Role of MATLAB in Crop Yield Estimation

Raorane A.A. and Kulkarni R.V.

Abstract: Agriculture provides food security and strong economy for any country. According to the Food and Agriculture Organization, "Food security exists when all people have access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life." Food security is a major determinant of national security and self- sufficiency Matlab plays an important role in crop yield estimation. In this paper an attempt has been made to estimate crop yield for selected cash crops (Rice, Groundnut, Soyabin, Ragi) from the sample data collected from twelve talukas of Kolhapur district, Maharashtra state. Analysis and estimation of yield was done using process tool constructed by Matlab. One can use this model for estimating yield of any crop. The advances in computing and information storage have provided major factors for calculating and assessing the results. The challenge has been to extract knowledge from this raw data; this has lead to new methods and techniques such as data mining that can bridge the knowledge of the data to the crop yield estimation. This research aimed to assess these new techniques and apply them to the various variables consisting in the database to establish if meaningful relationships can be found.

Keywords- Yield estimation, process model, regression analysis, crop cutting experiments, sql server,Matlab

Introduction-

India today ranks second, worldwide in average annual agricultural output. Agriculture and associated sectors like agro-forestry and fisheries accounted for 16.6% of the GDP in 2009 along with 50% of the total workforce. The economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth. Agriculture is demographically the broadest economic sector and plays a major role in weaving the socio-economic fabric of India.

India is the world's largest producer of many fresh fruits, vegetables, milk, spices, fresh meats, fiber yielding crops such as jute; millets and oil seeds, etc. as per FAO world agriculture statistics (2010)(ref.).

Raorane A.A. is with Department of computer science, Vivekanand College, Tarabai park Kolhapur INDIA. abhiraorane@gmail.com and Kulkarni R.V. is working as Head of the Department, Chh. Shahu Institute of business Education and Research Centre Kolhapur. 416006 INDIA Email: drrvkulkarni@siberindia.co.in

Current scenario-

Today, *precision agriculture* refers to the application of modern GPS technology in connection with small-scale, sensor-based treatment of the crop. This introduces large amounts of data which are collected and stored for later usage. Appropriate use of this data often leads to gain in efficiency and thereby economic advantage. However, the amount of data poses a problem – which should be solved using techniques. One of the tasks that remains incomplete is *yield prediction* based on available data. This can be formulated and treated as a multi-dimensional regression task. This paper deals with appropriate regression techniques and evaluates it on selected agriculture data of Kolhapur District in Maharashtra state, India.

Kolhapur District Agricultural

Fertility of soils is determined by various macro and micro nutrients available in the soil. The Panchganga Basin, a well watered and agriculturally developed region covers 45752.2 sq.km area and supports 26, 11,547 (2.6 percent of state) population. The index values of N, P & K are collected from government soil survey and soil testing Laboratory, Kolhapur at village level. These index values of N. P. & K. are grouped into six categories and tahsil wise areas in percentage in concern category are computed. To recognized the fertility level of the soils composite index is computed with the help of NPK values and is grouped into five categories. In Kolhapur district, there is large variation in the distribution of macronutrients of the soil. It is observed that most of the areas of the study region are fertile in nature. Low and very low fertility of soil is noted in some pockets only. The physiography, climate and agricultural activities have greatly influenced the nutrients status of soil. Specific fertilizers and addition of organic matters are recommended for nutrients deficient areas which will help to keep the balance of nutrients and to restore the fertility of soils. Moreover, it is observed during the fieldwork that the anthropogenic influences are degrading the soils in the region which needs further investigations.

