APPLICATION OF DATAMINING TOOL TO CROP MANAGEMENT SYSTEM

Raorane A.A., Assistant Professor
Vivekanand College, Kolhapur, India
E-mail: abhiraorane@gmail.com

Dr. Kulkarni R.V., Professor
Chh. Shahu Institute of business Education and Research Centre, Kolhapur, India
E-mail: drvkvkulkarni@siberindia.co.in

ABSTRACT

It is important to estimate crop production for various policy decisions relating to storage, distribution, pricing, marketing, import-export, etc. The estimation of major crops was issued by Directorate of Economics and Statistics. However, these advance estimates are only gustimates and not the objective estimates. Researchers have concentrated on estimating lots of subjective assessment based on several qualitative factors. Thus there is a need to develop statistical model to forecasts crop acreage and production. In this paper authors have developed integrated crop model 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 data mining 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 data mining techniques and apply them to the various variables consisting in the database to establish if meaningful relationships can be found.

KEY WORDS

Yield estimation; Data mining; Regression analysis; Crop cutting experiments.

In the past decades, information technology (IT) has become more and more part of our everyday lives. With IT, improvements in efficiency can be made in almost any part of industry and services. Nowadays, this is especially true for agriculture. Due to the modernization and better affordability of state-of-the-art GPS technology, a farmer nowadays harvests not only crops but also growing amounts of data. These data are precise and small-scale – which is essentially why the combination of GPS, agriculture and data has been termed precision agriculture.

However, collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information about a certain asset – here: soil and yield properties – which should be used to the farmer's advantage. This is a common problem for which the term data mining has been coined. Data mining techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer.

A common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. In the past, this yield prediction usually relied on farmers' long-term experience for specific fields, crops and climate conditions. However, this knowledge might also be available, but hidden, in the small-scale, precise data which can nowadays be collected in-season using a multitude of sensors. These sensors essentially aim to measure a field's heterogeneity. Therefore, the problem of yield prediction encountered here is one of data mining and, specifically, multi-dimensional regression. This article concentrates on the capabilities of regression techniques used in agricultural yield data.

Furthermore, this article can be seen as a well-suited regression model, which is connected to database. This model can easily take the input process the data through data mining model which is developed using SQL for database design and Matlab for statistical

analysis. The researcher tried to find the best prediction model. To accomplish this, the model output on site-year data from different years and sites is compared. Results on the parameterization of the different models are presented.

Research Target. The overall research target is to find those indicators of a field's heterogeneity which are suited best to be used for a yield prediction task. The sub-task here is one of multidimensional regression – predicting yield from past and in-season attributes. Furthermore, from the agricultural perspective, it is interesting to see how much the factor influences the yield in the current site-year. For this purpose, modeling techniques can be used, but have to be evaluated first. Therefore, this work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. For this purpose, regression techniques will be evaluated on different data sets. Since models usually are strongly parameterized, an additional question is whether the model parameters can be carried over from one field to other fields which are comparable in (data set) size. This issue will also be addressed in this work. This is especially useful when new data have to be evaluated using one of the presented models.

Data mining and agriculture. In developed countries Data mining is widely applied to solve agricultural problems. e.g. the prediction of wine fermentation problems can be performed by using a k-means approach. Knowing in advance that the wine fermentation process could get stuck or be slow can help the enologist to correct it and ensure a good fermentation process. Weather forecasts can be improved using a k-nearest neighbor approach, where it is assumed that the climate during a certain year is similar to the one recorded in the past. The same data mining technique can also be used for estimating soil-water parameters.

Regression. Regression uses existing values to forecast what other values will be. It uses standard statistical techniques such as linear regression. Many real world problems like stock prices, sale volumes and product failure rates are very difficult to predict because they may depend on complex interactions of multiple predictive vectors. Neural network and decision trees are used to solve this type of complex real life problems.

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 personnel connect to 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 Description. The data required for this work has been collected for the years 2005–2012 on twelve talukas for four crops of Kolhapur district in Maharashtra, India. This data includes the year wise production, productivity index, rainfall, soil fertility (Electrical conductivity, PH, N, P, K) of respected year.

Sample Design. Researcher has collected data from directorate of economics & statistics of India. State Governments Statistical & agricultural department as well as soil department.

Generally the government employee called as talathis, is collecting the required data for the department. In each village he use to select plot and the respective crops randomly, means the department is collecting the required information for yield estimation from each and every village.

For this research study researcher has selected following crops in Kolhapur district in Maharashtra state in India. Researcher selected these crops because maximum of the farmers are cultivating these crops throughout the district as cash crops: rice, ground nut, soybean, ragi.

