Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-981-07-6260-5

Region Specific Crop Yield Analysis A Data Mining Approach

D Ramesh B Vishnu Vardhan

Abstract— **Indian Gross Domestic Product (GDP) is largely depends on agricultural production. Agricultural production is the product of cultivated area and average yield per unit area. Its impact on the welfare of the country is much greater and nearly 70% of the working population depends on agricultural activities for their livelihood. Agrarian sector in India is facing a severe problem to maximize crop productivity. More than 60 % of the crop in India still depends solely on monsoon rainfall (Central Statistical Organization, 2008). An analysis of past climatic variations and its impact on agricultural production is presented in this paper to get the climate variability on agriculture. Recent development in Information Technology enabled new methods to adopt in agriculture sector. High end servers and the latest software in analytics are useful for predicting certain crop production. Clustering approach is used for estimating the future year's rice production based on average rain fall of specific region. Multiple Linear Regression is a statistical data mining technique which is adopted in this paper for predicting rice yield expectation in the regions of Andhra Pradesh, India.**

Keywords— ***Agrarian Sector, Clustering , Data Mining, Gross Domestic Product, Multiple Linear Regression.***

I. Introduction

A large network of rivers, suitable soils and the Biodiversity is making India rich in Agriculture sector. Indian Agricultural policy focuses on food self sufficiency and remunerative prices for farmers. Agriculture is the back bone of Indian Economy. The majority of farmers in India are not getting expected Crop Yield due to several reasons. Though, successful research practices are observed in Agricultural sector, expert advice regarding the rice cultivation is not reaching the farmers community.

In this context the farmers necessarily requires a timely advice to get a competitive production in their crops. Agriculture production is the product of cultivated area and an average yield per unit area. The maximization of the crop productivity has become a serious problem which needs to be addressed. India still depends solely on monsoon rainfall. The climate variations need to be addressed and an analysis is to be made in order to help the farmers to maximize the crop productivity.

Among the food grains cultivated in India, Rice ranks first both in terms of area as well as production. The average rice yield per hectare is very less when you compare with other countries like China, Japan, Korea and Australia and the world average is 27 q/ha. Rail fall and its duration is an important for rice cultivation. Rice is more suited to high rain fall regions because it requires abundant moisture. Rainfall conditions influences the rice cultivation. In general the influencing factors for rice production are rainfall, temperature, day length and humidity. They are known to cumulatively influence the total production and area of rice in the world. India has two crop seasons, Kharif and Rabi, based on the dependence of crop productivity on monsoon. The Kharif season is during summer months of monsoons lasting from April to September. The Rabi season is during winter month from October to March. The highest concentration of rain fed agriculture occurs in western and southern areas in India where much of the rice is rain fed.

In this paper our aim is to create a user friendly interface for farmers, which gives the analysis of rice production based on the monsoon rainfall. The data mining Techniques and statistical analysis tools are used for the approximation and maximizing the crop productivity. The hardware and the software related Data Mining and Ware Housing tools are useful to extract knowledge from huge databases and statistical methods were used to predict the future crop productivity.

Data Mining is a non trivial extraction of previously unknown, potential, useful and reliable patterns from a set of data. The process of Data Mining is analyzing the data from different perceptive and summarizing it. Many Data Mining techniques were developed over the years, out of them some are conceptually very simple and some other are more complex and led to formulation of the global optimization problem. The Agricultural Data involves many of the parameters which are in complex nature. In order to get the proper analysis on the Agricultural data, one needs to split the data in different categories. In general, the regression techniques and principal component analysis commonly used for finding patterns in Agricultural data sets.

Data Mining is widely applied to agricultural problems. Weather forecasting is improved by K Nearest Neighborhood approach where it is assumed that the climate during a certain year is similar to the one recorded in the past. Knowing the

D. Ramesh
Associate Professor of CSE,
JNTUH College of Engineering,
Nachupalli, Karimnagar Dist., Andhra Pradesh, India.

B. Vishnu Vardhan
Professor of CSE & Head Department of I.T.,
JNTUH College of Engineering,
Nachupalli, Karimnagar Dist., Andhra Pradesh, India.

weather a day or a weak in advance is very important especially in agriculture. The variability of the climate is one of the most important factors which impacts agricultural productions. Though, the present mediums are able to provide accurate forecast of the weather for the next few days. But forecasting the weather conditions four to six months ahead of time is a big challenge. The uncertainty about the weather can be devastating in agriculture and sometimes the farmers may not be prepared to face the weather conditions which results in poor productivity.

