir, Karnataka, Mizoram, Punjab, Rajasthan, Sikkim, and Uttarakhand**. These pilot studies played a key role in refining the eLISS system for broader deployment.

9.2 Support Material

For providing implementation support, various materials were provided to ensure a smooth transition and effective usage of the eLISS web portal and Data Collection App. The support materials included **step-wise PDF tutorials and an operational manual** to guide users through different processes. Additionally, **video tutorials** were made available to

demonstrate major activities on the portal. To address common concerns, **FAQs** were provided for Enumerators, Supervisors, and District Officers. Furthermore, an **In-App Help** section was integrated, offering step-by-step instructions for various activities directly within the application. These resources collectively ensured that users had access to comprehensive guidance, facilitating efficient implementation and usage.



9.3 Trainings

For implementing the developed end-to-end solution various online and offline trainings were conducted. The training modules were meticulously formulated to cover all important topics such as data collection methodologies, real-time data supervision, and techniques for improving overall estimates of number of animals and production. Detailed plan for conducting zone-wise training sessions were established, encompassing both online and offline formats to accommodate diverse geographical and infrastructural challenges.

Online Training Sessions

S. No.	Online Session	Date	No of Participants
1.	All India Training of Master Trainers on Web Portal and Android App for ISS Scheme	February 17-18, 2021	274
2.	Online Refresher Training for States/UTs in West & Central Zone	September 22, 2021	560
3.	Online Refresher Training for States/UTs in South Zone	September 23, 2021	360
4.	Online Refresher Training for States/UTs in North East Zone	September 24, 2021	180
5.	Online Refresher Training for States/UTs in East Zone	September 27, 2021	1005
6.	Online Refresher Training for States/UTs in North Zone	September 28, 2021	634
7.	Online Refresher Training for Puducherry & Lakshadweep	September 21, 2023	50
8.	Online Training on “eLISS data collection app V2.0”	March 23, 2024	1003
9.	Training on eLISS Web Portal and eLISS App	June 04, 2024	15

Offline Training Sessions

S. No.	Zone	Date	No of Participants	No. of States/UTs & Name of States/UTs	
1.	South Zone	February 02-03, 2023	167	1. Karnataka 2. Lakshadweep 3. Tamil Nadu 4. Telangana	5. Andhra Pradesh 6. Kerala 7. A & N Island 8. Puducherry
2.	North Central Zone	February 08-09, 2023	150	1. Uttar Pradesh	2. Madhya Pradesh
3.	West Zone	February 16-17, 2023	109	1. Gujarat 2. Maharashtra	3. Rajasthan 4. Dadar & Nagar Haveli & DD
4.	East Zone	February 23-24, 2023	153	1. Bihar 2. Chattisgarh 3. Jharkhand	4. Odisha 5. West Bengal
5.	North Zone	February 28-March 01, 2023	140	1. Haryana 2. Himachal Pradesh 3. Punjab 4. Uttarakhand	5. Jammu & Kashmir 6. Ladakh 7. Delhi
6.	North Eastern Zone	April 27-28, 2023	135	1. Assam 2. Arunachal Pradesh 3. Sikkim 4. Tripura	5. Meghalaya 6. Nagaland 7. Mizoram 8. Manipur
7.	North Zone	June 06, 2024	140	1. Jammu & Kashmir 2. Ladakh 3. Himachal Pradesh	

9.4 Technical Support

To ensure seamless operation and user assistance, **technical support for the eLISS end-to-end solution** is provided through multiple channels, including a **helpmail ID, zonal WhatsApp groups, and an Automated Query Module (AQM)**. The AQM is a **user-friendly system** designed to assist users at the **District, Supervisor, and Enumerator levels**. It allows users to **raise queries directly through the app and receive immediate responses**, ensuring the smooth functioning of data collection processes. The module provides structured query submission by enabling users to **select relevant tags, sub-tags,**

schedules, and rounds before entering a detailed message explaining their issue. This streamlined approach enhances efficiency in addressing technical concerns, minimizing disruptions in data collection and ensuring timely resolution of user queries.

Email Support	Whatsapp Support	Automated Query Module
helpmail.iss@gmail.com	Whatsapp Zonal Groups	Real Time Query Module in the help section of eLISS Data Collection app for quick response to the queries

Automated Query Module

To ensure the smooth functioning of eLISS data collection app, technical support is provided at District/ Supervisor/ Enumerator level using the automated query module. It is a user friendly module which allows users to raise the query and receive response to their queries immediately through eLISS data collection app. In this query module, users have options to select tag, sub-tag, schedule, round for which query needs to be raised. Further, users need to enter a detailed message explaining their query.

A unique ticket number is generated for each query which allows users to track the status of the query on the app. When a user receives the first response, the status of the query becomes "Response". Further, users have an option to mark whether the query is resolved with the provided response or not. In case, the user reports that the query is not resolved then another response would be provided and the status of the query changes to "In-Process". Further, when a new response to the query is provided then a notification appears on the home screen and the status of the query is updated to "Response". If the query is resolved by the response provided, users have an option to mark the query resolved which changes the status to "Resolved". If the users need to share the raised query on whatsapp then it can be done by using the share option provided in the Automated Query Module.

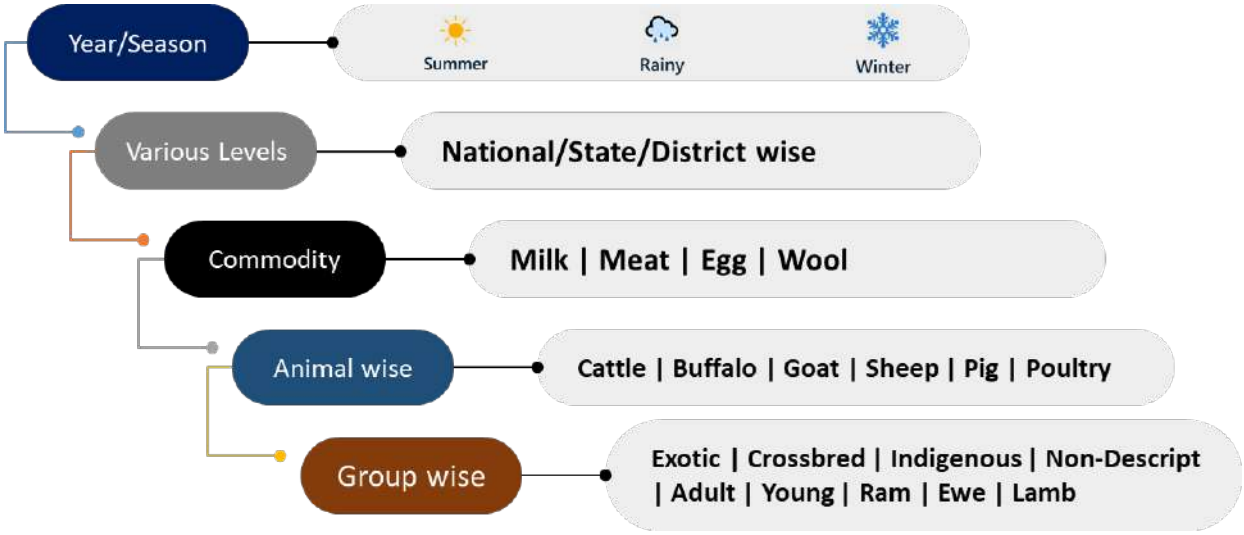
9.5 Released versions

Several versions of the **eLISS Data Collection App** have been released over time, incorporating improvements in functionality, user experience, and data quality. Each update introduced enhancements such as improved UI/UX design, better data validation mechanisms, and optimized data collection workflows. Key improvements included **restrictions to ensure structured data entry**, enabling one village or ward to be filled at a time and requiring Schedule-I to be completed before Schedule-II. Additional updates focused on **automating data entry for efficiency, adding over 100 new validation checks, and making certain fields optional for faster data collection**. Furthermore, a **new sync and refresh system architecture using microservices** was implemented, along with improvements to **memory and network issue handling**, ensuring a seamless and efficient data collection process.

S No.	Major eLISS released versions	Released Date
1.	eLISS Version V1.1	**12 July 2021**
2.	eLISS Version V1.2	**14 August 2021**
3.	eLISS Version V1.21	**23 October 2021**
4.	eLISS Version V1.3	**22 April 2022**
5.	eLISS Version V1.4	**25 December 2022**
6.	eLISS Version V1.5	**01 July 2023**
7.	eLISS Version V2.0	**1 March 2024**
8.	eLISS Version V2.1	**23 March 2024**
9.	eLISS Version V2.2	**12 July 2024**

10. Generation of Livestock statistics and preparation of reports using eLISS web portal

eLISS end-to-end solution consists of data analysis module which analyse the unit level data collected through eLISS data collection app for generating estimates for number of animals, average yield and production pertaining to milk, meat, egg and wool. Estimates are generated at various levels for all three seasons individually and then pooled for a year. eLISS End-to-End solution uses ISS estimation procedure for generating the estimates.



Preparation of reports using eLISS web portal

Various reports will be generated using the analysed data pertaining to milk, meat, egg and wool.

SURVEYS FOR ESTIMATION OF AREA AND PRODUCTION OF HORTICULTURAL CROP

Tauqueer Ahmad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1. Introduction

In recent years, horticulture sector has emerged as an important component of the Indian economy. This sector of agriculture contributes about one-fifth share in the economy of agriculture and allied sectors. Hence, the statistics of horticultural crops has become one of the priority programmes for the planning commission. Horticulture development is being given high priority in the Five Year Plans. For preparation of various developmental programmes and for policy formulations etc., the availability of adequate, reliable and timely statistics on area, yield and production estimates of horticultural crops is essential. At present, a scheme namely "Crop Estimation Survey on Fruits and Vegetables" is being implemented under Directorate of Economics & Statistics (DES), Ministry of Agriculture. This scheme has so far been implemented only in 11 States namely Andhra Pradesh, Gujarat, Haryana, Himachal Pradesh, Karnataka, Maharashtra, Orissa, Punjab, Rajasthan, Tamil Nadu and Uttar Pradesh and covers only 7 fruits and 7 vegetables namely, Mango, Apple, Banana, Grapes, Guava, Citrus, Pineapple, Cauliflower, Potato, Onion, Tomato, Cabbage, Ginger and Turmeric. For estimation of area and production of fruit and vegetable crops, methodology developed by Indian Agricultural Statistics Research institute (IASRI) is being used in these States. Here, present method of generation of data, existing methodology being used and an alternative methodology for estimation of area and production of different horticultural crops have been discussed in brief.

IASRI carried a series of surveys to evolve a sampling methodology for estimating area and yields of fruits and vegetables. A study on vegetables was conducted in rural areas of Delhi by Sukhatme *et al.* (1969). Singh *et al.* (1976) conducted a study on fresh fruits in Tamil Nadu and on vegetables in Bangalore district of Karnataka. Problems and issues related to Statistics of Horticultural Crops were discussed in detail in the Symposium organized during ISAS Conference in 2001. Details of the sampling methodology for estimation of extent of cultivation and production of fruit and vegetable crops are given below:

2. Sampling Methodology for Estimation of Extent of Cultivation and Production of Fruit and Vegetable Crops

2.1. Fruit crops

In view of the special features of fruit crops, estimation of extent of cultivation and production of fruit crops is somewhat different than other annual crops. Some of the features are:-

i. As against seasonal nature of field crops, fruits are perennial crops.

ii. Fruit trees, besides being grown in regular orchards, are also extensively grown on canal banks, field bunds, road sides, back yard of houses and even as stray trees.

iii. Different fruits are frequently grown in the same orchard.

iv. Fruit trees take quite a few years before they start bearing fruit.

v. All the trees in an orchard may not be of the same age i.e. an orchard may contain both bearing and young trees.

vi. Harvesting of fruits trees is done in a number of pickings extending over several weeks.

vii. Several fruits like citrus, guava etc. have two harvesting seasons in a year.

All these features are to be carefully considered while planning a sample survey to estimate the extent of cultivation and yield of fruits.

Unlike other crops, extent of cultivation of a fruit may be measured in terms of area under the crop or by the number of trees both bearing as well as young. However, only bearing trees contribute towards the production of the fruit. The number of young trees on the other hand provide an idea about the extent of cultivation of the crops in the future.

The choice of sampling design would depend upon whether only one fruit is of interest or more than one fruits are being studied. Normally, the survey may be planned to cover all important fruit crops simultaneously at the State level. However, if single fruit is to be covered for some specified area, say the district level, on the basis of importance of the crops, the sampling design for such surveys may be used. Accordingly, the sampling design for single fruit in a district and for several fruit crops at the State level are separately described below:

2.1.1 Sampling plan for surveys to estimate the extent of cultivation and production of a single fruit crop in a district

Each village in the district may be identified as “reporting or “non-reporting” for the crops on the basis whether the fruit is grown in the village or not. A list of “reporting” as well as “non-reporting” villages may be prepared along with area under the fruit. This information may be obtained from revenue records or from past years data.

2.1.1.1 Sampling design

The sampling design may be broadly defined as stratified three stage random sampling. The tehsils/taluks/blocks or groups thereof in the district may be taken as strata, villages as primary sampling units (psu’s), orchards as second stage units and clusters of trees as the ultimate units of sampling. The sample size of villages i.e. the number of villages to be selected in the district may be allocated to different strata in proportion to the area under the fruit in the strata. The “reporting” villages in a stratum may be regarded as p.s.u.’s and selection of allocated or the desired number of villages may be done by probability proportional to size (pps) with replacement, taking area under the fruit as the size measure.

Orchards in the selected villages and cluster of trees in the orchards are then selected with SRSWOR. Also since there may be errors in the reporting/recording of fruit cultivation or some fruit cultivation may be taken up in the "non-reporting" villages, a sample of villages may also be selected from the "non-reporting" groups of villages in each stratum. For determining the extent of cultivation, the selected villages may be completely enumerated to obtain information on the area under fruit orchards and the number of trees both in the orchards as well as stray tree. The trees may also be enumerated with respect to the varieties as well as status about bearing or non-bearing fruits. Apart from estimation of extent of cultivation of fruit complete enumeration would also provide a frame of orchards for further selection of orchards and trees for estimation of yield. For estimation of yield of fruit, five orchards may be selected by SRSWOR to record information regarding cultivation practices such as irrigation, manuring, inter-cropping and other practices followed by the cultivators throughout the year. From each of the selected orchards, three clusters of four trees each of bearing age may be selected at random for recording data on yield of a fruit throughout the harvesting season.