REVIEW OF LITERATURE

For examining the null hypothesis and its importance in crop, an extensive literature survey was conducted. Research papers in journals like American Journal of Agricultural Economics, Canadian Journal of Agricultural Economics, Agribusiness, and North Central Journal of Agricultural Economics were reviewed to analyze the findings of various researchers working in the general area of frequency distributions for historical crop yields.

In addition to the literature survey, a quantitative analysis of actual yields is considered necessary to test the null hypothesis. As part of this testing, data on historical a crop yields for sugarcane and soybeans were collected from India from Kolhapur districts for analysis. This data set used for testing the assumption of normality crops were selected for data collection and analysis because of their economic importance to Kolhapur district.

Day (1965) – He suggest that yield distribution in agricultural crop do not exhibit normality.[5]

Dorfan (1991) Argues that a large amount of agricultural economic data is inconsistent with the assumption of normality of Crop yield distribution.[6]

The importance of skew ness in crop yield forecasting is specified by Gallagher by Suggesting that ignoring skewed distribution will lead to underestimation of most likely yields.[7]

Just and Weinenger (1991) have studied country level data and based on their experiment the disagreed with crop yield distributions are non normal argued that the evidence available to date is not enough to disapprove normality of crop yields. [9]

Norwood B. Roberts, Lusk (2004) – By their studies it was observe that the semi parametric model ranked highest for forecasting purposes.[8]

Ramirez, Misra & Field (2003) – Studied yield distribution of these crops and conclude that they are non normal and left skewed. [10]

From the research article “Data mining of agricultural yield Data: A comparison of regression models” George RuB express that large amount of data which is collected and stored for analysis. Making appropriate use of these data often leads to considerable gains in efficiency and therefore economic advantage. This paper deals with appropriate regression techniques on selected agriculture data.[11]

“Classification of agricultural land soils: A data mining approach” In this research paper V. Ramesh and K. Ram explains comparison of different classifiers and the outcome of this research could improve the management and systems of soil uses throughout a large fields that include agriculture, horticulture, environmental and land use management.[12]

D. R. Mehata and others are worked on “Rainfall variability analysis and its impact on crop productivity” In this case study they collected the weekly rainfall data and number of rainy days recorded at the main Dry farming research station from 1958 to 1996 (39 yrs). The correlation and regression studies were worked out using rainfall(x) as independent variable and yield(y) as dependent variable to derive information on rainfall-yield relationship and to develop yield prediction model for important crops.

From “Generalized software tools for crop area estimation and yield forecast” Roberto Benedetti and others describes the procedure that leads to the estimates of the variables of interest, such as land use and crop yield and other sampling standard deviation, is rather tedious and complex, till to make necessary for statistian to have a stable and generalized computational system available. The SAS is also often ideal instrument to face with these needs, because it permits the handling of data effectively and provides all necessary functions to manage easily surveys with thousands of micro data. This paper focus on the use of this system in different steps of the survey: sample design, data editing and estimation. The information produced is however, available for one user only, the manager of the survey.[13]

“Risk in Agriculture: A study of crop yield distribution and crop insurance” by Narsi Reddy Gayam in his research study examines the assumption of normality of crop yields using data collected from INDIA involving sugarcane and Soybean. The null hypothesis (Crop yield are normally distributed) was tested using the Lilliefors method combined with intensive qualitative analysis of the data. Result show that in all cases considered in this thesis, crop yield are not normally distributed.[13]

RESEARCH DESIGN

It is necessary to study all aspects about crop for crop estimation. The following are the some of the study areas. The researcher concentrates on these areas to get direction for his research. Agriculture is very vast and versatile study area. So it is very complicated and difficult to find out proper variable and the methodology to suit these variables to get expected result. So researcher must have gone through vast literature survey, interviews and discussion with experts in agriculture field. Also researcher has study various ongoing and completed research work on national and international level.

- 1] Selection of crop for research.
 - Irrigated or rain fed crop.
 - Cash crop.
 - Productivity of crop.
- 2] Weather
 - Rainfall
 - Data available of actual rainfall.

- District wise rainfall, Taluka wise rainfall, village wise rainfall.
- 3] Requirement of rainfall- [stages]
 - Before sowing
 - After sowing
 - At cultivation
- 4] Soil fertility
 - Soil characteristic
 - Water holding capacity of soil.
 - talukawise samples of soil.
 - Electrical conductivity
 - Suitable condition for crop to specified study area.
- 5] External dosages to maintain PH.
 - Requirement of Nitrogen
 - Requirement of Potassium.
 - Area wise and location wise requirement of fertilizers.
 - Data available from farmers.

