Int. J. Agricult. Stat. Sci. Vol. 14, No. 1, pp. ....-....., 2018 *ISSN : 0973-1903*

ORIGINAL ARTICLE

AGRICULTURAL DATA ANALYTICS – SMALL TO BIG DATA

S. Ravichandran* and K. Kareemulla

ICAR- National Academy of Agricultural Research Management, Hyderabad-500 030, India.
E-mail : srgmravi@yahoo.com

Abstract : Agriculture, the lifeline of Indian economy, is highly dependent on the timely occurrence of monsoons, *viz*. South-West and North-East monsoons spread over nearly six months *viz*. June – December of the year. India receives 75% of rainfall during South-West monsoon in June to September and Indian agriculture is good during this time of the year. Statistical data is collected on various aspects of crop growth through various stages and these data need to be statistically analyzed using advanced statistical techniques with the help of highly sophisticated statistical software developed for various statistical applications. The present paper throws light on various statistical techniques for analysis of data collected from various agricultural experiments.

Key words : Agriculture, ARIMA, Big data, Vector ARIMA (VAR).

1. Introduction

Agriculture is unquestionably the largest livelihood provider in India, more so in the vast rural areas. Around, 60% of the Indian population, directly or indirectly, depends upon Agriculture. It also contributes a significant figure to the Gross Domestic Product (GDP). Sustainable agriculture, in terms of food security, rural employment and environmentally sustainable technologies such as soil conservation, sustainable natural resource management and biodiversity protection are essential for holistic rural development. Indian agriculture and allied activities have witnessed a green revolution, a white revolution, a yellow revolution and a blue revolution. Indian Agriculture gears up for the second green revolution. Food production in India is also witnessing a steady increase. Food production jumped up from 50.82 million tonnes in 1950-51 to 272.0 milllion tonnes in 2017 because of advancement in technologies.

2. Methodology

Agricultural data collection and subsequent analysis using appropriate statistical techniques plays a vital role in achieving higher productivity. Data, in particular, agricultural data is collected either by conducting primary surveys or collected from secondary sources such as websites, published reports, journals and other sources. Primary data is collected by conducting various studies and by applying various sampling schemes such as Simple Random Sampling (SRS), Stratified Random Sampling, Systematic Sampling and other sampling procedures which are widely followed in agricultural surveys. Sampling methodologies are adopted over census because certain advantages such as time, cost, labour and other factors [Cochran (2004)]. Agricultural data can also be generated by way of conducting field experiments by following three basic principles of experimentation such as randomization, replication and local control and by adopting various statistical designs such as Completely Randomized Designs (CRD), Randomized Block Designs (RBD) and Latin Square Designs (LSD) [Federer (1955)]. Experimenter can also adopt various advanced designs such as Balanced Incomplete Block Design (BIBD), Partially Balanced Incomplete Block Design (PBIBD) and other designs available in literature depending upon the need of the experiment and experimenter. Designs such as Split Plot Design and Strip Plot Design can be adopted depending upon the availability of experimental material. Hence, data generation is a very important phase of Statistical Data Analysis. Once, the experiment is laid, data from various stages right from sowing to harvesting

*Author for correspondence **Received** January 06, 2018 **Revised** April 11, 2018 **Accepted** April 30, 2018

stages are collected by the experimenter. Once, the data are collected, these data have to be fully protected and will have to kept carefully as these data are obtained over a long period of time say months and in case of few experiments, it is spread over years. There are various stages in statistical data analytics. Generation of appropriate data in the first and foremost stage since without data, it is impossible to make planning. Data are obtained on number of parameters such as production, yield, area, rainfall, price, stock over a long period of time and are called time series data. These time series data on various parameters over a period of time are required for formulation of policies.

Exploratory Data Analysis [Gomez and Gomez (1984)] is the second stage and this stage corresponds to pictorially viewing the datasets using any of the exploratory techniques such as bar diagram, histogram, pie diagram, line diagram and other pictorial form of representation of data. These techniques help one to identify whether right dataset is obtained. Exploratory technique such as Scatter diagram and Box diagram will help the researcher to identify whether the data obtained falls in the proper range or not. These explanatory techniques will provide the experimenter about the outliers present in the data. Outliers in the data needs to remove otherwise the data will mislead the researcher whenever any decision making is involved. Scatter plots will indicate the dependencies or independencies of the parameters of datasets. Plotting of time series variables help to identify the presence of positive or negative trends in the datasets.