Technological Changes-

Role of technology in increasing agricultural and food production in the country is well known. However, adequate and convincing evidence on impact of improved technologies and policies followed during different periods since 1951 in reducing variation in production and resulting risk has been lacking. The issue of instability attracted lot of attention of researchers, in the early phase of adoption of green revolution technology, who found that adoption of new technology had increased instability in food grains and agricultural production in India. This conclusion was based on the period when improved technology had reached very small area. This study shows that when a little longer period

is taken into consideration, which witnessed spread of improved technology to large area, the inference on increase in instability due to adoption of new technology get totally refuted. Yield variability in food grains crops as well as in non food grains crops was much lower in the first phase of green revolution extending upto 1988 as compared to pre green revolution period. Volatility in yield, away from trend, witnessed further decline during 1989-2007. Production of non food grains show increase in instability during last two decades but production of food grains and total crop sector was much more stable in the recent period compared to pre green revolution and first two decades of green revolution in the Country. This indicates that Indian agriculture has developed resilience to absorb various shocks in supply caused by climatic and other factors. Food grains production remained more unstable as compared to production of group of non food grains crops. Instability in yield of cereal and pulses declined over time. However, opposite holds true for oilseeds. Oilseed production is also found more risky as compared to cereals and pulses. Among individual crops wheat, paddy and sugarcane are found least risky whereas bajra, groundnut, rapeseed/mustard, jowar and gram involves high risk. Pattern in area, yield and production instability of food grains differs widely across states. Yield instability was major source of instability in food grains production in most of the states. Production was most stable in the state of Punjab followed by Kerala. Haryana, Uttar Pradesh and West Bengal have brought down instability in food grains production sharply. Foodgrains production is highly unstable in the states of Maharahtra, Tamil Nadu, Orissa, Madhya Pradesh, Rajasthan and Gujarat.

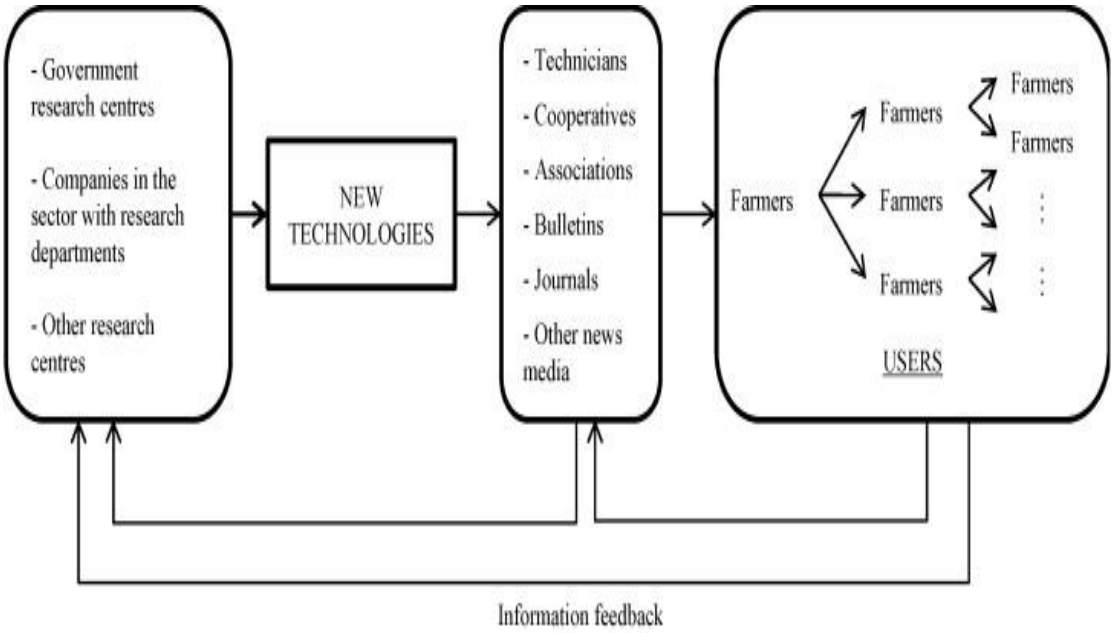
Researcher's own experience and situation of Agriculture in Kolhapur district

Yield of a crop is dependent on geo-climatic condition of an area. Farming in Kolhapur district is largely based on the past experiences of the producer in addition to the guidance by government departments. Yield estimates are annually predicted by the statistical department of the government. After analyzing the outcome or the yield, a farmer understands the problems in the process of cultivation, progress of crops, cropping pattern, effect of rainfall, soil parameters and effects of fertilizers and pesticides on the crop. Important question is whether the producers acquire this scientific data from various agriculture departments and if available, on time? And whether he has the capacity to analyze this information?