2.1.1.2 Sample size

A total of 150-200 reporting villages (psu's) may be selected in the district. As described above, this number may be allocated to different strata (tehsils) in proportion to area under orchards and the allocated number of villages in a stratum may be selected with pps with replacement. At the second stage of sampling 5 orchards may be selected at random and from each selected orchard, three clusters of 4 bearing trees may be selected at the ultimate stage of sampling. Earlier surveys have shown that with this type of design and sample size, the average yield at the district level is likely to be estimated with a Standard Error(S.E.) of about 5% and the area and total production with a S.E. between 5 to 10%. However, the efficiencies of various estimators would depend upon the amount of variability in different characters. Surveys conducted during initial years will provide an idea about these variability and accordingly the number of villages and orchards selected may be modified to achieve the desired degree of precision.

2.1.2 Sampling plan for estimation of extent of cultivation and production of more than one fruit crops in a state:

The important fruit crops whose production is to be estimated should be identified first. Normally, the previous years' area figures under different fruit crops are available at the tehsil/taluk level and these may be used to determine the important fruits in the State. Since the cultivation of fruits is usually not so evenly spread and may in fact be concentrated in a few districts/regions, the first step in the planning of fruit survey is to identify and delimit the important fruit growing regions or areas for different fruits. A district is considered too large a unit of area for this purpose. However, taluks or sub-divisions or equivalent areas in a district may be considered appropriate. Thus, taluks which are important at least for one of the fruit crops, may be identified as important fruit growing taluks. It may be mentioned that importance of a taluk with respect to a fruit is determined on the basis of area under that fruit and thus a taluk important for a given fruit may not be important for other fruits. As a broad guideline, for a given fruit, the important taluks are those which taken together cover 40-50% of the total area under that fruit in the entire State.

2.1.2.1 Sampling design and sample size

All taluks/sub-divisions, considered important fruit growing areas as described above, may be taken as strata. The remaining area or taluks may be further classified or grouped into 4 to 5 strata with respect to importance of individual fruit crops taking into account the geographical contiguity. In these strata, taluks may be considered as primary sampling units. Thus survey would then cover all important fruit growing taluks i.e. taluks in which fruit cultivation is concentrated as well as the selected taluks out of the rest.

In the selected taluks also, all the villages may not be growing all the fruits. A frame of villages growing different fruit in a stratum is, therefore, prepared. Accordingly, villages in a stratum may be classified into two categories (i) growing at least one fruit and (ii) growing no fruit at all. In category (i) on the basis of village-wise area under fruits, villages may be identified as 'reporting' or "non-reporting" for individual fruits. If the reported areas are considered as reliable, efforts may be concentrated only in the reporting villages for each fruit. However, experience shows that faculty reporting is not uncommon and therefore, adequate representation may be given to non-reporting group. From the reporting group of villages for a given fruit crop four villages may be selected with replacement and with probability proportional to area reported under the fruit crop. From the non-reporting group of villages (in which other fruits are grown), a sample of two villages may be selected in each stratum by SRSWOR. From the villages in category (ii) where no cultivation of fruits is reported, a sample of two villages may be selected by SRSWOR. The selected villages may be completely enumerated for the extent of cultivation and number of trees in orchards and also the stray trees.

For yield estimation, a sub sample of two villages out of four reporting villages may be retained in all the major fruit growing taluks/strata and from each village 5 orchards and 3 clusters of 4 trees each of bearing age may be selected for this purpose. The selected clusters of trees may be observed for entire harvest period both with respect to weight as well as number of fruit. However, exceptions to this procedure may be made for certain crops like banana and grapes. A uniform approach in this regard is essential for comparability as well as pooling of estimates over different areas.

2.2 Vegetable crops

The survey approach for estimation of area and production of vegetable crops is somewhat more complex due to special feature of cultivation of these crops. Some of these features are as follows:

i. The vegetables are short duration crops and their duration varies considerably from one vegetable to the other.

ii. Harvesting of vegetables involves a number of pickings

iii. Vegetable cultivation is more or less a continuous process with various operations like sowing, harvesting, etc. being done simultaneously in different fields of a village.

iv. Vegetables are highly sensitive crops and this normally adds to the variability in the yield rates of the crops.

It is also realized that due to perishable nature of the vegetable crops, production depends on availability of marketing facilities in the area. This is why cultivation of vegetables is normally concentrated around bigger town and cities. Accordingly, the methodology for estimation of area and production of vegetable crops has been developed at the district level in different surveys conducted so far in various States.

The sampling design for surveys for estimation of area and production of vegetables is described below:

2.2.1 Sampling design

The sampling design is a stratified multistage random sampling. Taluks or equivalent areas may be taken as main strata. Further, since area under vegetables may vary considerably from one village to another in a taluk, sub-stratification may be done on the basis of village-wise area under vegetables. For this purpose 3 to 4 substrata with equal area under vegetables may be formed. The data figures may be available in revenue records. If not available, then a preliminary survey may be conducted to obtain village wise area under vegetables. Within the strata, clusters of three villages may be taken as primary sampling units. For determining the extent of cultivation, a sampling fraction of about 20% may be used for selection of clusters of villages. The allocation of clusters of villages to different strata may be done in proportion to area under vegetables. The allocated number of clusters in different strata may be selected with simple random sampling without replacement (SRSWOR). For yield study, 50% of the clusters selected for area may be retained and fields growing vegetables may be selected in these clusters. The selected clusters of villages may be completely enumerated for area under vegetables. Vegetables being short duration crops, one time enumeration in a year may not be meaningful. To account for the short duration of crops and early and late varieties, a year may be divided into four periods of three months each. The area enumeration may be done in the beginning of each period. This will also provide a frame of vegetable fields for estimation of yield rates. For estimation of production, 6 to 8 fields of each important vegetables may be selected in each of the clusters selected for yield study. In each of the selected fields, a randomly located plot of 5m x 5m may be demarcated and observed for all the pickings in the respective periods. The yield of a vegetable for a selected field is obtained as the aggregate of all pickings in the period obtained from the c.c. plot. The average yield of the vegetable for the village is obtained as a simple mean of field wise yield and when multiplied by the area under vegetable in the village gives the vegetable production in the village. In this way the production for each period may be estimated separately. The average yield is then obtained from the estimated production and the area under a vegetable.

This sampling design is likely to provide estimates of average yield with less than 5% standard error and the area and production with less than 10% standard error for important vegetable crops at the district level.

3. Present Status

Country level estimates are developed only for important fruit and vegetable crops under central scheme "Crop Estimation Survey on Fruits and Vegetables". Although this scheme is in operation for the last several years, coverage in terms of fruits and vegetables as well as in terms of geographical coverage of the country, it is grossly inadequate. Since DES releases the estimates for only eleven States, users are depending on other sources of statistics. National Horticulture Board (NHB) of the Ministry of Agriculture is bringing out the publication entitled "Indian Horticulture Database". This publication contains recent data on area, production and prices of various horticultural produce. NHB data is relatively more comprehensive by way of covering data for most of the States, and is generally used for the purpose of GDP estimates etc. The estimates released by NHB are not based on sound statistical methodology and need to be examined.

It is evident from the above that reliable and timely estimates of area and production at all India level for these crops are not available from any source. Moreover, there are problems in the implementation of the pilot scheme on fruits and vegetables because of various reasons.

4. Recommendations of National Statistical Commission (NSC)

National Statistical Commission (NSC) - 2001 in its recommendation (refer 4.5.7 on page 122) has suggested that:

- "The methodology adopted in the pilot scheme of ***Crop Estimation Survey on Fruits and Vegetables*** should be reviewed and an alternative methodology for estimating the production of horticultural crops should be developed taking into account information flowing from all sources including market arrivals, exports and growers associations.

- Special studies required to establish the feasibility of such a methodology should be taken up by a team comprising representatives from Indian Agricultural Statistics Research Institute (IASRI), Directorate of Economics and Statistics (DES), Ministry of Agriculture (MoA), Field Operations Division (FOD) of National Sample Survey Organization (NSSO) and from one or two major states growing horticultural crops.

- The alternative methodology should be tried out on a pilot basis before actually implementing it on a large scale".

5. Alternative Methodology

In view of NSC recommendation, a Brainstorming Session on "Alternative Methodology for Estimation of Production of Horticultural Crops" was held on 24th November 2003 at IASRI, New Delhi under the Chairmanship of Economic and Statistical Advisor to discuss the issues regarding strengthening of horticulture database.

Accordingly, a project proposal entitled "Pilot Study to Develop an Alternative Methodology for Estimation of Area and Production of Horticultural Crops" after incorporating the suggestions made during the brainstorming session was submitted by IASRI to DES, MOA for funding. The project was finally funded by Ministry of Statistics and Programme

Implementation and was completed at IASRI. This study was conducted in two States namely, Maharastra and Himachal Pradesh covering important fruits and vegetables.

5.1 Sampling design and sample size

First of all, important districts were identified for conducting survey on the basis of district-wise area figures under fruits and vegetables of the state. As a broad guideline, the important districts are those which taken together cover 70-80% of the total area under fruits and vegetables in the entire state. The sampling design which was adopted for the survey may be described as stratified multistage random sampling. Taluk-wise area figures under fruits and vegetables were used for stratifying the taluks of the selected districts into two groups viz. high productive taluks and low productive taluks. High productive taluks are those which constitute 60-70 percent of the total area under fruits and vegetables of the district and rest of the taluks will fall under low productive taluks. A sample of two taluks was selected by simple random sampling without replacement (SRSWOR) from both the groups after rejecting taluks contributing less than 5% of total area under fruits and vegetables of the district.

From each of the four selected taluks, a sample of five villages was selected by SRSWOR. The selected villages were completely enumerated so as to record number of orchards under different fruits and cropping pattern with respect to vegetables. An orchard for selection process should have minimum of 12 fruit trees of bearing age of a single fruit crop. For fruits survey, a sample of five orchards was selected from each selected village by SRSWOR. In case, there are more than one fruit crop available in the village then orchards were selected in proportion to the number of orchards for two major fruit crops in each of the village with a minimum of two orchards for each fruit crop. Major fruit crops were decided on the basis of number of orchards of different fruits available in the village. From each selected orchard, a sample of three clusters each consisting of four trees of bearing age was selected randomly out of the total number of trees of bearing age. The yield of selected trees was collected through enquiry and yield of any four trees was collected through physical observation.

For vegetable survey, a sample of 10 vegetable growers was selected out of qualified vegetable growers of a village. For this, after complete enumeration of selected village, a list of qualified vegetable growers was prepared. Qualified growers are those vegetable growers who have 0.1 ha and above gross cropped area under vegetables in case of Maharashtra and 0.01 ha and above in case of Himachal Pradesh. Ranking of qualified vegetable growers was done as per grass cropped area and then qualified vegetable growers were divided into two groups after ranking. If number of growers is odd, the first group will have one more grower than the second group. A total of six vegetable growers were selected from the first group and rest four from the second group. In case total number of qualified vegetable growers in any village is less than or equal to ten, all the growers were selected for detailed survey enquiry. The produce of all the vegetables crops grown by the selected vegetable grower was recorded through enquiry and physical observation was taken on the day of visit. The FI must get in touch with the grower of the selected field from time to time and ascertain the date of harvest. He must be present on the day of harvest. He must locate the experimental plot of specified size (5mx5m) before the cultivator starts harvesting the field. In each selected field,

the experimental plot of the specified size must be located at random beginning with South-West corner of the selected field.

6. Project Implementation

The major steps followed for the successful completion of the study are as under:

- Data collection work was done in two phases i.e. complete enumeration for three months and detailed survey for one year with the help of Commissionerate of Agriculture, Pune and Directorate of Land Records, Shimla. Seventeen (17) Field Investigators (FIs) were hired in Maharashtra and eight(8) were hired in H.P. with the help of respective State Govts.

- Schedules along with instruction manual were prepared for data collection. Training for data collection was provided to the hired FIs in both the States in two phases.

- Supervision of data collection was done by the IASRI officials at regular intervals and all necessary support for this purpose was provided by the State Govts.

- The data was coded and entered. The entered data was scrutinized.

- Estimation procedures were developed for area, production and productivity of fruits as well as vegetables.

- Data analysis was done and estimates of area, production and productivity of important fruits and vegetables were obtained for surveyed districts of both the states.

- The district-wise market arrival data for Maharashtra and H.P. for last ten years for important fruits and vegetables under the study was obtained from Maharashtra Agricultural Marketing Board, Pune and from H.P. Agricultural Marketing Board, Shimla respectively.

- The district-wise data for last ten years pertaining to area, production and productivity of important fruits and vegetables in Maharashtra and H.P. was also obtained from the respective State departments.

- Regression Analysis was done for each district of the State for each major fruit and vegetable crop using total production and market arrival data for the last 10 years.

- The total production of important fruits and vegetables for non surveyed districts of both the states for the year 2006-2007 was predicted using the market arrival data for the year 2006-2007.

- State level estimates were obtained by pooling the estimates of surveyed and non surveyed districts.

7. Salient Findings

The salient achievements of the study are as under:

- It is worth mentioning that based on the present study the recommended sample size i.e. no. of villages to be selected from a district is 80. Hence, there is a decrease in sample size i.e. from 150-200 villages per district to 80 villages per district.

- The survey procedures under the alternative methodology have been simplified. These are cost effective and less time consuming.

- The alternative methodology provides estimates for more than one fruit/vegetable at district level, whereas, the methodology used under CES-F&V provides estimates for a single fruit/vegetable at district level.

- The developed methodology is simple and easy to implement both for fruits and vegetables.

- The estimates of no. of stray trees have also been obtained and used in finding out estimates of total production of fruit crops.

- The correlation coefficient obtained between production figure and market arrival data is high.