Statistical data analysis involves converting raw data to information. When data is converted to information, meaningful inferences can be drawn. Analytics deals with discovery, interpretation, and communication of meaningful patterns in data. Data generated from agriculture has to be processed for various measures of central tendencies such as mean, median, mode, geometric, harmonic mean for the dataset. It is also possible to find out the deciles and percentiles for the given data set. Analyzing data for various central tendency measures can be done using various software programs such as Statistical Package for Social Sciences (SPSS), Statistical Analysis System and R. R is a source is open source programming environment and analysis of data can be easily carried out using R and the codes for the execution of various statistical data analysis is freely available in web. Various R users' community is also available in public domain and hence data analysis can easily be done using R programming language. Measures of dispersions such as range, standard deviation, coefficient of variation with regard to the collected datasets could be easily be obtained using any of the standard software packages. Probability, a measure uncertainty, can also be found out using these software packages. Agricultural data collected from various experiments follow certain distributions such as Binomial, Poisson, Normal and other advanced distributions. Parameters such as mean, variance etc. can be calculated using various statistical software packages. When an experiment is conducted, the researcher not only collect data on parameters of interest (like yield, plant height, weight etc.), the researcher collects data on other parameters such as input parameters *viz*. flowering days, irrigation, rainfall etc. It may at sometimes necessary to calculate various such as covariance, correlation and regression of important parameters. Analysis of data for some of the above-mentioned measures could be carried out using SAS 9.3 software or by writing programming codes using R programming language. Exploratory data analysis for finding out correlation using Scatter Diagram could also be done for agricultural experimental data using R and SAS. In agriculture, ranking of varieties for their performances is also important and this task can easily be carried out using R and SAS. Spearman's Rank correlation for some of the important descriptors could also be carried out using the above mentioned software packages. Simple linear models are extensively utilized by agricultural scientists using data generated from agricultural experiments. Statistical analysis of data for certain parameters for fitting linear and non-linear regression modelling could be carried out using standard statistical software packages. Linear statistical models *viz*. linear regression models are mainly utilized by the agricultural scientific community where as scientists engaged in fisheries and animal science research employ lot of non-linear modelling techniques such as Fox model, Schaeffer model, Pella-Tomlinson model etc. Data generated in many of the fisheries' experiments have to be fitted using various software packages such as SPSS, SAS, R etc. Programming codes for fitting the above-mentioned non-linear models using data obtained from fisheries is available in public domain and can be utilized by the scientific community. Statistical modelling of generated data can be utilized

for forecasting since forecasting is an essential component of model fitting. Forecasting for future is important because of the fact that policy formulations can be carried out. This forecasting helps to tackle eventualities in case of exigencies.

As mentioned earlier, for generation of agricultural data, researchers need to conduct their experiments in the pot culture, laboratory or under field conditions using any of the standard experimental designs such as CRD, RBD, LSD, Split Plot, BIBD, PBIBD etc. In many of these experiments, aim of the researcher is to identify the best treatment among the available treatments. This task is easily carried out using any of the standard statistical packages such as SPSS, SAS, R etc. Conducting the experiment, selection of treatment and other methodological issues can be found out from various research papers published over a period of time by various researchers. Sample surveys are widely utilized in various agricultural and fisheries experiments. Conducting sample surveys and estimation of various parameters is available in literature [Cochran (2004)]. Estimation of parameters from various sample surveys can be done using SPSS, SAS and R. The estimation of parameters is very important in various cases in agriculture. For example, estimation of total food production of the country using multi-stage stratified sampling adopted by the scientific community is important because of the fact estimation of food production statistics would help the planners in formulating suitable strategies in case of drought or flood and other exigencies.