From the researcher's point of view, for estimation of yield three factors are important viz., area under cultivation, rainfall and soil fertility. After collecting this information, relation between these variables is established to predict crop production.

Methodology:

Literature survey and personal interviews are taken from various personality concerns with agriculture. In literature survey books, phd thesis, research papers, articles, conference papers were studied. In doing so the dialogue with the researcher, agriculturist, and government agriculture statistician carried away time to time for coming to final perfect model.



The researcher aiming to build a model should keep in mind that it can be handled easily by the end user. There should be a standard input form to incorporate data through the desk top. Later on the data is processed with appropriate statistical formulae to get results in the form of table, graph, etc.

After incorporating the data through input screen and entering the data in the database tables, the data in the database table is processed using various queries of SQL. This data is then connected to MATLAB for data mining. MATLAB simplifies the task of calculation by using various statistical libraries of formulae.

To check the validity of the data, before making the model using MATLAB, the data is analyzed in excel worksheet using all statistical functionality.

DATA MINING PROCESS MODEL

In the course of this project, analysis of real data sets, primarily agricultural data sets, provided by various agriculture organizations were carried out. From this experience a process model was developed for applying data mining techniques to data, with the goal of incorporating the induced domain information into a software module (Figure 1). The key points of this model (Garner et al, 1995) are;

- A two-way interaction between the provider of the data and the data mining expert. Both work together to transform the raw data into the final data set(s) input to the machine learning algorithm's-with the domain expert providing information about data semantics and 'legal' transformations that can be applied to the data, and the data mining expert guiding the process so as to improve the intelligibility and accuracy of the results.
- An iterative approach. Machine learning is an exploratory process; it generally takes several cycles through the process model to find a good "fit" between a representation of the data and a data mining algorithm. In addition, distinct attribute combinations running through different schemes can produce wildly different data models, even though the predictive accuracy of the results may be equivalent. These alternative views may provide valuable insights into patterns covering different subsets of the data.

In the model presented in following Figure, activity flows in a clockwise direction. In the pre- processing stage, the raw data is represented as a single table, as required by the data mining algorithm's. This table is translated into the ARFF format, an attribute/value table representation that includes header information on the attributes data types. The data may also require considerable 'cleansing', to remove outliers, handle missing values, detect erroneous values, and so forth.

At this point the data provider (domain expert) and the data mining expert collaborate to transform the cleansed data into a form that will produce a readable, accurate data model when processed by a data mining algorithm. These two analysts may, for example, hypothesize that one or more attributes are irrelevant, and set aside these extraneous columns. Attributes may be manipulated mathematically, for example to convert all columns containing productivity index of crop measurements to a common scale, to normalize values in a given column, or to combine two or more columns into a single derived attribute.