- Market arrival data has been used for obtaining State level estimates.

- The percentage standard errors of the estimates obtained for fruits are between 5 to 20 and for vegetables are between 5 to 30 percent at district level based on sample size of 20 i.e. 20 villages per district as per the proposed sampling design.

- It is expected that at district level, the total production of important fruits can be estimated with less than 10 percent standard error and the total production of important vegetables can be estimated with less than 15 percent standard error at 95 percent confidence interval, if 80 villages are selected from each selected district.

- The estimated sample size for obtaining less than 15 percent standard error of the estimates of total production in case of fruits and less than 20 percent standard error in case of vegetables at 95 percent confidence interval is 43 i.e. 43 villages per district.

8. Concluding Remarks

This study is the first pilot study in the direction of developing alternative methodology. The study has revealed very encouraging results as shown above and demonstrated the feasibility of estimating the production of fruits and vegetables with much smaller sample size. In view of above, it is recommended that the methodology needs to be tested in few states before it is implemented at large scale. It may be indicated here that the National Statistical Commission has also recommended that the alternative methodology should be tried out on a pilot basis before actually implementing it on a large scale.

Once the alternative methodology is tested, the developed methodology will bridge the data gaps in official statistics related to horticultural crops. Once the methodology is developed, the greater commitment and willingness of all concerned to strictly comply with prescribed procedures and time schedules will bring about remarkable improvements in the system.

9. IASRI component of CHAMAN program under MIDH

The ICAR-IASRI has been declared as National Level Agency (NLA) under Mission for Integrated Development of Horticulture (MIDH) for taking up a project entitled "Study to test the developed alternative methodology for estimation of area and production of horticultural crops: IASRI component of CHAMAN program under MIDH" funded by Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Govt. of India. Testing and validation of the methodology for estimation of area and production of horticultural crops developed by ICAR-IASRI is being carried out in five states of the country namely, Andhra Pradesh, Tamil Nadu, Maharashtra, Himachal Pradesh and Karnataka with the help of Directorate of Horticulture of the respective State Govts. under Coordinated Programme on Horticulture Assessment and Management using Geoinformatics (CHAMAN). The developed methodology is also being tested and validated in two more states namely, Madhya Pradesh (M.P.) and Haryana from respective State Govts. own funding. Primary data collection is being carried out by hired Field Investigators (FIs) through Service Providers by the State Govts. except M.P. and Tamil Nadu where State Govts. nominated staff are engaged.

The objectives of the study are:

- To test the developed alternative methodology for estimation of acreage under each major fruit and vegetable crops
- To test the developed alternative methodology for estimation of yield rates and total production of major fruit and vegetable crops grown in the State
- To validate the accuracy of estimates of area under major fruits and vegetables using remote sensing techniques with the area estimates using complete enumeration

Implementation Plan and Status:

- First Year (2014-15) : Initial preparatory works by IASRI and states for the survey - completed
- Second Year (2015-16): Conduct of the field work in all the 7 states for area and production assessment - field work in progress in 6 states
- Third Year (2016-17): Field work to be continued, validation and analysis of data collected and submission of report - field work to be initiated in Karnataka State and to be continued in rest of the 6 states till December 2016/June 2017.

The proposed sampling design adopted for the survey is stratified multistage random sampling. Selection of blocks/ mandals/ taluks and villages within the selected blocks/ mandals/ taluks was carried out as per the proposed sampling design. Request for Proposal (RFP) document for hiring Field Investigators in the states through an agency, Schedules and Instruction Manuals (English and Hindi version) for primary data collection were prepared

and sent to the states along with List of selected blocks/mandals/taluks and villages for all the selected districts.

Class room training as well as Field training for filling up schedules and different methods of conducting Crop Cutting Experiments in case of horticultural crops were imparted to the hired FIs/nominated staff as well as Master trainers (State officials) by ICAR-IASRI officials in six states. The data collection work is in progress in 6 states. Field work will be initiated in Karnataka State shortly due to delay in hiring FIs by the State. Supervision of data collection was done by the ICAR-IASRI officials in all the states under study. Infrastructure report on Organizational structure and man power requirements, their role and functions along with budget requirement for implementation of this methodology in all the states of the country from 2018-2019 onwards was prepared and submitted to the funding agency. As desired by Union Agriculture Minister, Honorable Shri Radha Mohan Singh ji, the progress of the project till date was presented before him on 20th July, 2016 and 4th October 2017 at Krishi Bhawan, New Delhi.

The schedules developed for primary data collection are being designed using CAPI (Computer Assisted Personal Interviewing) designer. Data entry software has been developed and is being tested. Development of data analysis software is in progress.

Under this study, primary data collection was carried out by hired Field Investigators (FIs) through Service Providers by the respective State Govts. except Tamil Nadu where State Govts. nominated staff were engaged.

Data collection was carried out in two phases:

Phase-1 (Area enumeration) and

Phase-2 (Detailed survey for yield data collection).

Field work for both the phases has been completed in all the four states under study. Data entry and scrutiny of data is in progress in Andhra Pradesh, Tamil Nadu and Maharashtra and Himachal Pradesh.

10. Integrated approach using Remote sensing and GIS

Unfortunately, Remote Sensing and GIS, which are very potential tools for estimation of area and production, at least for major fruits and vegetables in the districts were not included in the previous study due to fund constraint. It is quite possible to prepare digital map of orchards of different age groups for important fruits in a district based on spectral reflectance using satellite multi-spectral digital data. This will not only provide complete list (sampling frame) of the orchards but also provide an estimate of area under important fruits of the district.

It is possible to estimate area under important vegetables through the spectral data of remote sensing satellites. The Remote Sensing and GIS technology based approach may not only help in reducing the workload of field survey but also provide a technique of improving the quality of information in this sector by integrating both the sources of information. In case of

vegetable crops, area and production estimation using remote sensing and GIS may not be very reliable. The major reason for this inaccuracy is smaller field sizes of vegetable crops, which creates problem of spatial detection of the fields due to poor spatial resolution of the satellite sensors. This problem may be further complicated when some of the vegetable crop signatures are non-separable from other vegetation classes of the season.

In case of Kharif vegetables, the cloud cover in most of the season does not allow to capture multi-spectral signatures of the crop by the sensors of satellite. Therefore, in case of estimation of area and production of vegetables, an attempt may be made on pilot basis, which may be further extended at large scale depending on the outcome.

References

1. Cochran, W.G. (1977): Sampling Techniques. 3rd Ed., Wiley Eastern Limited, New Delhi.

2. Proceedings of the Symposium on "Statistics of Horticultural Crops: Problems and Issues. *Jour. Ind. Soc. Agril. Stats.*, 54(1), 119-138.

3. Report of the National Statistical Commission(2001): Ministry of Statistics and Programme Implementation publication.

4. Singh, D., Manwani, A.H. and Srivastava, A.K. (1976): Survey on fresh fruits in Tamil Nadu. IASRI publication.

5. Singh, D., Srivastava, A.K., Singh, P. and Pal, S. (1976): Survey on vegetables in Banglore district of Karnataka State. IASRI publication.

6. Sukhatme, B.V., Manwani, A.H. and S.R. Bapat (1969): Survey on vegetables in rural areas of Delhi. IASRI publication.

CROP YIELD ESTIMATION INITIATIVES FOR CROP INSURANCE UNDER PMFBY

Dr. Sunil Dubey & Dr. S. Bandyopadhyay

Mahalanobis National Crop Forecast Centre, Pusa Campus, New Delhi-110012

Background

Pradhan Mantri Fasal Bima Yojana (PMFBY), being implemented in the country from 2016, with the aim to support agriculture production by providing crop insurance against comprehensive risks throughout the crop season. The Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India (GoI), is the nodal agency for executing PMFBY. It is a yield-guarantee scheme over an insured area. Hence, the data on crop yield estimates over Insurance Units for the current and past years are crucial for crop loss assessment and indemnity payout.

Area-yield insurance is a yield guarantee scheme over the specified area. It adopts an area approach defining the Unit Area of Insurance known as an Insurance Unit (IU), generally coinciding with a group of villages called Gram Panchayat (GP). Crop yield estimates of the IU for the current and past years form the basis for crop loss assessment and indemnity pay-out. A certain per cent, 70%, 80%, or 90% of the average yield of the past five best out of seven years of an IU called Threshold Yield (TY) is guaranteed for the current year.

In PMFBY, crop yield estimation is done by carrying out Crop Cutting Experiments (CCEs), i.e., manual yield measurements at a randomly selected limited number of field plots for each crop in each IU. The CCE process are time-consuming and manpower intensive and vulnerable to human errors. The limited number of measurements leading to higher standard errors of estimates and their proneness to subjectivity have become major constraints in generating reliable yield estimates. As a result, the estimated yield of an IU tends to be biased leading to disputes and delays in claims settlements. Thus, the challenge in the area-yield crop insurance in India continues to be improving the crop yield estimation system.

Technology Intervention in PMFBY

Technology interventions, in the form of using satellite data, long-term weather datasets, and usage of mobile apps for CCEs and collecting crop pictures, to improve crop yield estimation and crop damage assessment, are largely recognized and promoted. Technology based estimation is largely based on documenting the edaphic factors, weather and its changes as it occurs in the entire period the crop is on ground (from seed to harvest) and how the health of crop was impacted by it, and is reflective of the potential of the crop yield that has materialized in the season, cutting across all varieties and farming practices. These technology-based estimations are unbiased and when used in conjunction with CCE estimates results in betterment of final yield estimates.

MNCFC, DA&FW has taken up many initiatives to improve crop yield estimation procedures ever since the launch of PMFBY. Technology agencies from both Government and Private sectors have been engaged for developing new yield estimation methods using various datasets and models through pilot studies.

The objective of these pilot studies was to generate reliable yield estimates through scalable models for paddy and wheat crops at IU level. These studies were conducted during 2019 and 2020 for both the kharif and the rabi seasons for selected crops and districts across

different agro-climatic zones of India. The Expert Committee, comprising of experts from Remote Sensing & GIS, Agriculture and Data Science domains, carried out detailed evaluation of pilot studies and recommended five approaches for yield estimation for nation-wide rollout covering paddy and wheat crops from Kharif 2023.

Towards enabling large scale adoption of technology-based yield estimates in PMFBY system for crop loss assessment, DA&FW has conceptualised a special initiative i.e., "Yield Estimation System based on Technology (YES-TECH)" under PMFBY. YES-TECH advocates the blended use of modelled and CCE yield estimates for insurance claim assessment. MNCFC is acting as secretariat of YESTECH and coordinating with different stakeholders involved in YESTECH initiative.

YESTECH Framework

For smooth implementation of YESTECH, the role of various stakeholders such as DA&FW, YESTECH Secretariat, State Government, Technology Implementation Partner (TIP), Mentor Institute for Technology Rollout (MITR) and Insurance Companies (ICs) are clearly defined.

In the YESTECH initiative, rollout of technology in the State is the responsibility of TIPs selected through rigorous scrutiny process. TIP includes, Central/State Government technology agencies actively involved in developing technological solution for Agriculture, Academia (State Agriculture Universities & CGIAR Institutes), Private Sector agencies having proven capability of technology development and implementation. MITR is basically the nominee of Central Government in State to mentor the TIPs in successful rollout of initiative.

Models identified under YESTECH

Based on large scale pilot studies and proactive steps already taken by some of the States to adopt technology-based yield estimation under PMFBY five models are recommended under YESTECH for technology based yield estimation. All the approaches which have been included under YES-TECH, are well established supported with peer reviewed research and these models are in open source and thoroughly tested under Indian conditions. The models are- 1. Semi-physical model, 2. AI/ML models, 3. Crop simulation models, 4. Ensemble models (AI/ Crop Simulation/ Semi-physical etc.,) and 5. Parametric index of crop performance (Indirect approach).

Each and every models identified under YESTECH has some unique feature with certain limitation. Semi Physical model is based on Radiation Use Efficiency (RUE) concept proposed by Monteith, 1977. This model is based on the bio-chemical process of plant–light absorption for photosynthesis, radiation use efficiency, stress factors, accumulated biomass, and grain yield. Its strength lies in adopting a process-based framework with limited parameterization. It is recognized as a better approach than simple empirical modelling. However, crop and region specific RUE and Harvest index (HI) is crucial factor for this model.

Machine learning models are non-parametric and captures the non-linear relationships between the yield and the features influencing the yield. Machine learning can determine pattern and correlations and develop the predictive model. The predictive model is built using several features, and as such, parameters of the models are determined using historical data during the training phase. In recent years, machine learning models have been used in various researches to improve the accuracy of crop yield by incorporating satellite derived vegetation indices, meteorological data, hydrological variables, edaphic factors etc. Good quality yield data at granular scale is very crucial for training the model.

Crop Simulation models simulate the plant processes to estimate various bio-physical parameters and final crop yield. These models need intensive parameterization starting from genetic coefficients of crop variety under cultivation, crop sowing time, crop management practices – fertilizer applications, irrigation supplies, pest/disease occurrence, etc. These are highly reliable point-based or location-specific models due to the availability of input parameters in experimental plots. Model calibration and validation approaches are required for for the selected crop. The simulation models are sensitive to the genetic coefficients of the crop. A proper varietal-based approximation of genetic coefficients is required at the disaggregated level before simulating the crop yield. The modelling approach often remains lumped in absence of spatial input parameters unless space-based inputs are assimilated or externally incorporated.

An ensemble approach to yield estimation involves combining the predictions of multiple models to make a final estimation. This can be done through various techniques such as model averaging, model voting, or more sophisticated methods such as stacking. One advantage of using an ensemble approach is that it can potentially improve the accuracy of the final prediction by reducing the variance of the individual models. This is because the individual models may make different errors, and by combining their predictions, the errors may cancel out to some extent. To use an ensemble approach for yield estimation, one has to train multiple models on the same dataset.

Parametric Index - Crop Health Factor is a composite index of crop performance incorporating multiple physical and biophysical parameters related to crop health. It is a quantitative measure of crop health and its overall performance. The model requires seasons' maximum NDVI, season's maximum LSWI, season's maximum VH backscatter, integrated VH backscatter, integrated FAPAR, rainfall and rainy days as a positive functional relationship whereas requires crop condition variability as a negative functional relationship. This model is not providing yield however it gives a composite index of crop performance which is a quantitative measure of crop health and its overall performance.