In agriculture, data on various parameters such as production, consumption, yield, prices of commodities are collected over time, which we call them as time series variables. Analysis of these time series variables can be handled using various time series data analysis approaches. Some of these approaches include Auto Regressive Integrated Moving Averages (ARIMA) modelling approach, Structural Time Series modelling Modelling, Vector ARIMA (VAR) modelling approaches. Some of these modelling approaches require that the data under consideration is stationary (both mean and variance are constant over time) and some are not. Some of the time series modelling approaches such as TAR, ARCH, GARCH, GARCH modelling [Tong (2013)] approaches are also utilized for modelling time series data in agriculture are collected over time. The analysis of these time series data by employing some of the above models could be carried out using SAS and R. Programming codes developed by users using R programming language are available in various user online forums.

Agricultural data collected on various parameters involving varieties and locations could also be grouped using techniques such as cluster analysis, can be performed using SAS and R codes. Grouping of varieties for various agro-climatic zones are required by the planners and hence this particular tasks can be carried out easily by the agricultural researchers. Performance of stability analysis for understanding the performance of varieties in various agro-ecological zones using techniques such as Eberhart and Russel, Perkins and Jinks can be easily done through various statistical software such as SAS and R very efficiently. Stability Performance of developed agricultural crop varieties are very important for the researchers, since this would help the breeding scientists involved in development varieties to plan their crop improvement policies.

In agriculture, data on various quantitative and qualitative descriptors are collected. Data analysis using multivariate statistical techniques such as Path Coefficient Analysis, MANOVA, Discriminant Analysis, Canonical Correlation Analysis can easily be carried out by utilizing SAS software by writing macros in SAS language. Statistical data analysis of the various multivariate statistical analysis can also be performed using R programming language by writing programs. Lucid description on various multivariate statistical techniques utilized by researchers in various fields of science are available in literature [Anderson (2003)].

Bioinformatics is an emerging area in agriculture. Analysis of chromosomal data etc. can be easily analyzed using various statistical techniques by making use of software like SAS, R and other advanced software. Agricultural data analysis for small datasets can be efficiently managed by using many of the statistical techniques discussed earlier. Results obtained by the above techniques for small datasets by employing various software applications are highly accurate, efficient and can be performed in less time. The memory capacity of present computer systems don't pose major challenges for performing analysis of small datasets. However, huge data is generated by agriculture in various locations of the country and the statistical methodologies developed for small data sets cannot be

utilized for the voluminous data. In this direction, effort is made to analyze huge data sets using Big Data Techniques employed by the researchers engaged in other areas of research.

3. Conclusion

Agriculture generates lot of data, proper analysis of these data by utilizing advanced statistical techniques, either small or big data, helps in reliable and efficient estimation of various agricultural related parameters, such as area, production and productivity of crops and also estimation of price, supply, demand and other market influencing parameters. For efficient and accurate estimation of parameters and subsequent forecasting, researchers engaged in agriculture should start using advanced statistical techniques, for both small and large data. Most of the statistical techniques used in statistical data analysis have been in use for a long period time. All these techniques are employed for analyzing small datasets, since availability of small datasets itself was difficult earlier. But nowadays, data generated in any field of science, especially, data generated in agriculture is huge and voluminous. Also, data analysis of these voluminous datasets is relatively easy and computer systems for analyses of big data is available [Bart (2014)]. Hence, researchers engaged in agricultural research and other activities should develop big data platforms in agriculture and also analyze the voluminous data generated using big data analytical techniques. Statistical analysis of big data can also be performed by using statistical software such as SAS and programs can also be written for solving Big Data analysis by making use of appropriate big data statistical analysis techniques.

References

Anderson (2003). *An Introduction to Multivariate Statistical analysis*. John Wiley and Sons.

Bart, Baesens (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. John Wiley & Sons, New York.

Cochran, W.G. (2004). *Sampling Techniques*. John Wiley and Sons, New York.

Federer, W.T. (1955). *Experimental Design Theory and Application*. The Mcmillan and company, Canada.

Gomez, K.A., and Arturo A. Gomez (1984). *Statistical Procedures for Agricultural Research*. John Wiley & Sons, New York.

Tong, Howell (2013). *Threshold Models in Time Series Analysis- Some Reflections*. Research Report. Serial No. 503. University of Hong Kong.