Crop. Irrigated crops e.g. sugarcane is not dependent on rainfall because essential water is provided through various water sources to these crops. So these crops were not considered for research. Researcher only concentrates on rain fed crops for his study.

Researcher considered only cash crops for his research. Researcher verifies the availability of data in the study area. i.e. cultivated crops. The impact of other variable and financial suitability of the selected crop is checked. The crops which have large stake in market are considered for research. Researcher also selected major cultivated crops in his study area where the past data is available.

Rainfall. Rainfall plays very important role in crop cultivation. Timely rainfall is very important for crop. But the accurate data of rainfall is not available. So researcher generates these data from interviewing the experts and farmers. He finalizes the data using various yearly rainy patterns.

There are stages in growing of crops. Sunlight is essential in growing period, the plants requires sunlight for fast development. For this reason data of rainfall in respected period is important for study.

Soil fertility. In respect of the fertility of soil, the following characteristics of soil are considered. i.e. water holding capacity, electrical conductivity, PH.

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 enormous data was collected for yield estimation. But, the entire data is not immediately relevant for model making. We have to clean the data and rearrange it until it is suitable for statistical method. Then different experimental models were designed. If the model does not interpret as expected, we rearrange the model structure or check the viability of the data for our model. This procedure is repeated until the final model designed. 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 did not match the actual crop yield. So, model was rearranged. Now this model give suitable output at taluka level. Again taluka wise data was collected and analysed.

The above models are used for analysis of yields of four crops viz. Paddy, Soybean, Groundnut and Ragi, because they are rain fed crops. There are twelve tehsils in Kolhapur district and four crops were selected from each tehsil. So 48 models were designed for the data analysis. Some crops were not cultivated in some talukas, their data was not available. e.g. in Shirol, Ragi is not cultivated, in Gaganbawada Soybean is not cultivated, etc.

Data analysis using Data mining Model. Collected data is analyzed using the statistical model. The result is the estimation of crop yield. This estimated yield is compared with the observed yield.

Researcher used Microsoft Excel for data analysis. It is time consuming to arrange the data on various different pages and implement all the formulas and check the result of this enormous data.

There are specific steps followed to carry out for the data analysis.

1. Collect the data.
2. Arrange the data according to subject.
3. Filter the data. So that unnecessary data should take out from the analysis.
4. Apply suitable methodology.
5. Modify and settle the model according to the result.

Initially above tasks were carried out using Excel. But for the enormous data, it is very difficult to analyze it using same method. It is difficult to arrange the data and memorize the various steps. So to overcome this difficulty researcher checked various options. And he came to a last approach which is describe below:

1. Consider the data as database.
2. Select the suitable RDBMS for the data.
3. Complete the table design for database.
4. Access the data in database using data mining tool [Matlab].

The average yield, standard deviation, partial and multiple correlation coefficients have been calculated and fitted in multiple regression planes. The principal component analysis for the components was analyzed. Statistical Method used for data analysis are Correlation, Multiple regression, principal component analysis.

Correlation. The correlation between two variables may be the following types:

1. Positive Correlation:- If an increase (or decrease) in the value of one variable is followed by increase (or decrease) in the value of other variable then we say that there is a positive correlation between two variables. Thus positively correlated variables change in the same direction. For example – income and expenditure, height and weight of group of people are positively correlated variables.
2. Negative Correlation: - If an increase (or decrease) in the value of one variable is accompanied by decrease (or increase) in the value of other variable then we say that, there is a negative correlation between two variables. That is negatively

correlated variables deviate in the opposite directions. For example, Supply and Price of the commodity.

3. No Correlation:- If the value of one variable increase or decrease, the value of the other variable remains constant on the average then we say that there is no correlation and the variables are said to be uncorrelated.

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. Productivity index is based on several variables such as agricultural yield, rainfall, soil fertility (includes PH, EC, N, P, K, water holding capacity of soil) etc. A variable whose numerical value is to be predicted is taken as dependent variable or response variable (productivity of crop) 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.

Principal Component Analysis. This application of PCA is simple: calculate the principal components and choose from them rather than the original data to construct the empirical model (regression, neural network, etc.). The (hoped for) advantage of doing this is that since PCA squeezes information into a subset of the new variables, less of them will be necessary to construct the model. In fact, it would not be unreasonable to simply step through the first so many principal components to build the model: First, use just the first principal component, then try the first and second, then the first, second and third, etc. A nice side benefit is that all the principal components are uncorrelated with each other.