Weightage to Tech based yield in PMFBY

It is recommended that at least 30% weightage is to be assigned to modelled yield. Composite yield estimate will be generated by assigning 70% weightage to CCE yield and 30% weightage to modelled yield, and this blended yield estimate will be used for arriving at yield loss assessments. Whereas in case of Crop Health Factor (CHF), 70% weightage is to be assigned to CCE-yield deviation from its threshold and 30% weightage to CHF deviation from its threshold. Further, it is in discretion of States to increase the weightage for modelled yield and there is no upper cap. States who are implementing YESTECH has to declare weight to modelled yield in the beginning of crop season before the start of every season.

Current Status & Way forwards

Currently three major crops i.e. Paddy, Wheat and Soybean crops are notified under YESTECH, provision is to bring more number of crops such as Cotton, Jowar, Bajra, Mustard Gram and Maize. Studies on these crops is in progress. As of now, 9 states (Assam, Andhra Pradesh, Haryana, Karnataka, Madhya Pradesh, Maharashtra, Odisha, Uttar Pradesh, Tamilnadu) are actively involved in the YESTECH and claim for the three identified crop is being settled based on the blended yield of CCE and technology based model yield.

Mahalanobis National Crop Forecast Centre as a YESTECH secretariat implementing the initiative and strive to achieve the objective to bring more number of crops under through continuous research studies as well as to make the scheme more transparent, and efficient.

FASAL 2.0 - AN OVERVIEW

Shri Karan Choudhary, Ms Preeti Tahlani and S. Bandyopadhyay

Mahalanobis National Crop Forecast Centre, Pusa Campus, New Delhi-110012

Background

Timely availability of reliable information on agricultural output, drought and other related aspects are of great significance for planning and policy making -particularly in the management of critical areas such as food security, price stability, international trade etc. The information is extremely useful in identifying problematic and the nature of required spatial, temporal and qualitative interventions. However, the existing system of agricultural statistics, in spite of established procedures and wide coverage, has inherent limitations in the matter of providing an objective assessment of crops at the pre-harvesting stages with the desired details.

In order to enhance capabilities of the existing system of crop forecasts and crop estimation, the Department of Agriculture & Farmers Welfare (DA&FW) considered introduction of technological advancements and adoption of emerging technologies such as Remote Sensing (RS), Geographic Information System (GIS) etc. Accordingly, in the year 1987, the DA&FW sponsored a project called "Crop Acreage and Production Estimates (CAPE)" with the objective of developing methodologies using Remote Sensing (RS) techniques for crop area and production forecasting.

Besides Remote Sensing, other important inputs such as weather data, land based observations and economic parameters influencing farmers' decisions also serve as complementary and supplementary inputs for making crop forecasts. While Crop forecasting with Remote Sensing technique is carried out when crop has grown sufficiently, forecasting at sowing stage is attempted through econometric models using previous years' crop acreage and production data, market prices, current season weather data etc. Thus, an approach which integrates inputs from these diverse sources was needed to make forecasts of desired coverage, accuracy and timeliness and the concept of "Forecasting Agricultural output using Space, Agro-meteorology and Land based observations" or FASAL was revised. FASAL project was implemented since 2006 by Space Applications Centre, in collaboration with India Meteorological Department and Institute of Economic Growth.

After the validation of the approach for FASAL, the technology was transferred to the newly created centre under DA&FW, Mahalanobis National Crop Forecast Centre (MNCFC), which was established in 2012 to operationalize the FASAL project. Since 2012, the project is successfully providing National/State/District level forecast of Rice (Kharif and Rabi), Jute, Sugarcane, Cotton, Rapeseed & Mustard, Wheat and Rabi Sorghum in their major growing regions, in the country. Rabi Pulses and tur crop were also included under FASAL project in the recent years.

Objectives of the FASAL Project

- Implement methodologies developed by ISRO for crop production forecasting and drought assessment.
- Maintain a database comprising data from various sources such as IMD, IEG, and state departments.
- Assimilate crop forecast data from other projects like flood and drought monitoring.

- Coordinate the use of geomatics for other agricultural aspects such as soil health and horticulture.

Crop Production Forecasting

The FASAL system integrates various approaches and organizations for creating a hierarchical information system, related to crop condition and crop production at any time of the season from sowing to harvest. Thus, multiple forecasts of major crops namely Rice(Kharif & Rabi), Jowar (Rabi), Jute, Cotton, Sugarcane, Rapeseed & Mustard, pulses(Rabi) and Wheat were generated at National/State/District level by MNCFC since its inception. Tur crop was also included in the recent years for the assessment.

Econometric and weather-based models were used for forecasting the total crop production, early in the season, before it becomes amenable to remote sensing data. Mid-season assessments were supplemented with multi-temporal, remote sensing data. In the latter half of crop growth, direct use of remote sensing data to estimate the acreage and forecast of the yield was done by integrating remote sensing based indices and meteorological data. In addition, use of field information and weather inputs at various stages increases the accuracy of forecasts.

FASAL 2.0

In the recent years, various developments in technology producing new datasets, tools of analysis have taken place. It was felt to revamp the methodology of generating crop estimates under FASAL by effectively utilizing new datasets and emerging technologies to produce accurate and timely crop estimates. Hence, FASAL project was revamped in the form of FASAL 2.0 to make it more robust and sink it with latest technology.

The FASAL 2.0 focuses on major crops viz. Paddy, Wheat, Soybean, Rapeseed and Mustard, Tur, Gram, Lentil, Cotton, Sugarcane, Rabi Sorghum and Jute.

The two major segment of activities under FASAL 2.0 are-

- Crop Mapping, inventory and health monitoring
- Crop yield estimation by different methodologies and generation of harmonized estimates.

FASAL 2.0 involves collaboration among various technical institutes/organizations and private agencies for better utilization of the latest technology available with different agencies. The role of various organizations under the project is outlined below-

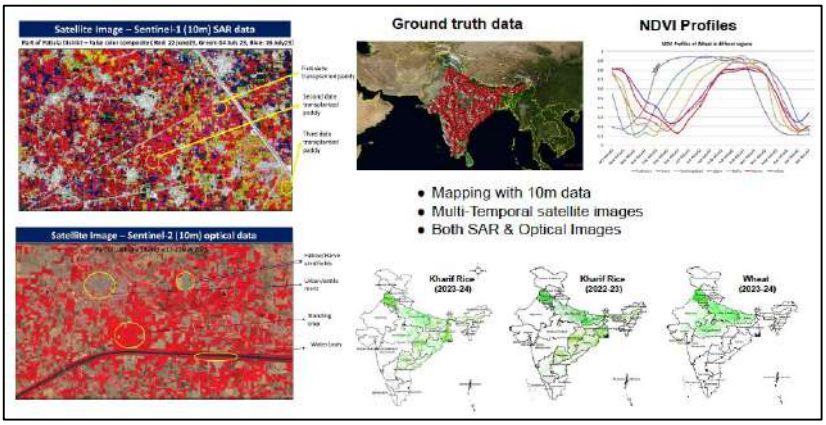


Figure 1: Satellite based Crop Mapping

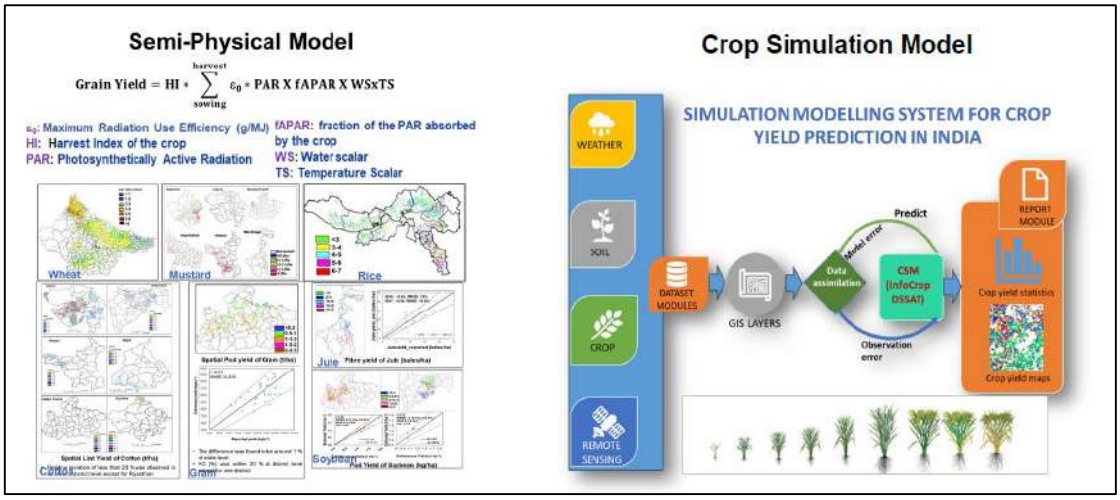


Figure 2: Crop yield Estimation

- **Department of Agriculture and Farmers' Welfare (EA,DES)**
  - Overall Coordination & Programme funding
  - Development of Mobile App for ground truth (GT) data collection
  - Coordinate with Knowledge partners such as IEG, ISI, IITs and IASRI for development AI/ML models for yield estimation
  - Examination of yield estimates provided by various institutes

- **Mahalanobis National Crop Forecast Centre (MNCFC), DA&FW**
  - Operational acreage forecast for Jute, Sugarcane, Rice (3 seasons), Rabi Sorghum and Wheat
  - Engaging industry for Crop Mapping and GT collection of 6 crops viz. Cotton, Soybean, Tur, Gram, Lentil and Rapeseed & Mustard
  - Crop Health monitoring and generation of Crop Health factor
  - Provide inputs to lead agencies for yield estimation
  - Coordination with State remote sensing centres and state departments of Agriculture for support in digital analysis and GT collection
  - Use of Digital Crop Survey (DCS) data for augmenting crop signatures.

- **ISRO as Lead Agency –** ISRO (SAC and NRSC) is identified as Lead agency to provide operational Semi Physical model based Yield for various crops such as Paddy, Soybean, Wheat and Mustard.
- **Indian Council of Agricultural research (ICAR)/Indian Agricultural Research Institute(IARI) as Lead agency -**Identified as Lead agency to provide operational Crop Simulation model based Yield for various crops such as Paddy, Wheat and Mustard.
- **Institute of Economic Growth (IEG) as knowledge partner -**Knowledge partner for providing yield based on econometric & AI/ML models

- **Knowledge partners IASRI, ISI, IITs -** Provide Crop yield based on AI/ML models
- **India Meteorological Department (IMD) –** Agency for providing weather data for yield estimation to various lead agencies and knowledge partners

The generation of productions estimates is done through field data as well as from FASAL 2.0. Recently, Digital Crop Survey (DCS) for crop area and Crop cutting experiments (CCEs) under Digital General Crop estimation(DGCES) for Crop Yield have been started to augment the production estimates using field data.

MNCFC plays a major role in generation of Crop maps and area statistics using remote sensing data which are regularly submitted to the department and also shared with lead agencies for generation of yield estimates through crop simulation and semi-physical models.

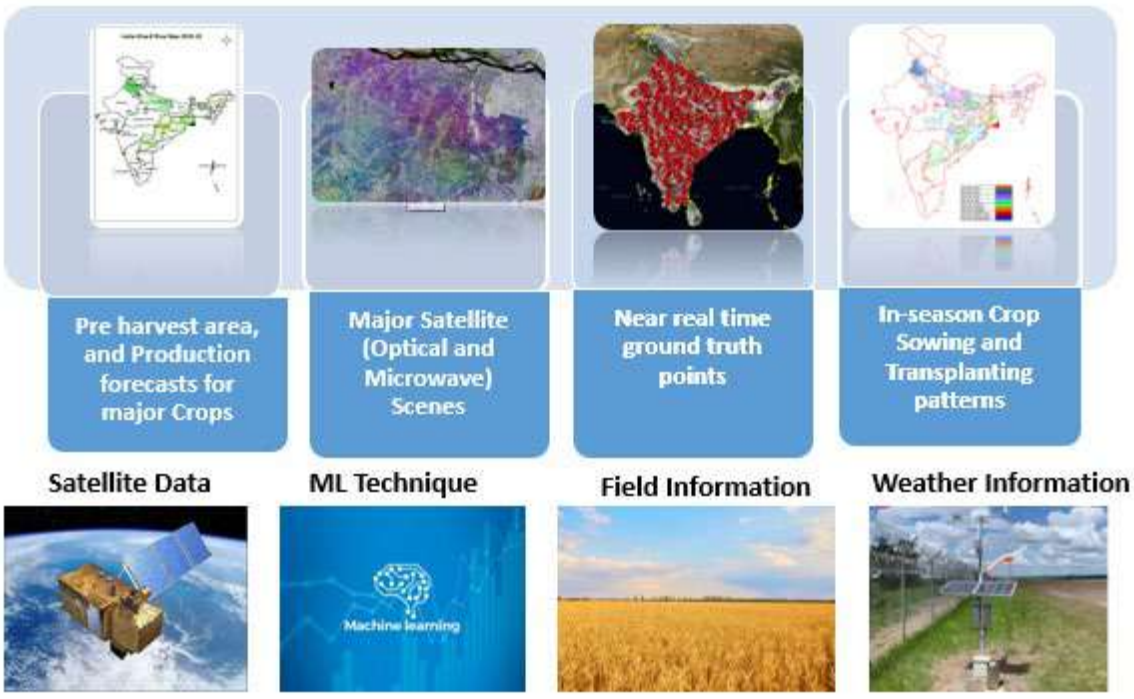


Figure 3: Remote sensing based crop production estimation

The major outputs generated under FASAL 2.0 are -

- Area and Yield of major crops at National, State and District level
- Crop Maps
- Maps for Sowing and Transplanting patterns
- Comparison of Satellite Images for Increase and decrease in crop area
- Remote Sensing based drought monitoring and Support to states.
- Regular inputs to Crop Weather Watch meetings i.e Crop conditions, Soil moisture status, Crop harvesting status for Paddy & wheat crop, Impact of episodic events (hailstorm/heavy rainfall, flood etc.) on crop area, Crop Residue burning, Change in cropping pattern w.r.t previous year
- Ground Truth data collection