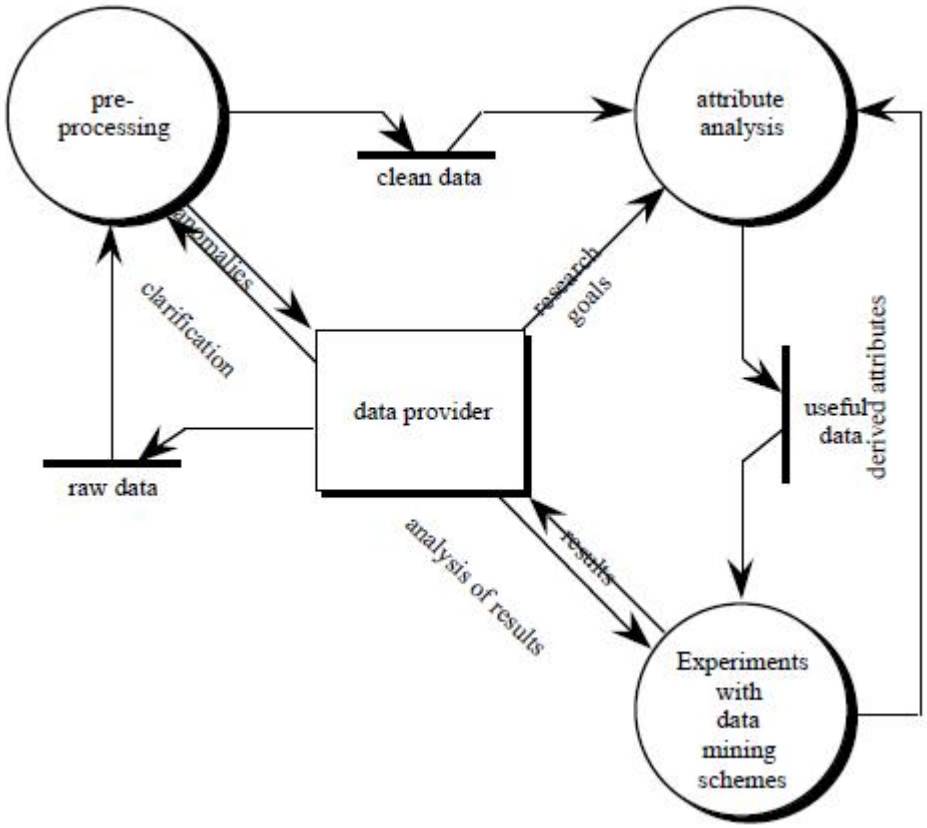


Figure- Process model for a machine learning application (data flow diagram)

One or more versions of the cleansed data are then processed by the data mining schemes. The domain expert determines which portions of the output are sufficiently narrative or interesting to deserve further exploration, and which portions represent common knowledge for that field. The data mining expert interprets the algorithm's output and gives advice on further experiments that could be run with this data.

Project Aim

Development of System

Hence, aim is to provide, not only an analysis of selected data mining tools available within MATLAB and a synthesis of these tools, but more importantly, a means to analyze and synthesize further data mining tools, thus providing an increasingly holistic view of the data mining capabilities of MATLAB.

Essentially then, researcher wish to discover the extent to which each of a number of MATLAB data mining tools is capable of carrying out the different stages of the data mining process. He wish to integrate these tools in order to bring greater clarity to the potential of MATLAB in the data mining arena. And, as he do this, to clearly define the methodology used in carrying out this work, in order that it might be used in future work in this area.

In summary, researcher aim is to create a means for obtaining a holistic view of the data mining capabilities of MATLAB. Researcher is accomplishing this by setting forth the methodology of this process and by demonstrating this methodology by investigating and synthesizing several data mining tools available for MATLAB.

Project Motivation

MATLAB is a powerful and flexible tool, of performing data mining. It is clear that MATLAB has not been given appropriate attention in this area. MATLAB is not yet in the league of packages such as Clementine, Weka and even Excel. In addition, though MATLAB is chosen more frequently than Oracle, it is generally used in conjunction with other tools. Whereas Oracle is implemented as a standalone tool over 50% of the time, MATLAB is used on its own just over 12% of the time.

Data Mining Model building [Regression analysis using MATLAB]

Objective of the present work is to formulate a perfect model for the agricultural yield estimation. With the objective in mind vast data was collected in relation to the yield estimation. But, as the entire data is not immediately relevant for model making, first we have to clean the data and rearrange it as suitable for the selected statistical method. Then different experimental models are worked out and if the model does not interpret as expected researcher rearranged the model structure or check the viability of the data for his model. This procedure is repeated until the final model works out. For example, construct the fitting of the regression model for estimation or prediction of the crop yield. Initially there was only one model for the district, but the results obtained w.r.t. the estimation of crop yields by using the model did not match the actual crop yield. So, the model was rearranged for working suitably at taluka level. Again taluka wise data was collected and analyzed.

The above models have been used is for analysis of yields of four crops viz. Paddy, Soybean, Groundnut and Ragi, because they all are rain fed crops. Twelve tehsils and four crops, meaning 48 models were designed. As some crops were not cultivated in some talukas, there is no available data, e.g. in Shirol, Ragi is not cultivated, in Gaganbawada Soybean is not cultivated, etc.