Application and Interpretation. After finalizing the method for crop estimation the researcher concentrated upon the data analysis. So for this analysis researcher select the components (variables) from which he should finalize the correct module for crop yield estimation. For component finalization researcher adopted principal component analysis concept. From this he finalizes the main component for the model.

Following are the stages in component finalization. Researcher analyzes the data for selected crop using productivity index, rainfall and soil fertility index year wise.

Following are the results. The procedure for calculation of productivity index is not correct. i.e. Productivity Index X cultivated crop area = yield. There is no correlation seen between the components. r12 and r13 generated faulty results. Here regression must be in .5 to 1. But researcher not gets this range from the analysis. Only three year data is analyzed, which is not sufficient for the analysis. The data of soil fertility index is not suitable in the method.

rice5-6			
Yield (ton) PI X area	x1 (yield)	x2 (rainfall)	x3 (soil)
4369430	43.6943	1210	1.4
1539450	15.3945	974	1.53
29688750	296.8875	2472	1.57
22287600	222.876	3300	1.18
35289000	352.89	5075	1.76
3406000	34.06	6823	1.41
18147000	181.47	1558	1.82
13740300	137.403	1505	1.54
19609260	196.0926	1388	1.44
33437400	334.374	2531	1.55
26557128	265.5713	2987	1.35
39704310	397.0431	3279	1.31
AVG	206.4797	2758.5	1.488333
STDV	129.435	1734.923	0.18115
Corelation	r12 0.17361	r13 0.06356	r23 0.10186

-	r12	r12^2	r13	r13^2	r23	r23^2	r12r23	r13r23
rice 5-6	-0.124	0.01547	-0.271	0.07360	-0.101	0.01036	0.01266	0.02761

!R!	!R11!	!R12!	!r13!	!R22!	!R33!
0.893686	0.989637	0.152018	0.283964	0.926396	0.984525

In this stage researcher observed no effect on yield with respect to parameters like EC, N, P and K. This is because the value of these parameters for analysis is remaining same for different year. So the mean is constant. Variance, Standard deviation and correlation is zero, which is ineffective for regression analysis. There is no correlation between the components.

RICE 5-6							
Taluka	PR INDEX	RAINFALL	SOIL				
			PH	EC	N	P	K
			Neutral	Neutral	INDEX	INDEX	INDEX
Karveer	1725	1104.9	82.19993	85.66075	1.87	1.82	2.43
Radanagri	2614	3932	58.9025	95.16697	2.06	1.76	2.63
Kagal	1454	873.8	92.86523	95.55807	1.77	1.54	2.47
G. Bavda	1310	4626	32.05357	98.83929	2.05	1.41	2.55
Shahuwadi	1812	2398.3	66.42108	94.71778	1.91	1.18	2.4
Hatkangle	1769	651	90.60439	91.56712	1.86	1.4	2.45
Shirol	2799	469	87.84247	76.78082	1.88	1.53	2.57
Panhala	2925	1746	76.93556	97.97965	1.83	1.57	2.46
Gadhinglaj	2701	863	89.90585	97.7009	1.83	1.44	2.82
Bhudargad	2423	1753	60.42292	98.19974	2.04	1.55	2.52
Ajara	1989	2137	51.87808	99.50739	2	1.35	2.44
Chandgad	2763	2397.1	28.04106	99.36718	2.04	1.31	2.46

AVG	2190.33	1912.59	68.17271	94.25380	1.9283333	1.4883333	2.5166667
STDV	574.2558	1297.673	22.51343	6.777116	0.103294	0.18115	0.116333

Matrix of total Correlation coeff								
R	-	r1	r2	r3	r4	r5	r6	r7
	1	-0.16915	-0.16915	0.016704	-0.07018	0.079322	0.12785	0.401554
	r1	-0.16915	1	-0.80091	0.490997	0.764105	-0.06734	0.020541
	r2	-0.13801	-0.71758	1	-0.52493	-0.87206	0.316876	0.164138
	r3	0.042235	-0.33565	0.612463	1	0.347422	-0.35434	0.033255
	r4	0.079322	0.764105	-0.77777	-0.22106	1	-0.08664	-0.00126
	r5	0.12785	-0.06734	0.396372	0.098789	-0.08664	1	0.184346
	r6	0.401554	0.020541	-0.26478	-0.56047	-0.00126	0.184346	1

In this stage the same problems occurred in the last two stages. Considering these problem researcher again reformat procedure. Researcher collected five year data in place of 3 year data. Again he seen same problem of EC, N, P and K. He realized that the data which he collected from concern organization remain same for last ten years. It is not updated. So the result from the data is not suitable for analysis.