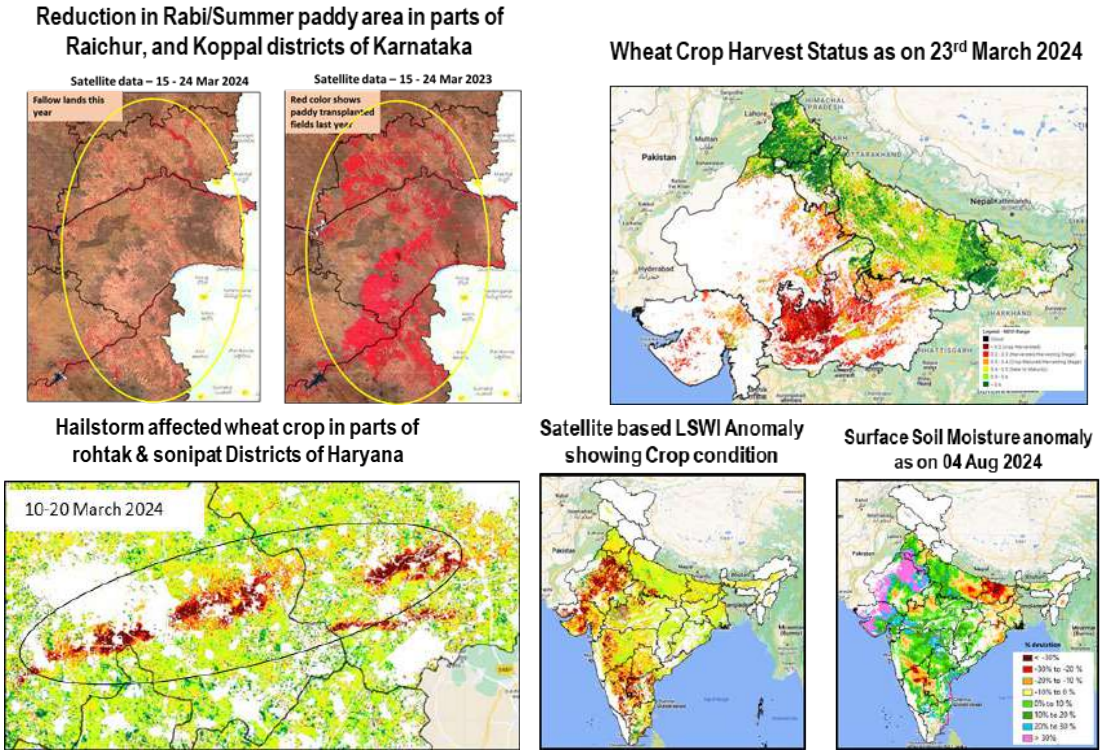


Figure 5. Crop Surveillance & monitoring (Various Data products generated for Crop Weather watch weekly meetings)

Recent achievements under FASAL

- Crop production forecasts using remote sensing and agro-meteorology for 10 major crops – Total 18 Forecasts issued during 2023-24.
- Use of National and International Satellite data (Resourcesat 2, RISAT 1, Sentinel 1, Sentinel 2 and Landsat 8) for crop area estimation, in collaboration with state remote sensing centres.
- **Advanced machine learning models** are being used in MNCFC for crop mapping.
- Around 110000+ GT observations have been collected during the 2024-25 season under the FASAL project.
- Crop Health GT are also being collected under the project.
- Semi-physical based crop yield forecasts were generated by SAC, ISRO.
- Crop Simulation model based yield forecast generated by ICAR/IARI.
- IEG, IASRI, ISI have also been involved under FASAL project for improving yield forecasts using machine learning models and econometric modelling.
- Crop Map Generation is done for Rice, Wheat, Cotton, Jute, Sugarcane, Soybean Rapeseed & Mustard, Rabi Sorghum, Tur, Rabi Rice, Gram and Lentil.
- A new initiative **Krishi Decision Support System (Krishi DSS)** platform is being developed where automatic crop map generation for Rice and Wheat will be done using the pre-trained models.
- All the maps will be published on **Krishi DSS platform** for users.
- Users can also use the platform for crop classification using GT data.
- A new polygon based GT collection app KrishiMapper have been developed and being used for collection of GT from the current year. Training have been given to State agricultural departments for using the **KrishiMapper app**.

Conclusion

FASAL plays a crucial role in agricultural planning, price stabilization, and food security in India. By leveraging remote sensing, econometric models, and field data, it provides reliable crop production estimates, helping policymakers make informed decisions. The project continues to evolve with advancements in satellite technology, AI, and mobile-based data collection.

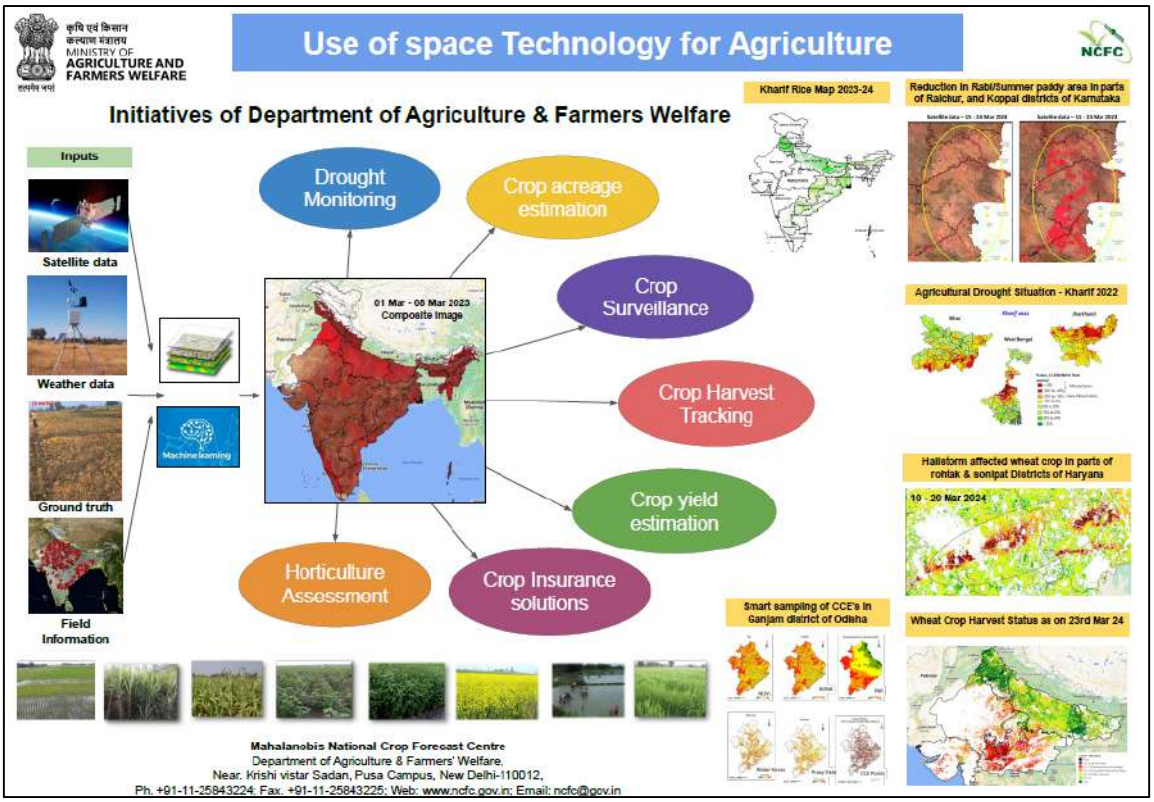


Figure 6: Use of Space Technology for Agriculture

ENERGY AUDIT SURVEYS IN AGRICULTURE

Kaustav Aditya and Bharti

ICAR-Indian Agricultural Statistics Research Institute, New Delhi - 110012

1. Introduction

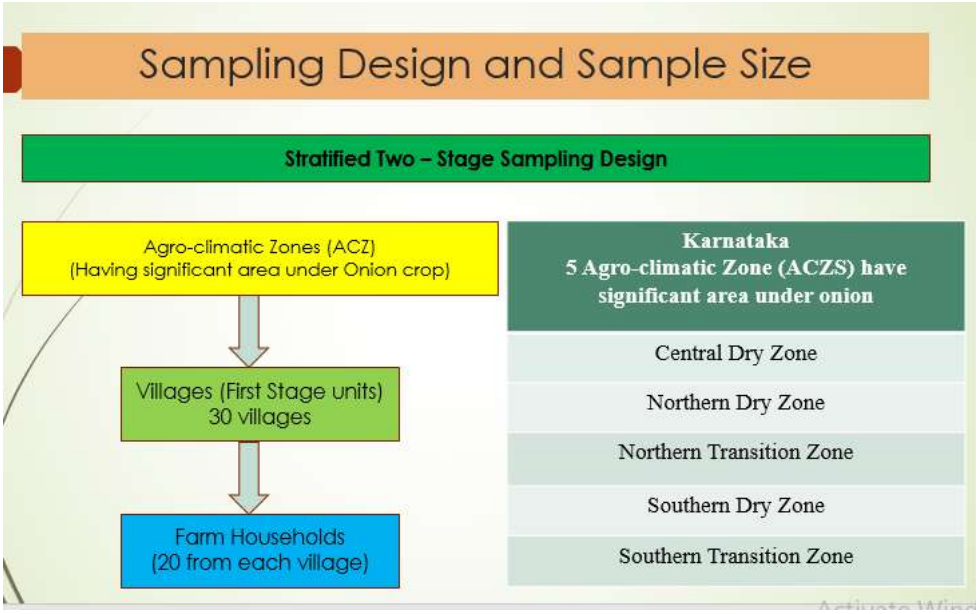
In production agriculture, maximizing crop yield requires the efficient use of various inputs. Different types of energy—direct and indirect, renewable and non-renewable, commercial and non-commercial—are utilized in farming. Direct energy use primarily involves petroleum-based fuels, which power tractors, power tillers, combine harvesters, self-propelled machinery, and equipment used for land preparation, irrigation, weeding, harvesting, fertilizer application, plant protection, and transplantation. Indirect energy sources include seeds, fertilizers, and pesticides, with fertilizers and fuel being the dominant energy consumers in agricultural production.

Assessing energy use and identifying high-energy-consuming operations are key strategies for improving energy management in agriculture. Enhancing energy efficiency and reducing energy intensity are crucial to optimizing energy use, leading to lower production costs and improved energy utilization without compromising crop yields. The adoption of renewable energy systems offers an effective solution for minimizing energy consumption and reducing pollution in agricultural production. Optimal energy use in farming not only boosts crop yields and lowers costs but also conserves fossil fuels and reduces air pollution.

Conducting an energy audit is a crucial first step toward optimizing energy use in agriculture. A systematic approach is essential for identifying energy-intensive operations and implementing effective conservation measures. Additionally, a well-defined methodology is necessary to standardize energy audits in agro-industrial systems. To ensure the effectiveness of energy audit surveys, methodologies must be developed for sampling design, determining appropriate sample sizes, allocating samples across various strata and sub-strata, selecting samples at different stages, and structuring data collection and estimation processes. This chapter presents a clear and structured framework for conducting energy audit surveys in the agricultural sector. It was specifically designed for use by ICAR-AICRP centers on Energy in Agriculture and Agro-based Industries (EAAI) to ensure consistency and comparability of results.

2. Sampling Design and Sample Size

In the energy audit survey, a stratified two-stage sampling design has been implemented. Agro-climatic zones (ACZs) within a state that have a significant cultivation area for the target crop are considered as strata. ACZs with negligible or no cultivation of the selected crop are excluded from the strata. For example, a state has five agro-climatic zones, but two of them have minimal target crop cultivation. Therefore, only the remaining three ACZs are included as strata for the energy audit survey of that crop. Within each stratum, villages serve as the first-stage sampling units (FSUs), while farm households constitute the second-stage sampling units (SSUs). Within each stratum, villages are designated as the first-stage sampling units (FSUs), while farm households serve as the second-stage sampling units (SSUs). A total of 30 villages has to be selected for the major crop of the state. From each selected village, 20 eligible farm households are chosen, ensuring that these households are evenly distributed across the five landholding categories.



3. Sampling Design and Sample Size

In this survey, a total of 600 farm households has be selected for energy auditing in the state's major crop. The allocation of these 600 farm households has to be carried out across different stages, namely first-stage sampling units (FSUs) and second-stage sampling units (SSUs), as outlined below:

- **Allocation of number of villages to Agro-climatic Zones**

Villages has to allocated to agro-climatic zones according to area under crop in that agro-climatic zone.

$$n_i = \frac{A_i}{\sum A_i} \times n$$

where,
n_i = Number of villages to be selected from i-th ACZ, i = 1, 2, …, I
A_i= Area of the onion crop in the i-th ACZ (stratum)

- **Allocation of number of households in land holding categories in villages**

In each of the selected village, a sample of 20 eligible farm households has to be selected. These 20 farm households has to be selected from five landholding categories in the village. These five landholding categories are known as substrata. Allocation of 20 farm households in these three land holding categories (sub-strata) has to be done based on the proportion of number of farm households in these categories (sub-strata) in the selected village. The information on number of eligible farm households in five categories (sub-strata) in the selected villages will be available from the listing exercise where information will be collected in schedule-1

4. **Collection of Data:** There are three stages of data collection in the survey as given below.

- **Schedule-I: Listing/Enumeration of Selected Villages:** In this schedule, we list all the eligible farm households of the selected villages. Eligible farm households would be those who are growing that identified crop of the centre/state. This information will

then be used for the preparation of land holding category-wise list of the eligible farm household in the selected village. The primary data will be collected with village, taluk/block, district, agro-climatic zone, farmer's land holding area as per data questionnaire in Schedule 1.