II. Related Work

Soil profile descriptions were proposed by Verheyen [1] for classifying soils in combination with GPS based technologies. They were applied K-Means approach for the soil classification. In a similar approach crop classifications using hyper spectral data was carried out by Campus Valls [2] by adopting one of the data mining approach i.e. Support Vector Machines. Meyer [3] et al. used an intensified fuzzy cluster analysis for classifying plants, soil and residue regions of interest from GPS based colour images. Most of the work carried out by the researchers involved with data mining techniques coupled with image processing. Similarly independent component analysis for spatio temporal data is applied by Basak [4] to mine for patterns in weather data using North Atlantic Oscillation (NAO).

In an another exploration Jiawei Han [5] used K- Means algorithm for getting temperature and rainfall as initial spatial data and analyzed agricultural meteorology for the enhancement of crop yield. With this process the object is assigned to a cluster to which it is most similar based on the distance between object and cluster mean. Then it computes new mean for each cluster and thereby iterates until the criterion function converges. While experimenting the spatial data Chi-Farn Chen [6] et al. were used clustering techniques where frequent item sets were found using cluster analysis. In clustering the objects with high similarity are grouped together as a cluster. In general dissimilarity is assessed based on the attribute values describing the objects. Using such clustering mechanism a knowledge discovery is carried out in their work.

Similar clustering analysis is carried out by Wang [7] which is used to identify clusters embedded in the data. It can be expressed by distinct function, specified by users or experts. Guha [8] et al. was carried out their research on clustering of data streams which produces high eminence clusters to make it certain that inter-cluster resemblance is low and intra-cluster similarity in high.

In this paper clustering is being considered for making similar objects together. With respect to the available agricultural data we made clusters based on the mean of the rainfall and similar objects are made as a single cluster. As we are aware that the climate during certain year in similar to the one recorder past. If one could able to get cluster of years of common rain fall then it is easy to find out mean yield for those clusters. Similarly using Expectation Maximization Theory one can extract the expected mean crop based on the monsoon period.

III. Hierarchical Clustering Technique

Clustering techniques are divided in hierarchical and partitioning. The hierarchical clustering approach builds a tree of clusters. The root of this tree can be a cluster containing all the data. Then, branch by branch, the initial big cluster is split in sub-clusters, until a partition having the desired number of clusters is reached. In this case, the hierarchical clustering is referred to as divisive. In order to have randomness in the estimations a single parameter rain fall is considered. A process is being evolved to divide the data in to cluster format and evaluation is carried out which can give the near approximation results of rice production for a particular year for a specific region.

In this process four clusters were formed by considering the rainfall for the years 1958 to 2008. Here rain fall means average rainfall for that year for the specific region. Once the process of clustering is completed by taking the parameters of the rain fall, each cluster may be grouped with different years. This clusterization is carried out by taking the Centroid and median of the rainfall values of a specific region. After forming clusters based on rainfall, the average production is mapped to each cluster.

TABLE I. CLUSTERS ARE FORMED BASED ON AVERAGE RAIN FALL OF A SPECIFIC REGION

S.No	Years of Cluster 1	Avg Rain fall	Years of Cluster 2	Avg Rain fall	Years of Cluster 3	Avg Rain fall	Years of Cluster 4	Avg Rain fall
1	1960	1.52	1958	1.71	1955	2.29	1957	3.36
2	1963	0.95	1959	1.71	1956	2.22	1962	2.79
3	1965	0.66	1961	1.78	1964	2.16	1975	2.75
4	1968	1.51	1967	1.63	1966	1.9	1981	2.34
5	1970	1.33	1969	1.7	1973	1.82	1988	3.03
6	1971	1.44	1972	1.74	1974	2.29	1991	2.9
7	1976	1.02	1983	1.58	1977	2.22	1996	2.44
8	1982	1.46	1984	1.79	1978	2.12	1998	2.42
9	1985	1.26	1986	1.77	1979	2.24	2000	2.54
10	1994	1.39	1987	1.71	1980	1.98	2001	2.37
11	1995	1.39	1989	1.69	1993	2	2004	2.52
12	2002	1.08	1990	1.64	1999	1.98	2005	6.43
13	2003	1.51	1992	1.78	2006	2.33	2007	2.93
14			1997	1.63	2008	2.08		
15			2009	1.58				

Table 1 presents the four clusters along with the average rain fall of that year where cluster 1 and 4 consists of 13 years cluster 2 comprises with 15 years, cluster 3 comprising of 14 different years. In the cluster 1 the least average rain fall being 0.66 and highest being 1.52. Similarly cluster 2 has 1.58 as minimum and 1.78 as maximum average rainfall. Cluster 3 is ranges in between 1.9 to 2.33 and finally cluster 4 is ranging between 2.34 to 6.43. Now Multiple Linear Regression mapping is applied on these clusters with their corresponding productions.