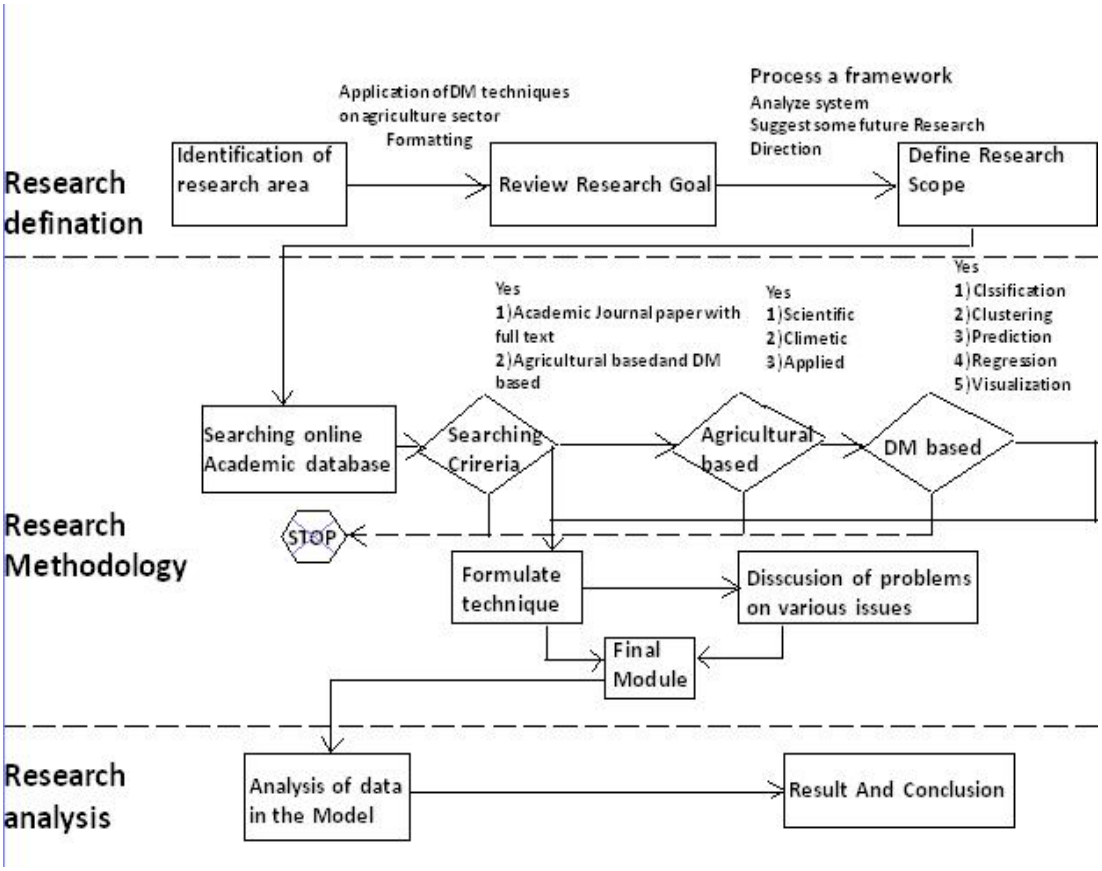
These above tasks are carried out using Excel. But when the complex data analysis for huge data have to be carry out, its difficult to arrange the data and memorize the various steps which we carried out past. So to overcome this difficulty researcher checked various option to tackle this. So he came to final strategy which is describe below

1) Consider the data as database.
2) Select the suitable RDBMS for the data.
3) After completing the table design in database.
4) Access the data in data mining tool [Matlab] for final calculation using various predefined tools available in MATLAB.

The average yield, standard deviation, partial and multiple correlation coefficients have been calculated and fitted in multiple regression planes. In the end principal component from the variables was calculated.

Multiple Regression: Some statistical methods serve as forecasting (or estimation) techniques. One of such techniques is regression analysis. We are familiar with linear regression and correlation of two variables. It is called as simple correlation and regression.