- **Schedule-II: List of operational holdings in selected village (Re-tabulation from Schedule-I):** This is the list of eligible farm households, categorized into five landholding categories as per size of the operational holdings recorded in Schedule-I.
- **Schedule-III: Detailed operational holding survey - Farmer's questionnaire:** In each village, 20 eligible farm households will be selected from Schedule-II, and detailed operational holding survey is to be done using farmer's questionnaire Schedule-III. Thus, a total of 600 farm households will be surveyed in the State. In these questionnaires, information on basic energy inputs and outputs, detailed energy consumption will be collected and verified with measurements for both source and operations wise. Data collection from the farm household has to be done two times, one at the beginning of the crop season and another at the harvesting time. For example, in case of sugarcane, the operations are land development, land preparation, seed bed preparation, planting, irrigation, weeding, fertilizer application, chemical application, earthing-up, de-trashing, harvesting, transportation and trash management. In addition to the energy consumption, information may be collected with cost details for the type of equipment/implements used, fuel/electricity consumption and man power used.

5. Estimation Formula and Data Analysis:

Energy consumption is to be calculated based on source-wise and operation-wise and also analysis is to be carried out from the collected field data. Besides, estimates of different parameters, standard error and 95 per cent confidence interval will be calculated. From the analysis, energy intensive operation will be identified for better energy conservation and efficiency improvement. Recommendation with economic analysis may be given for effective implementation.

- The estimates of average value of the variables in the selected ACZ is as:

$$\hat{\bar{Y}}_h = \frac{\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik} y_{hik}}{\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}}$$

- The estimate of average value of the variable in the state is as

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik} y_{hik}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}}$$

- The estimates of variance of the estimate of average value of the variable in the state is as

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}(w_{hik}-1)\left(y_{hik}-\hat{\bar{Y}}\right)^2}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}\right)^2}$$

DIGITAL AGRICULTURE: ICAR PERSPECTIVE

Anil Rai

Indian Council of Agricultural Research, New Delhi

Introduction

Digitalization in agriculture has a great potential to bring about significant innovation and transformation in India's agricultural sector. The adoption of digital technologies can help India to increase agricultural productivity, reduce waste, increase agricultural export, increase farmers' income and improve food and nutrition security. This will also help in protecting environment and sustainable development of overall agricultural sector. The country is ready for digitalization of agriculture as the growth and penetration of digital infrastructure in India is quite high. India has 1.2 billion mobile phone subscribers having penetration rate 86% with monthly growth rate of 0.21%. The penetration rate of smartphones is 71%. The internet penetration rate in India during year 2022 was around 49%. The internet penetration in rural India is 41% as opposed to 71% in the urban India. This clearly shows that now Indian agricultural is ready for digital transformation.

Digital Agriculture: The science of Digital Agriculture (DA) has been gaining prominence with the advent of fast-paced advances in cutting edge technologies that can have significant impact, specially, to small and marginal farmers. DA offers a wide range of technological solutions for farmers *i.e.* smart precision agriculture, data-driven decision support system and Information and Communication Technology (ICT) based relevant services including financial services. Smart precision agriculture refers to agriculture system which (i) is ecologically sound, (ii) has enhanced productivity, (ii) has lower cost of production, (iii) is climate resilient and sustainable, (iv) has adoption of scientific proven technologies and (v) is conservative use of inputs without hampering the environmental health. The digitalization of agriculture can drive innovation and transformation in broadly four sectors of Indian agriculture namely (i) Precision agriculture, (ii) Climate-smart agriculture, (iii) Supply chain management, and (iv) Financial inclusion.

Precision Agriculture is a set of digital technologies are being used such as sensors, drones, and satellite imagery. Using digital equipment, tools, software, process, farmers can monitor their crops, orchards, animals, aquaculture systems, soil, water, weather, post-production management of produce in real-time, enabling them to make more informed decisions about planting, irrigation, fertilization, pest management, harvest and post-harvest management of agricultural produce. This can help to reduce waste, improve yields, improve quality, and save costs. Briefs about some major precision agriculture digital technologies are given in brief.

1. Applications of sensors in agriculture:

Sensors are being increasingly used in Indian agriculture to improve productivity, reduce waste, and optimize resource use. Soil moisture sensors are used to measure the amount of water in the soil. By providing real-time data on soil moisture levels, farmers can optimize their irrigation practices, reducing water waste and increase crop yields. Nutrient sensors can provide real-time data on nutrient levels in the soil, enabling farmers to optimize their

fertilization practices and reduce nutrient waste. Weather sensors can provide real-time data on weather conditions, including temperature, humidity, wind speed, and rainfall. This information can help farmers make more informed decisions about planting, harvesting, and other farm operations. Crop health sensors are used to monitor the health and growth of crops. They can detect early signs of disease, nutrient deficiencies, and pest infestations, enabling farmers to take timely action to prevent crop losses. Livestock sensors are used to monitor the health and behaviour of animals. They can detect early signs of illness or stress, enabling farmers to take timely action to prevent the spread of disease or reduce animal mortality. Water quality sensor are being used to monitor the water quality specially dissolved oxygen and ammonia in water. They are helpful in making aquaculture system efficient and high productive. The use of sensors in agriculture helps farmers to take more informed decisions and optimize their operations and resources, leading to higher yields, reduced waste, and improved profitability. However, it is important to ensure that these technologies are accessible and affordable to small-scale and marginalized farmers to avoid exacerbating existing inequalities in the agricultural sector.

Sensor-based post-harvest management of agricultural produce can help to reduce post-harvest losses, improve quality, and increase profitability for farmers. Different sensors are being used for this purpose. Temperature sensors: Temperature sensors can be used to monitor the temperature of harvested produce during storage and transportation. By providing real-time data on temperature, farmers can identify potential issues (such as overheating or chilling injury) and take corrective action before it is too late. Humidity sensors: Humidity sensors can be used to monitor the moisture content of harvested produce. By maintaining optimal humidity levels during storage, farmers can help to reduce the risk of fungal and bacterial growth, prolong shelf life, and maintain product quality. Ethylene sensors: Ethylene sensors can be used to monitor the production of ethylene gas, which can accelerate the ripening process and reduce shelf life. By monitoring ethylene levels, farmers can take timely action to slow down the ripening process and extend the shelf life of their produce. Quality sensors: Quality sensors can be used to measure the quality of harvested produce, including factors such as size, colour, and texture. By monitoring quality, farmers can identify potential issues and take corrective action before the produce is shipped to market. GPS sensors: GPS sensors can be used to track the location of harvested produce during transportation, enabling farmers to monitor temperature, humidity, and other factors that can impact quality. Further, sensor-based storage management is an effective method of post-harvest management that can help Indian farmers to increase their productivity and income, reduce post-harvest losses, and improve the quality of their produce. However, it is important to ensure that the technology used for sensor-based storage management is affordable, easy to use, and appropriate for the needs of small-scale farmers. Additionally, farmers should be trained in the use of the technology to ensure that they are used effectively. The major advantage of this are (i) Improved storage conditions: Sensor-based storage management enables farmers to monitor the temperature, humidity, and other environmental factors in real-time. This helps to maintain the optimal storage conditions for different types of crops, preventing spoilage, and increasing the shelf life of the produce, (ii) Reduced wastage: By maintaining the optimal storage conditions, sensor-based storage management helps to reduce post-harvest losses due to spoilage and damage. This can result in higher profits for farmers and a more sustainable food supply chain, (iii) Increased efficiency: Sensor-based storage management automates the monitoring and control of storage conditions, reducing the need for manual labour and increasing efficiency. This saves time and

reduces labour costs for farmers, (iv) Real-time data: Sensor-based storage management provides real-time data on storage conditions, enabling farmers to make informed decisions about when to harvest and sell their produce. This can help to optimize the timing of sales and maximize profits (v) Quality assurance: By maintaining the optimal storage conditions, sensor-based storage management helps to ensure that the quality of the produce is maintained. This can increase consumer confidence in the produce and help to build a brand reputation for quality. The use of sensor-based post-harvest management in Indian agriculture can help to reduce losses, improve quality, and increase profitability for farmers. However, it is important to ensure that these technologies are accessible and affordable to small-scale and marginalized farmers, who may not have the resources to invest in such technologies. This may require government support in the form of subsidies, training programs, and infrastructure investments.

2. Variable Rate technology (VRT):

Variable Rate Technology (VRT) is a precision agriculture practice that involves the use of technology to apply different amounts of inputs (e.g., seed, fertilizer, water, pesticides, feed, chemicals) to different parts of a field, orchards, or any agricultural production system including livestock and fisheries, based on the specific needs. VRT is becoming increasingly popular in agriculture as a way to improve yields, reduce input costs, and minimize environmental impact. Some popular example in which VRT can be used in Indian agriculture are (i) seed planting, (ii) fertilizer application, (iii) pesticide application and (iv) water management. The VRT can be used to optimize seed planting by applying seeds at variable rates based on soil characteristics, topography, and historical yield data. This can help to ensure that each part of the field receives the appropriate amount of seeds, improving crop yields and reducing waste. The fertilizers can be applied using VRT at variable rates based on soil nutrient levels, crop requirements, and environmental conditions. This can help to reduce fertilizer waste and runoff, minimize nutrient leaching, and improve soil health. VRT can also be used to apply pesticides at variable rates based on pest pressure/infestation, crop growth stage, and environmental conditions. This can help to reduce pesticide use, minimize the risk of resistance development, and improve pest control efficacy. The optimize irrigation practices can be followed using VRT by applying water at variable rates based on soil moisture levels, soil texture, crop water requirements, and weather conditions. This can help to conserve water, reduce irrigation costs, and improve crop yields. The use of VRT in agriculture can help to improve productivity, reduce input costs, and minimize environmental impact. However, it is important to ensure that cost of these technologies are reduced to make it accessible and affordable to small-scale and marginalized farmers, who may not have the resources to invest in precision agriculture practices. This may require government support in the form of subsidies, training programs, and infrastructure investments.

3. Drones:

Drones, or unmanned aerial vehicles (UAVs), have become increasingly popular in agriculture due to their ability to collect data quickly and efficiently of an object without physical contact with object. Major applications of drones in agriculture are as follows:

a. Crop monitoring: Drones equipped with high-resolution cameras can be used to monitor crop health and growth throughout the growing season. This can help farmers

to identify issues such as nutrient deficiencies, pest and disease outbreaks, and water stress, and take corrective measures before crop losses occur.

b. Mapping: Drones can be used to create detailed maps of agricultural land, providing farmers with information on soil quality, topography, and drainage patterns. This can help farmers to optimize their planting and irrigation practices, leading to higher yields and reduced input costs.

c. Crop spraying: Drones can be used to spray crops with precise applications of pesticides and fertilizers based on infestation and crop requirements at very micro level, reducing the need for manual spraying and minimizing exposure to chemicals for farmers. This can also reduce the amount of chemical waste that is produced, improving environmental sustainability.

d. Precision agriculture: Drones can be used in conjunction with other precision agriculture technologies, such as GPS-guided equipment and variable rate technology (VRT), to optimize input application based on the specific needs of each crop. This can help to reduce input costs, minimize waste, and improve yields.

e. Crop damage assessment: Drones can be used to assess crop damage caused by natural disasters such as floods, droughts, and wildfires. This can help farmers to estimate losses and make insurance claims more quickly and accurately.

f. Aquaculture management: It can be successfully employed in water sampling from rivers and large reservoirs to monitor the water health and ecological conditions including fish health and biomass.

However, it is important to ensure that drones are used safely and responsibly, and that farmers are trained in their use. Additionally, regulations around the use of drones in agriculture should be put in place to ensure that they are used legally and ethically.

4. Robotics:

Robotics is the field of technology that involves the design, construction, and operation of robots. In recent years, robotics has been increasingly used in agriculture to improve efficiency, reduce labour costs, and increase productivity. Robots can be successfully employed for harvesting crops such as fruits and vegetables can increase efficiency and reduce labour costs. For example, robotic arms can be used to pick fruits and vegetables, while robots equipped with sensors which can identify ripe produce along with assessment of its quality. Planting and seeding can be done with specially designed robots with precision and accuracy. This can improve crop yields and reduce waste. Robots can also be used to detect and remove weeds, reducing the need for herbicides and manual labour. Autonomous robots equipped with sensors and irrigation systems can be used to water crops with precision, reducing water waste and increasing efficiency. It has been demonstrated to use robots in laborious and hazardous operations of agriculture such as spraying chemicals which are dangerous to human health. Overall, robotics has the potential to revolutionize agriculture in India by increasing efficiency, reducing labour costs, and improving productivity. However, the use of robotics in agriculture requires significant investment in technology, research, and development. Additionally, farmers and farm workers must be trained in the use of robotics technology to ensure that it is used effectively.

5. Artificial Intelligence (AI):

Artificial Intelligence (AI) has the potential to revolutionize Indian agriculture by enabling farmers to make data-driven decisions and optimize their crop yields. It can be effectively used for (i) predictive analytics, (ii) precision agriculture, (iii) pest and disease management, (iv) soil health management, (v) livestock herd management and dairy automation, (vi) aquaculture automation and management (vii) market prediction. In case of predictive analytics, AI can be used to analyse historical weather data and crop data to make predictions about future crop yields. This can help farmers to plan their planting and harvesting schedules and optimize their crop management practices. AI can be used in precision agriculture in conjunction with other precision agriculture technologies, such as drones and GPS-guided equipment, to optimize input application based on the specific needs of each crop. This can help to reduce input costs, minimize waste, and improve yields. Pest and disease management can be done using AI to identify pests and diseases early on and recommend appropriate treatment options. This can help farmers to minimize crop losses and reduce the need for chemical treatments. AI can be used to analyse soil data and provide recommendations for optimizing soil health. This can help farmers to reduce soil erosion, improve water retention, and increase crop yields. Applications of AI has been successfully demonstrated to predict and identify animal and fish diseases through models and image analysis. It is also successfully used in complete automation of livestock and fish production and its automation.