Clustering techniques are used when there is no previous knowledge about the data. When a training set is available, classification techniques can be applied. In such cases, the training set is exploited for classifying data of unknown classification. The training set can be exploited in two ways: it can be used directly for performing the classification, or it can be used for setting up the parameters of a model which fits the data. A test case is been evaluated by considering a specific year's average rain fall there by mapping it to the existing clusters. Once the cluster is being identified then based on the average rain fall, one can predict the approximate rice production for that year of any region. In general Indian rice production is consisting of two seasons i.e. kharif and Rabi, in order to achieve the approximate rice production for future years it can be estimated by considering the average rain fall of a specific region for that year. We can give tentative rice production for coming years with one variable i.e. average rain fall of the year of the specific region.

IV. Analysis of Proposed Model

Multiple Linear Regression (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variable(s). The dependent variable is sometimes termed as predictant and independent variables are called predictors. MLR is based on least squares and probably the most widely used method in climatology for developing models to reconstruct climate variables from tree ring services. In this present paper a model is presented using MLR technique where the predictant being the rain fall of the year and there are five predictors namely "Area of Sowing", "Yield", "year", rainfall and soil parameters which includes Nitrogen(N), Phosphorous(P) and Potassium(K).

The proposed model is defined as follows:

Model Equation :

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \dots\dots + b_k x_{i,k} + e_i \qquad (1)$$

where
$x_{i,k}$ = value of k^th^ predictor in year i
b_0 = regression constant
b_k = coefficient on the k^th^ predictor
k = total number of predictors
y_i = predictand in year i
e_i = error time

Prediction equation :

The model is estimated by least squares which yields parameter estimates such that the sum of squares of errors is minimized. The resulting equation is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + \dots\dots + \hat{b}_k x_{i,k} \qquad (2)$$

where '^' denotes estimated values.

V. Model Implementation

Three data sets were acquired from three Govt. departments. First data set procured from Indian Meteorological department for the years 1951 to 2011. This rain fall data consisting of year wise rain fall of the above said years and specific to all the regions of India. This data is processed in such a way that the regions of Andhra Pradesh (In terms of longitude and latitude) were mapped and stored in the data base. The rail fall data of 25 districts of the region are the processed zones for the analysis purpose. The second phase of the data collection is made in Agriculture department of Andhra Pradesh where specific information such as rice yield, area of sowing, yield per hectare, year, and soil parameters were obtained. In the last phase of data collection is with Indian Statistical department where the data collection is made specific to crop yield for a specific region and specific to a year. The data collected from various places are well suited for classifying decision making behaviour of the region.

In this micro level analysis the data is processed with respect to district level where each district parameters were mapped with years. Average "rain fall" of the year and "Rice yield" were the two parameters of interest and clusters were made based on these parameters.

The estimation process of crop production is carried out by considering two variants analysis first being rainfall for 58 years and the rice production i.e. yield per hector for all the districts of A.P state in India. Using the equation of Multiple Linear Regression an analysis is made 5, 10, 15, and 20 cumulative years to consolidate the approximate estimation of the crop. For the years of consolidation, 3, 4, 5 and 6 variant (predictor) analysis is carried out and y_i is been calculated which is the outcome from the equation (1). The estimation is carried out by the least square and $\hat{y}_i$ is the estimated production of predictant year. A 10 years 5 variant analysis of a specific region is given in the Table 2.

The stability aspect of rice production is involved with certain temporal variables such as soil conditions, rain fall and area being sowed. The soil conditions necessarily involves with the usage of Nitrogen, Phosphorous and Calcium values. All the above said temporal variables processed with Multiple Linear Regression model for a span of 10 years.