However, in practice, we observe that, the variable under study is influenced by two or more variables. Hence, two variables are not sufficient to describe it. e.g. National income is based on several variables such as agricultural yield, industrial production, import, export, production of minerals, marine wealth etc. A variable whose numerical value is to be predicted is taken as dependent variable or response variable (National income) and remaining all variables are treated as independent variables or explanatory variables.



The regression analysis based on the dependent variable and two or more independent variables is referred as multiple regressions.

The correlation coefficient between X_1 and X_2 is

$$r_{12} = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2} = \frac{\sum X_1 X_2}{n\sigma_1 \sigma_2} \quad \text{or}$$

$$\sum X_1 X_2 = n\sigma_1 \sigma_2 r_{12}$$

The correlation coefficient between X_2 and X_3 is

$$r_{23} = \frac{Cov(X_2, X_3)}{\sigma_2 \sigma_3} = \frac{\sum X_2 X_3}{n\sigma_2 \sigma_3} \quad \text{or}$$

$$\sum X_2 X_3 = n\sigma_2 \sigma_3 r_{23}$$

The correlation coefficient between X_1 and X_3 is

$$r_{13} = \frac{Cov(X_1, X_3)}{\sigma_1 \sigma_3} = \frac{\sum X_1 X_3}{n\sigma_1 \sigma_3} \quad \text{or}$$

$$\sum X_1 X_3 = n\sigma_1 \sigma_3 r_{13}$$

r_{12}, r_{13}, r_{23} are called total correlation coefficients.

Note that R is symmetric matrix. |R| denotes determinant of R is given by,

$$|R| = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 1\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} - r_{12}\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} + r_{13}\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}$$

$$= 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}$$

Co-factors of the elements in the first row are given by,

$$R_{11} = (-1)^{1+1}\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} = 1 - r_{23}^2$$

$$R_{12} = (-1)^{1+2}\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} = r_{13}r_{23} - r_{12}$$

$$R_{13} = (-1)^{1+3}\begin{vmatrix} 1 & r_{13} \\ r_{13} & 1 \end{vmatrix} = 1 - r_{13}^2$$

Fitting of Regression Planes:

Considering Researchers regression equation, the equation of regression plane of X_1 on X_2 and X_3 where X_1 : Productivity Index, X_2 : Rainfall and X_3 : Soil fertility index

The equation of regression plane of X_1 on X_2 and X_3 is given by

$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$

Where a, $b_{12.3}$ and $b_{13.2}$ are constants to be determined by the method of least squares.

The required equation of regression plane of X_1 on X_2 and X_3 is given by,

$$\therefore \frac{R_{11}}{\sigma_1}\left(X_1 - \overline{X_1}\right) + \frac{R_{12}}{\sigma_2}(X_2 - \overline{X_2}) + \frac{R_{13}}{\sigma_3}(X_3 - \overline{X_3}) = 0$$

It is used to estimate X_1 : Productivity index, when X_2 : Rainfall and X_3 : Soil fertility index for the crop for given year are known. Then value of X_1 : Productivity Index obtained is the estimated value for that year. It compared with the actual i.e. observed value of X_1

Stages in data analysis for crop estimation

The crop estimation involved the following important steps,

1] Finalization of variables

2] Finalization of Formulization

3] Model fitting using Excel

4] Summarization of results from Excel analysis

Data Mining

From the analyzed datasets, researcher extract the knowledge i.e. estimation of crop yield considering historical data.

Database design

Arrange the data in database for better data transformation. For this process researcher selected Sql Server version 2.5.

1] Establish E-R diagram.

2] Normalize E-R diagram.

3] Draw table description diagram.

4] Convert this into table with primary key, foreign key, and other constraints.