6. Protected Cultivation:

Protected cultivation is the practice of growing crops under a controlled environment using structures such as greenhouses, shade net houses, and polyhouses. This method of cultivation has become increasingly popular in Indian agriculture due to its ability to protect crops from adverse weather conditions, pests, and diseases**.** These practice of cultivation is quite beneficial to the farmers such as (i) Extended growing season: It allows farmers to extend the growing season and cultivate crops throughout the year, irrespective of seasonal changes. This enables farmers to produce more crops and increase their income, (ii) Higher yields: Protected cultivation provides a controlled environment for crops, which helps to optimize their growth and increase their yields. This is particularly beneficial for high-value crops such as vegetables and fruits, (iii) Efficient water and nutrient management: allows farmers to efficiently manage water and nutrients, as they can control the amount of water and fertilizers supplied to the crops. This helps to reduce water and nutrient wastage and improves crop yields, (iv) Pest and disease control: It provides a barrier against pests and diseases, reducing the need for chemical pesticides and herbicides. This reduces the cost of crop management and improves the quality of the crops, and (v) Better quality produce: Due to controlled environment for crops, it helps to improve their quality. This is particularly beneficial for high-value crops such as vegetables and fruits, which require high-quality produce to fetch better prices in the market. The protected cultivation is an effective method of cultivation that can help Indian farmers to increase their productivity and income. However, it is important to ensure that the structures used for protected cultivation are sustainable and do not have adverse effects on the environment. Additionally, farmers should be trained in the use of protected cultivation techniques to ensure that they are used effectively.

7. Vertical Farming:

Vertical farming is a method of growing crops in vertically stacked layers using artificial lighting and a controlled environment. This method of cultivation has become increasingly popular in Indian agriculture due to its ability to produce high yields of crops in a small space and with minimal environmental impact. It has almost same benefits as protected cultivation. Apart from the benefits of protected cultivation it has efficient use of space i.e. Vertical farming allows farmers to grow crops in a small space, making it ideal for urban agriculture and areas with limited land availability. This can help to increase the productivity of agriculture in cities and reduce the carbon footprint of transporting food from rural areas. It can be seen that vertical farming is an effective method of cultivation that can help Indian farmers to increase their productivity and income, particularly in urban areas. However, it is important to ensure that the equipment used for vertical farming is sustainable and does not have adverse effects on the environment. Additionally, farmers should be trained in the use of vertical farming techniques to ensure that they are used effectively.

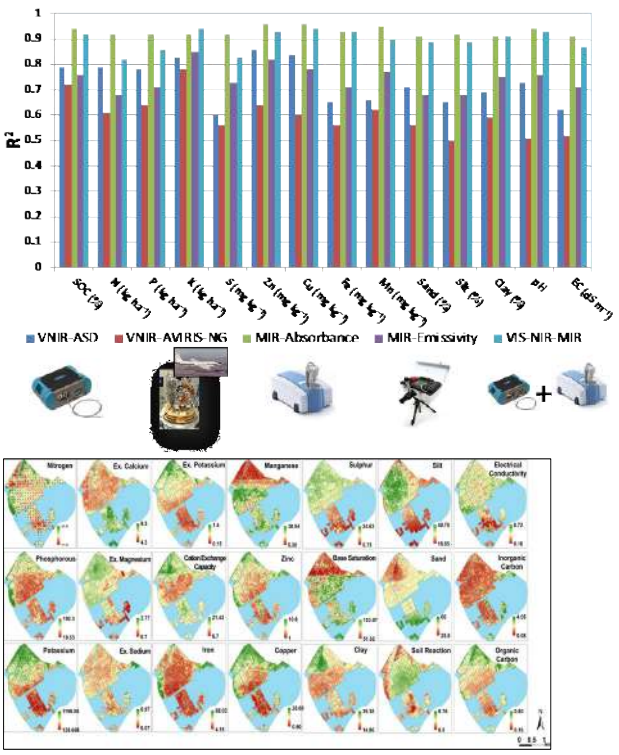
8. Aquaponics and Hydroponics:

Aquaponics and hydroponics are two modern methods of cultivation that have become increasingly popular in recent years. Both methods involve growing plants without soil, using water as the primary growing medium. Aquaponics is a method of cultivation that combines aquaculture and hydroponics. In this method, fish are grown in tanks, and the waste produced by the fish is used to fertilize the plants grown hydroponically, whereas, Hydroponics is a method of cultivation that involves growing plants in a nutrient-rich solution without soil. The major benefits of these systems are (i) Efficient use of space: Hydroponics allows farmers to grow crops in a small space, making it ideal for urban agriculture and areas with limited land availability, (ii) Efficient use of resources: Aquaponics combines fish farming and plant cultivation in a closed-loop system, which uses water and nutrients efficiently and minimizes waste, (iii) Reduced water usage: Aquaponics uses up to 90% less water than traditional soil-based agriculture, making it ideal for areas with water scarcity, whereas, hydroponics uses up to 70% less water than traditional soil-based agriculture, (iv) High-quality produce: The nutrients supplied by the fish waste help to produce high-quality crops, making aquaponics ideal for high-value crops such as herbs and leafy vegetables, whereas , in case of hydroponics we apply highly reduced chemical pesticides and herbicides. The aquaponics and hydroponics are effective methods of cultivation that can help Indian farmers to increase their productivity and income, particularly in urban areas or areas with water scarcity. However, it is important to ensure that the equipment used for aquaponics and hydroponics is sustainable and does not have adverse effects on the environment. Additionally, farmers should be trained in the use of aquaponics and hydroponics techniques to ensure that they are used effectively.

Indian Council of Agricultural Research (ICAR) initiated a network program in the field of precision agriculture named as "***ICAR- Network Program on Precision Agriculture (NePPA)"*** , initially with 16 ICAR Research Institutes with IARI as Lead. The program is focused on exploring potential applications of recent developments on technologies related to sensors, IoTs, drone and ICTs, variable rate technologies (VRTs) for precision smart agriculture. The major objectives span its scope bringing precision in monitoring and managing soil fertility, crop health, livestock farming, post-harvest operations, aquaculture and upscaling

in farmers' field scale to enhance input use efficiency and optimal production system. During last two years' of its operation, number of digital technologies have been developed and validated. Some of them which are ready to roll out in the field are briefly given below

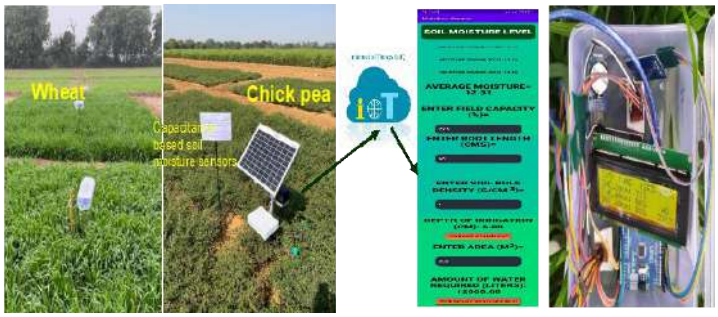
- Rapid sensor based appraisal of soil health at field scale has been developed. Presently, soil health assessments are performed under laboratory conditions using wet chemistry with tedious and time-consuming sample preparation and analysis. This will accelerate the national soil health card program by Government of India. Based on evaluation of different imaging and non-imaging sensors for estimating soil fertility attributes, it was found that the developed MIR sensor is the best in terms of prediction accuracy of important soil attributes. Apart from this, digital soil mapping of soil fertility using remote sensing satellite derived environmental co-variates can also be implemented using machine learning and generated soil fertility maps at regional level with reasonable accuracy. In this regard, an android based app was also developed for crop specific fertilizer recommendation based on inherent soil fertility and target yield.



- Drone Remote Sensing technology was used to monitor near real time crop condition through quantitative assessment of plant biophysical parameters such leaf area index (LAI), canopy chlorophyll (CCC) and nitrogen content (CNC) for site specific nitrogen application using variable rate technology. This is quite cost effective and environment friendly technology which can be immediately used by the farmers.

- A mobile based RiceNXpert was developed, which recommends timing and dose of N fertilizer based on 10 photographs of rice leaves. Similarly, low-cost smart phone based Pusa N manager was also developed for real time fertilizer recommendation in the farmers' fields in all weather conditions.

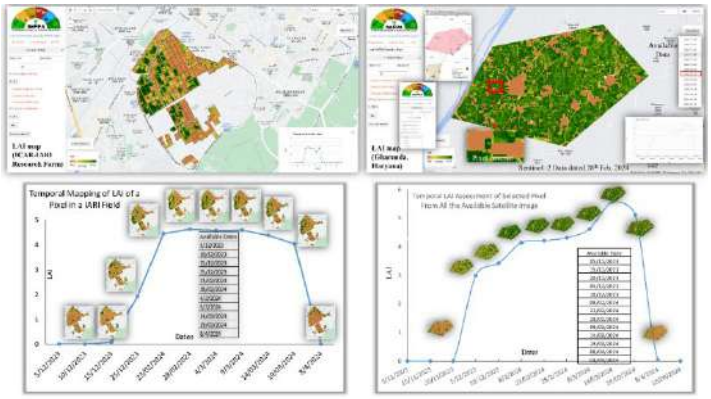
- Artifical Inteligence (AI) based intelligent Irrigation System for field crops using low-cost sensors and IoT technology was developed for precision water management and enhancing water use efficiency. It was tested and found to be very effective in wheat and chick pea crops.



- Sensor-based system is being developed for monitoring the perishable fruit like banana during transportation. In this system sensor data, traceability related data and location etc. can be fetched using web based software.
- In case of precision livestock farming (i) infrared thermography could be used to identify mastitis in the Sahiwal cows, (ii) oestrus detection in dairy cows is being done using wireless accelerometer system (pedometer), and (iii) IoT based herd environment mainly monitoring temperature, humidity and of the herd and managing them through automatic misting and fanning actuated as per the desired level.

- A prototype for IoT based monitoring of fish aqua-system was developed for monitoring and management of dissolved oxygen, temperature, salinity, pH and TDS of water. IoT based mobile operated smart fish feeder was developed which can be programmed to dispense desired rations 12 times/day. The device is fully operated through android based mobile App. Also, feeding and fish behaviour can be monitored through IP CAMERA via smart fish feeder mobile App.

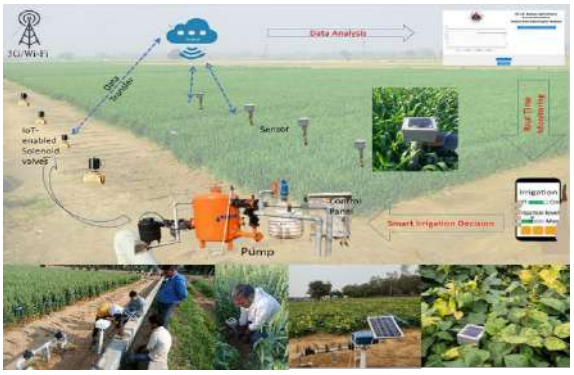
- **Near Real time Monitoring from a single field to large scale using Remote Sensing;** Models for near real-time crop monitoring utilize Sentinel-2 data, offering high-resolution insights into LAI (Leaf Area Index) at various scales. Using Google Earth Engine and a Gaussian Process Regression (GPR) model trained on PROSAIL simulations, researchers validated the approach with in-situ data from IARI farms in New Delhi, producing an LAI map for large croplands in Haryana.



- Drone based 5G Captive Network on demand for Smart Farming: The Network-in-a-Box (NIB) is a self-contained solution designed for the swift deployment of private 5G networks in a limited . It combines a fully-functional 5G SA Core and 5G RAN into a compact and portable unit, enabling users to establish a complete 5G network with minimal configuration. The pre-integrated design and intuitive user interface of NIB expedite network setup, minimizing downtime and simplifying the deployment process. NIB's versatility allows for adaptation to diverse deployment scenarios and spectrum requirements. The NIB's controlled environment makes it perfect for testing and simulating 5G network behavior in a variety of conditions.

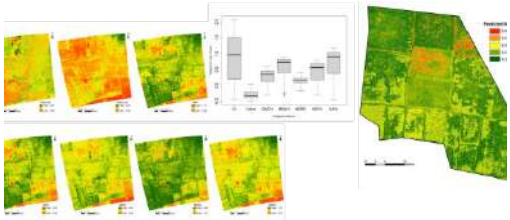
- Soil Moisture Sensor-Based Automatic Basin Irrigation System: The sensor based automatic irrigation system consisting of sensing unit, communication unit, and control unit was developed and evaluated IARI research farm with wheat crop. Low cost capacitance based soil moisture sensor, microcontroller, check gate with control unit were integrated together and powered from solar panel. Communication among them was established using LoRa and GSM. The system was evaluated in the research field of IARI in wheat crop. A total of 25% of water was saved through real-time soil moisture status-based automatic irrigation systems than the conventional manually controlled system under wheat crop.

- IoT based Sub-surface Drip Irrigation Automation in Cereal based Cropping System: Automatic sub-surface drip irrigation system having dripper installed at 20 cm depth, fertigation system, electric water pump, soil moisture sensor (SMS), solenoid valve enabled with solar panel and cloud server with a simple smartphone interface for soil moisture monitoring and irrigation scheduling was done in IARI field. Due to the sub-surface drip system water application efficiency achieved > 95%, water saving 30-40%, nutrient saving 40-50%, crop yield increased by 25% and B:C ratio: 1.85-2.1.



- Sensor-based irrigation automation for alternate wetting and drying (AWD) in transplanted rice: A sensor-based irrigation automation system for alternate wetting and drying (AWD) in transplanted rice was developed using ultrasonic sensors to monitor water levels. Field tests in 2023 showed similar grain yields (4.21 t/ha) with 18% less water use (1100 mm) and a 15% increase in water productivity (0.38 kg/m^3).

- Developing precision nitrogen management protocols for rice using remote sensing and geospatial tools: Upscaling the Nitrogen recommendation to farmers' field using multispectral satellite sensors and/or drone mounted sensors for enhancing nitrogen use efficiency. Our methodology for this specific objective, encompasses a comprehensive data collection and processing phase, followed by the development of a predictive model using Random Forest for nitrogen estimation in crop fields. The final model's performance was evaluated on a validation dataset, and we identified key influential variables. Additionally, the model's predictions were used to generate nitrogen level maps for the study area, highlighting the practical application of our work



- AI integrated IoT for pest forewarning system; ICAR-CICR developed an AI-based smart pheromone trap for real-time monitoring of pink bollworm in cotton. Using a YOLO-based machine learning algorithm with 91.4% detection accuracy, the trap captures insect images, counts them, and transmits data along with weather parameters hourly to a remote server. The system has been successfully deployed in Punjab.