TABLE II. A TEN YEAR FIVE VARIANTS ESTIMATION OF A SPECIFIC REGION

SNO	Area of sowing In Hectars	Rainfall	Yield per Hectar	P	K
1	247195	3.04	2622	53877	17059
2	261992	3.02	2536	51701	22446
3	267083	2.47	2638	65571	25490
4	272901	2.2	2911	63152	21039
5	222720	2.28	2589	53378	21552
6	214554	1.81	3072	45925	25970
7	232235	2.88	3499	50098	35350
8	244235	2.52	3448	56164	47451
9	253521	4.27	1877	63638	40493
10	245236	3.33	2781	61270	34002
11	253544	3.38	3254	65309	50046
12	255444	3.44	3048	67728	53038
13	212887	1.75	3101	65985	50881
Estimation time	**Year of prediction**	**Estimation By MLR approach**	**Exact production of that year**	**Variation or Diff**	**Per_%**
1987-1996	1997	634557.8	630371	-4186.8	-0.7
1988-1997	1998	653122	646371	-6751.01	-1.0
1989-1998	1999	672587.1	685440	12852.86	1.9
1990-1999	2000	765123.4	794415	29291.62	3.7
1991-2000	2001	548821.9	576622	27800.12	4.8
1992-2001	2002	676454.5	659110	-17344.5	-2.6
1993-2002	2003	821578.3	812590	-8988.27	-1.1
1994-2003	2004	804429.7	842122	37692.32	4.5
1995-2004	2005	462459.7	475859	13399.34	2.8
1996-2005	2006	695255.7	682001	-13254.7	-1.9
1997-2006	2007	821926.4	825032	3105.603	0.4
1998-2007	2008	782273.6	778593	-3680.57	-0.5
1999-2008	2009	688959	660163	-28796	-4.4

The next year estimation and actual production was compared and the difference was obtained in terms of percentages presented in the Table 2. In the percentages, negative sign means over estimation. But all the estimations for specific region are varying in between 0.4 % to 5 %. And the estimation of rice production for a specific region West Godavari is presented for the years 1997 to 2009. Such analysis is carried out for all the regions of A.P. for 5, 10, 15 and 20 years of estimation.

VI. Conclusions

In this paper an attempt is made by applying data mining techniques on Agrarian sector with the exponential growth of information pertaining to various fields, it is necessary to incorporate certain data mining techniques for various fields. In this process certain statistical methods were adopted in order to estimate Crop Yield analysis with existing data. Though there are different parameters were tested, our focus is 3,4 and 5 important variables. The comparison is made with existing Crop Yield to that of estimated results obtained from our analysis.

In the process of estimation we carried out the analysis by different ranges of year boundaries such as 5 year, 10 year and 15 year estimations. As we have explained earlier a 5 year analysis is obtained by taking 5 year span intervals and then the prediction is made by applying the model. These estimations were compared with the subsequent year's actual production . The net result we obtained is ranging between 90% to 95% accuracy.

References

[1] Verheyen K, Adriaens D, Hermy M, Deckers S., 2001,"High-resolution continuous soil classification using morphological soil profile descriptions". Geoderma Vol.101: pp. 31–48

[2] Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero JD, Moreno J., 2003, "Support vector machines for crop classification using hyper spectral data". Lect Notes Comp Sci 2652: pp. 134–141.

[3] Meyer GE, Neto JC, Jones DD, Hindman TW, 2004, "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images". Computer Electronics Agric Vol. 42: pp. 161–180.

[4] Basak J., Sudharshan, A., Trivedi D., M.S.Santhanam. 2004."Weather Data Mining Using Independent ComponentAnalysis". J. of Machine Learning Research 5: pp. 239-253.

[5] Jiawei Han, Member and Yongjian Fu, Member, "Mining Multiple-Level Association Rules in Large Databases" , IEEE transactions on knowledge and data engineering, vol 11, no.5, September/October, 2000.

[6] Chi-Farn Chen; Ching-Yueh Chang; Jiun-Bin Chen **"**Spatiial knowledge discovery using spatial datamining method", Geoscience and Remote Sensing Symposium, IEEE International Volume 8, Issue 25, pp. 5602 - 5605 July 2005.

[7] Wang, K., Xu, C., & Liu, B., 1999. Clustering transactions using large items. International Conference on Information and Knowledge Management, CIKM 99 Kansas city , Missouri United States, pp. 483–490.

[8] Guha, S., Mishra, N., Motwani, R., & O Callaghan, L.,2000. "Clustering data streams" Symposium on Foundation of Computing Science, pp. 359-366.

About Author (s):

D.Ramesh was graduated from ANU, Guntur, Post Graduate from JNTU Hyderabad, pursuing Ph.D from JNTU Kakinada and having 14 years of experience in Teaching. Presently working as Associate Professor of CSE in the Department of IT, JNTUH College of Engineering Karimnagar, a constituent college of JNTU Hyderabad.

B.Vishnu Vardhan received Doctorate in CSE in 2008 from JNTU Hyderabad and published 21 research papers in National / International Journal / Conferences. He has vast academic experience in Teaching and presently working as Professor of CSE and Head of the Department of IT, JNTUH College of Engineering, Karimnagar, a constituent college of JNTU Hyderabad.