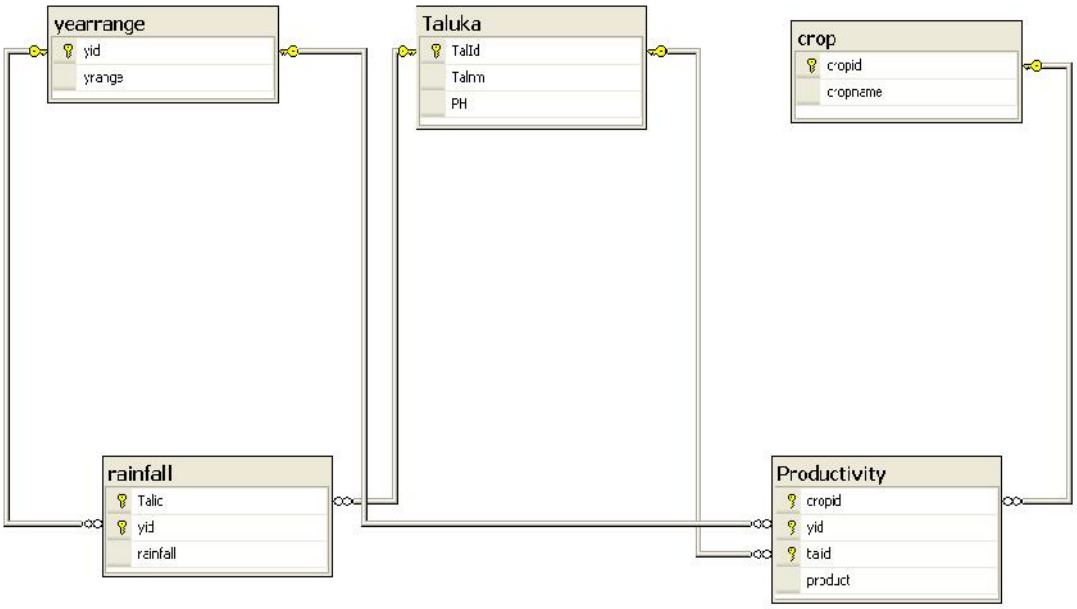


Fig. Data Base

Using Sql server

After completing the Database design which include factors like Productivity index, rainfall, soil fertility index for four crops Rice, Groundnut, Soybean and Raggi for 12 Talukas. Researcher concentrated mainly to design database in such way that user can easily input the data of any variable associated with crop estimation into developed software. Researcher has taken due care about this data which is analyzed properly through the software, to get correct results of estimation.

For Data connectivity to Sql server, researcher used ODBC (Open database connectivity) with user DSN (Data source). An ODBC user data source stores information about indicated data provider.

Using MATLAB

Researcher connected this database to the MATLAB for statistical analysis. Researcher has developed best algorithm which suit the estimation model. Researcher used MATLAB version R2010 for his software development.

MATLAB code for connectivity

Conn = database('data source name';'user name',' password')

This connects a MATLAB software session to a database via. ODBC driver and assign the returned connection object to conn. The arguments passed are as follows

1) Datasource name- The data source to which you connect.
2) Username and password – Password and user name to connect the database.

Sql query execution in MATLAB

The query written for yield estimation in MATLAB, the syntax is as follows

Exec(conn, 'sql query')

e.g. exec (conn, 'use yield Estimation')

A database object called curser has been used. Running exec returns the cursor object to the variable curs and returns additional information about the cursor object.

A datareturn format is property of MATLAB which converts the table into structure format because default conversion will be array which is inconvenient for formulation. This is carried by means of following code

Setdbprefs('property','value')

-setdbprefs('datareturnformat','structure')

MATLAB Cursor

Using the function fetch, converts output of select queries into appropriate format. Researcher used structured format for data analysis, so following commands are used

Curs=exec(conn,'select query');

RS=fetch(curs); RS= Record set

RS.data.column1

Rs.data.column2

Above code represents columnwise data of selected queries.

e. g. Taluka(TalID, Talnm)

TalID	Talnm
1	Ajara
2	Bhudargad
3	Gadhinglaj

RS = fetch(conn,'select *from Taluka')

RS.Data.TalID	RS.Data.Talnm
1	Ajara
2	Karvir
3	Gadhinglaj

Advantages of using Matlab

The researcher had made choice of MATLAB because it provides efficient and accurate platform for the research work for Data mining. Some of the advantages of MATLAB are as follows

1] Already predefined complex statistical modules available. So there is no requirement of writing the separate code.
2] The code is compact because of predefined modules used in the code.
3] Less time required to test the complex procedure.
4] User can write his own function which he can use many times when required for data analysis.
5] Very large database can be handled.
6] Open source system.