- Development of precision Soil fertilizer recommendation system: An e-Precision Fertilizer Recommendation System has been developed integrating soil analysis, digital mapping, and target yield data to provide personalized fertilizer advice for farmers. Supported by a mobile and web-based platform, it uses geospatial tools and digital soil maps to help optimize crop yields and farming practices.
- Drone based water sampling and quality assessment for Inland open waterbodies:
- ICAR-CIFRI developed a UAV-based water sampling system for collecting water samples from large or inaccessible water bodies. This approach enables assessment of water quality and emerging contaminants, supporting aquatic health monitoring and raising awareness about contaminants' impact on biodiversity and public health.

- Smart fish feeder for pond, RAS, and Aquarium; Smart Feeder is a viable solution to address feed optimization in pond-based commercial aquaculture, RAS, BioFloc, koi ponds, aquariums and Ornamental fish rearing tanks. By automating feed dispensation, this innovative device helps reduce costs, improve feed efficiency, and enhance sustainability. There is an Integrated IP camera that monitors fish and feeding activity and gives informed decisions and timely issue resolution. It also allows IoT Integration and Remote Monitoring to manage feeding activities remotely. It is monitored and controlled via CIFA AquaMegh App (Android & iOS).
- Machine Vision System for Mango sorting and grading ; A Machine Vision System was developed for inspecting mango features, providing real-time data via a GUI and PLC for tasks like sorting blemished fruit and optimizing production process. It enhances consistency in sizing and sorting while enabling automated defect detection. Preliminary

trials using fluorescence imaging showed promise for predicting ripeness and detecting internal defects like spongy tissue in Alphonso mangoes.

- Artificial Intelligence based mastitis disease identification in Cattle and Buffalo: The symptoms of Mastitis infection are seen in mammary glands of cow (teat part) and received thermal images of Sahiwal breed of cow from ICAR-National Dairy Research Institute, Karnal. These thermal images are in three categories i.e., Normal means healthy animal, Clinical means infected animal and Sub-clinical means initial stage of infection in animals. Deep Learning based Convolutional Neural Network model was applied and model was trained in two conditions namely Normal vs Clinical and Normal vs Subclinical. A web-based tool was also developed by implementing these models and options were provided to user to upload the thermal image of udders for identification of mastitis disease in the animal.

Digital agricultural Knowledge Dissemination System:

Agricultural knowledge dissemination systems are important for providing Indian farmers with the information they need to make informed decisions about their farming practices. Digital technologies can play an important role in improving the efficiency and effectiveness of these systems. In India, following systems are already being implemented by different organizations responsible for the same:

a. Mobile-based applications: Mobile applications can be developed to provide farmers with information on crop management practices, weather forecasts, market prices, and other relevant information. These applications can be designed to be user-friendly and accessible even for farmers with low literacy levels.

b. Interactive voice response (IVR) systems: IVR systems can be used to provide farmers with information on demand, using a telephone. This can be useful for farmers who do not have access to smartphones or the internet. IVR systems can also be designed to provide personalized recommendations based on the specific needs of each farmer.

c. Digital kiosks: Digital kiosks can be set up in rural areas to provide farmers with access to information on crop management practices, weather forecasts, market prices, and other relevant information. These kiosks can be staffed by trained personnel who can assist farmers with accessing and interpreting information.

d. Farmer helplines: Farmer helplines can be set up to provide farmers with advice and support on a wide range of issues related to agriculture. These helplines can be staffed by trained agronomists and other agricultural experts who can provide personalized recommendations based on the specific needs of each farmer.

e. Social media: Social media platforms such as Facebook and WhatsApp can be used to disseminate agricultural information to farmers. Farmers can join groups and communities on these platforms where they can share information and advice with each other, as well as receive updates on new agricultural practices and technologies

Indian Council of Agricultural Research (ICAR) developed a digital platform for agricultural extension system named as ***KISAN SARATHI - System of Agri-information Resources Auto-transmission and Technology Hub Interface.*** **This has been developed** to support this emerging need of multi ways and multilingual communication among various agricultural stakeholders, "Kisan Sarathi" an Information Communication and Technology (ICT) based interface solution. The ultimate goal of this project is to implement an intelligent online platform for supporting agriculture at local niche with national perspective. A total of 731 KVKs are enrolled with the system, where, more than three thousand six hundred agricultural scientists and subject matter experts are registered with Kisan-Sarathi. The services of Kisan-Sarathi for the farmers is available through an IVR based calling system via toll free numbers 1800-123-2175 and a short number 14426. More than 250 Lacs farmers are registered on this portal.

KRISHI - Agricultural Knowledge Resources and Information System Hub for Innovations portal developed by ICAR to bring its knowledge resources from all stakeholders at one place. The portal is being developed as a centralized data repository system of ICAR consisting of technology, data generated through experiments/ surveys/ observational studies, geo-spatial data, publications, learning resources etc. The KRISHI Portal, with over 41 lakh publications in the ICAR Research Data Repository, facilitates data-driven decision-making, technology adoption, and agricultural advancements. It grants easy access to extensive research and data for farmers, researchers, and policymakers**.** ICAR was also honoured for this initiative with the Gold Icon Award in the Open Data Championship Category by MEITY, Govt. of India in 2020.

ICAR institutes developed total 278 mobile apps for dissemination knowledge related to different fields of agriculture including, field, horticulture, and planation crops, animal and livestock, fisheries, farm mechanization and post-harvest management etc.. In order to ensure easy access of these mobile apps, an integrated mobile app named as ***"Krishi Integrated Solution for Agri Apps Navigation i.e. KISAAN 2.0"*** has been developed by ICAR. Further, ICAR developed number of AI based servers/ solutions for empowering agricultural research and extension system including AI-based disease identification system (AI-DISC), an AI-powered mobile application that automatically identifies diseases in 19 crops. Apart from this, all 113 ICAR institutes have their own Website for dissemination of information related to their respective domains. ICAR has more than 360 social media handles with huge number of followers on each. Total 63 call centres were established by ICAR for responding and addressing questions/queries of stakeholders. ICAR institute also developed online portals for selling their products and outputs. Recently, a MoU has been signed by ICAR with Amazon to accelerate this process. Further, ICAR institutes developed 360 databases and provided assistance to 832 major start-ups in the field of agriculture.

e-Governance: ICAR implemented e-office, Electronic Human Resource Management System (e-HRMS), Smart Performance Appraisal Report Recording Window (SPARROW), Court Case Management System (CCMS) software developed by NIC besides software developed at ICAR such as Land Record Management Information System (LRMS), Agricultural Research Management System (ARMS), Personal Information Management System (PIMS) to bring more objectivity, accountability, and transparency in office governance and management. ICAR Darpan Dashboard is customized using DARPAN portal developed by NIC to transform complex government data into compelling visuals. All ICAR Schemes/Projects are classified into 12 projects consists of 25 Key Performance Indicators (KPIs) which is available online. The Data Governance Quality Index (DGQI), a toolkit of NITI Aayog, provides a unique framework for self-assessment of data preparedness levels across the Government of India. DGQI is based on internationally accepted data preparedness assessment models from private and public sectors but appropriately contextualized for India. Due to number of digital initiatives taken by ICAR, the current value of DGQI Index of Department of Agricultural Research and Education (DARE)-ICAR is 4.43 out of 5.0, i.e. DARE-ICAR has been recognized as among top most Departments of Government of India in terms of Digitalization by NITI Aayog.

Climate Smart Agriculture: Digitalization is rapidly transforming the agriculture sector, and climate-smart agriculture (CSA) is no exception. Digital technologies can help farmers make better decisions, increase productivity, and reduce the environmental impact of agriculture. Brief Some of the important area of digitalization related to above field are given below:

a. Weather monitoring and forecasting: Digital tools such as weather stations, sensors, and satellite imagery can provide real-time weather data and forecasts. This information can help farmers make informed decisions about planting, harvesting, and irrigation

b. Precision farming: Digital technologies such as GPS, drones, and remote sensing can help farmers optimize the use of resources such as water, fertilizers, and pesticides. This can reduce input costs and environmental impact while increasing yields. Details for the same is given above.

c. Crop management: Digital platforms can help farmers monitor crop growth and detect pests and diseases early. This enables them to take timely action and reduce crop losses.

d. Market access: Digital platforms can help farmers access markets and get fair prices for their produce. For example, e-commerce platforms can connect farmers directly with consumers, cutting out intermediaries.

e. Data management: Digital tools can help farmers collect, store, and analyze data on crop yields, weather patterns, soil health, and other factors. This information can inform decision-making and improve farming practices over time.

f. Farm management: Digital platforms can help farmers manage their farms more efficiently. For example, farm management software can help farmers plan tasks, track inputs and outputs, and manage labour and equipment.

Digitalization in CSA is still in its early stages, and there are challenges to overcome, such as the digital divide in rural areas and the cost of technology. However, the potential benefits are significant, and digitalization can play a vital role in making agriculture more sustainable and resilient in the face of climate change.

Supply chain management: Digitalization is transforming supply chain management in agriculture by improving transparency, traceability, and efficiency. There are number of ways by which digitalization is quite beneficial for supply chain management in the field of agriculture. Digital technologies can capture and store data at the farm level, such as crop yields, soil health, and pesticide use. This information can help supply chain managers better understand the origin and quality of the produce they are sourcing. Digital tools such as block chain can help track the movement of produce along the supply chain. This enables supply chain managers to trace the origin of the produce, identify potential sources of contamination, and ensure compliance with regulations and certifications. Digital platforms developed for agriculture logistics can help optimize logistics, such as transportation and warehousing, by providing real-time visibility into inventory levels, shipping schedules, and delivery status. This can reduce waste, improve delivery times, and lower costs. Use of digital sensors and imaging technologies can detect defects and abnormalities in produce, such as bruises, discoloration, or molds. This can help supply chain managers identify and remove low-quality produce before it reaches the end consumer. Apart from this, digital tools can analyse data from across the supply chain to identify trends and patterns. This can help supply chain managers make informed decisions about sourcing, inventory management, and pricing. Further, digital e-commerce platforms can connect farmers directly with consumers, cutting out intermediaries. This can provide farmers with better prices and consumers with fresher produce. Digitalization in supply chain management in agriculture can help increase transparency, reduce waste, and improve efficiency. However, there are challenges to overcome, such as the cost of technology and the need for standardization across the industry. Nonetheless, the potential benefits of digitalization in agriculture supply chain management are significant, and it is an area that is likely to see continued growth and innovation in the years to come

Financial Inclusion: It is well known that Digitalization is a powerful tool for promoting financial inclusion in agriculture, particularly in developing countries where smallholder farmers often lack access to formal financial services. Success of Unified Payment Interface (UPI) which is a fast payment system, used largely for peer-to-peer payments and is also the most popular peer-to-merchant retail payment system in India with monthly transaction volumes of 8 billion in December 2022 (NPCI, 2022) is quite motivational for adoption of digital platform of financial inclusion agriculture. Digital technologies such as mobile banking can help farmers access financial services such as savings accounts, credit, and insurance. This can reduce the cost and complexity of accessing financial services, making it easier for smallholder farmers to participate in the formal economy. Digital platforms can facilitate payments between farmers, buyers, and financial institutions. This can reduce the need for cash transactions, which can be risky and time-consuming. Applications of digital technologies can analyse data such as farm yields, crop prices, and weather patterns to assess creditworthiness. This can enable financial institutions to make more informed lending decisions, reducing the risk of default and improving access to credit for smallholder farmers. The micro-insurance through digital platform assist farmers to purchase micro-insurance policies, which can protect them against risks such as crop failure, weather-related losses, and illness. This can reduce the financial vulnerability of smallholder farmers and provide a safety net in times of crisis. Also, digital platforms can enable farmers to participate in savings groups, which can provide a source of capital for investment in their farms. Digital tools can facilitate communication and record-keeping within these groups, making them more efficient and transparent. Further, these tools can provide farmers with access to financial education and training, enabling them to make informed decisions about financial management and investment. Digitalization is transforming the financial landscape of agriculture, and the potential benefits for smallholder farmers are significant. However, there are challenges to overcome, such as the need for reliable digital infrastructure and the importance of ensuring that digital financial services are accessible and affordable for all. Nonetheless, digitalization has the potential to improve financial inclusion and support the economic empowerment of smallholder farmers in agriculture.

Challenges: It was reported in studies that around 80% farmers adopted digital technologies in selected Asian countries. There is also increasing use of electronic extension services and social media in developing countries to share planting advice, weather updates, early disaster warnings, pest outbreak etc. Studies shows that technologies are being adopted separately instead of integrated or packaged together to address the specific problems of agriculture. The major challenges in adoption of these technologies are (i) extent of profitability should be beyond a threshold through in usage of the technology, (ii) amount of uncertainty and risk involved in adopting a technology, (iii) availability and suitability of the technology in the agricultural production cycle, (iv) requirement of skills and learning curve for adoption of the technology, (v) eco-environment and digital infrastructure availability in a particular region, (vi) level of agricultural production system, (vii) ease of flow of agricultural credits, and (viii) overall policy support for promotion of a technology.

There is a need of the inclusive digital transformation of Indian Agriculture empowering farmers at the centre of innovation. Awareness is to be brought on challenges and advocating is required for the improvement of the conditions and policies that limit the effective and inclusive use of digital tools to enhance agricultural production and allow positive social, environmental, and economic impacts. The development of digital agriculture needs to be

supported through dedicated funds for infrastructure and seed support to startups, easing regulatory compliances, investing in capacity-building programs, and developing digital infrastructure. A common platform for all agricultural activities, such as an AgriStack, would benefit all stakeholders. Collaboration between technology platforms and research organizations is necessary to develop digital tools for Indian agriculture. With concerted efforts and support, the digital agriculture sector in India has the potential to transform the agricultural landscape and pave the way for sustainable growth.