After completing these stages, researcher processed the data in data mining model, which is constructed using Matlab.

Summery

In this paper the author has exposed to basic technologies of Data Mining and basic description of how Data Mining architecture can be develop to deliver value of data mining to the end user.

One of the task remain incomplete is yield prediction based on available data. The author used Data Mining perspective which can be formulated and treated as a multidimensional regression task. This paper deal with appropriate regression techniques and evaluate it on selected agriculture data.

Refernces

[1] Data mining Techniques for Predicting Crop Productivity – A review article S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh IJCST Vol. 2, Issue 1, March 2011

[2] Chapman P. Gleason LARGE AREA YIELD ESTIMATION/ FORECASTING USING PLANT PROCESS MODELS By Chapman P. Gleason For Presentation at the 1982 Winter Meeting AMERICAN SOCIETY OF AGRICULTURAL ENGINEERS Palmer House, Chicago, Illinois December 14-17, 1982

[3] R S Deshpande AN ANALYSIS OF THE RESULTS OF CROP CUTTING EXPERIMENTS Agricultural Development and Rural Transformation Unit Institute for Social and Economic Change February 2003

[4] Ramesh Chand, Sanjeev Garg and Lalmani Pandey "Regional Variations in Agricultural Productivity A District Level Study" *in 2009* for National Professor Project

[5] Day, R. H., "Probability distribution of field crop yields" Journal of farm Economics 47(1965) 7B – 741.

[6] Dorfman J.H. "Should normality be a normal assumption?" Economic letters 42 (1993) 143 – 147

[7] Gallagher P. "U.S. Soyabean yields: estimation and forecasting with non symetric disturbance" American Journal of Agriculture Economics 69. (Nov. 1987) : 796 .803.

[8] Norwood B. Roborts, M.C. "Lusk J. L. "Ranking Crop yield model using out of sample likely wood functions" American Journal of Agriculture Economics 86. (4) (Nov. 2004)1032.1043

[9] Just R. E. Weinenger Q. "Are crop yields normally distributed" American Journal of Agriculture Economics 81. (May 1999): 287. 304.

[10] Ramirez, Misra & Filed – " Crop yield distribution revisited"
American Journal of Agriculture Economics 2003. Volume 85: 108
[11] Georg Ruß Data Mining of Agricultural Yield Data: A Comparison of Regression Models, ICDM'09,. Leipzig, Germany, July 2009
[12] V. Ramesh and K. Ramr Classification of agricultural land soils: A data mining approach" International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 1 Jan 2011 379
[13] Rainfall variability analysis and its impact on crop productivity Indian, Indian agriculture research journal 2002 29,33.,8) SPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing support to crop yield forecast and area estimates GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST by Roberto Benedetti A, Remo Catenaro A, Federica Piersimoni B
[14] "Risk in Agriculture: A study of crop yield distribution and crop insurance" by Narsi Reddy Gayam Thesis (M. Eng. in Logistics)--Massachusetts Institute of Technology, Engineering Systems Division, 2006. Includes bibliographical references (leaves 52-53).
[15] Gazetteer of Kolhapur District (2001)
[16] Aditya, Kaustav (2008). Forecasting of crop yield using discriminant function technique. M.Sc. thesis, PG School, IARI, New Delhi.
[17] Agrawal, Ranjana, Jain, R.C. and Singh, D.(1980). Forecasting of rice yield using climatic variables. Ind. J. Agric. Sci., 50 (9), 680-684.
[18] Agrawal, Ranjana and Jain, R.C. (1982). Composite model for forecasting rice yield. Ind. J. Agric.Sci., 52 (3), 189-194.
[19] Agrawal, Ranjana, Jain, R.C. and Jha, M.P. (1983). Joint effects of weather variables on rice yields. Mausam, 34 (2), 177-181.
[20] Agrawal, Ranjana, Jain, R.C. and Jha, M.P. (1986). Models for studying rice crop weather relationship. Mausam, 37 (1), 67-70.