RICE 5-6							
Taluka	PR INDEX	RAINFALL	SOIL				
			PH	EC	N	P	K
			Neutral	Neutral	INDEX	INDEX	INDEX
Karveer	1725	1104.9	7458	7772	1.87	1.82	2.43
Radanagri	2614	3932	3510	5671	2.06	1.76	2.63
Kagal	1454	873.8	7380	7594	1.77	1.54	2.47
G. Bavda	1310	4626	718	2214	2.05	1.41	2.55
Shahuwadi	1812	2398.3	5319	7585	1.91	1.18	2.4
Hatkangle	1769	651	5082	5136	1.86	1.4	2.45

Shirol	2799	469	5130	4484	1.88	1.53	2.57
Panhala	2925	1746	5217	6644	1.83	1.57	2.46
Gadhinglaj	2701	863	4106	4462	1.83	1.44	2.82
Bhudargad	2423	1753	4229	6873	2.04	1.55	2.52
Ajara	1989	2137	3370	6464	2	1.35	2.44
Chandgad	2763	2397.1	1994	7066	2.04	1.31	2.46

Rice							
	x1	x2	x3	x4	x5	x6	x7
1440.8	3072.2	1037.952	5425	4772	1.87	1.82	2.43
2750	3637.2	3781.25	6510	5671	2.6	2.76	2.88
1390.6	2783.8	966.8842	5380	5594	1.47	1.54	2.12
1423	2690.8	1012.465	4980	2214	1.05	1.41	2.09
1296	2979.6	839.808	4819	5585	1.91	1.8	2.4
1172.7	2533.8	687.6126	4882	5136	1.86	1.4	2.45
1244	2931	773.768	4830	6484	2.88	1.93	2.57
1176	3674.4	691.488	5717	6644	2.83	2.57	2.46
1486	3390.2	1104.098	5906	6462	1.83	2.44	2.82
1957	3009	1914.925	6429	6873	2.04	2.55	2.52
1370	2469.6	938.45	6070	5464	2	1.75	2.44
1865	3681.6	1739.113	6994	7066	2.04	2.31	2.46

Matrix of total Corelation coeff						
r1	r2	r3	r4	r5	r6	r7
1	0.497317	0.58973	0.523525	0.50849	0.839349	0.543934

The mean is constant. Variance, Standard deviation and correlation is zero, which is ineffective for regression analysis. There is no correlation between the components. Up to this stage researcher only verifying the result by considering one model for the district. Researcher now changes his assessment towards the model. So he redesigned model by concentrating talukas in place of district.

	yid	Talid	pindex	rainfall	ph	ec	n	P	k
1989	1	Ajara	1280	1401	3841	3457	2.56	1.6	2.84
1852	2	Ajara	1281	1400	3840	3456	2.56	1.61	2.85
2029	3	Ajara	1279	1399	3839	3454	2.5596	1.59	2.83
2155	4	Ajara	1278	1398	3838	3455	2.5592	1.58	2.82
1815	5	Ajara	1281	1404	3843	3458	2.5616	1.62	2.84
2508	6	Ajara	1282	1402	3842	3457	2.5608	1.63	2.86

R	1	0.775731	0.835214	0.723077	0.783421	0.980469	0.960769
	0.775731	1	0.989743	0.901525	0.982213	0.841282	0.589188
	0.835214	0.989743	1	0.907841	0.961572	0.885714	0.680336
	0.723077	0.901525	0.907841	1	0.846094	0.762587	0.576461
	0.783421	0.982213	0.961572	0.846094	1	0.862949	0.587095
	0.980469	0.841282	0.885714	0.762587	0.862949	1	0.907115
	0.960769	0.589188	0.680336	0.576461	0.587095	0.907115	1
R	R 11	R 12	R 13	R 14	R 15	R 16	R 17
Value	2.47E-21	6.48E-21	2.91E-21	-3.6E-22	4.19E-21	2.71E-21	1.93E-21
S.D.	1.47196	2.160247	1.870829	1.47196	0.000867	0.018708	0.014142
R		-1.62262	-0.17977	1.144001	-0.69672	-0.09813	0.2175
	987.9005	1402	3842	3457	2.5608	1.63	2.86

Researcher carried out his study talukawise on selected crops. In all 48 models for twelve talukas and 4 crops in each talukas were developed. Researcher sees effective

results from this analysis. Now researcher concentrates upon fitting the model for collected data. Again he collects the effective data for data analysis. i.e. rainfall.

After Studying various aspects about the data its correlated effects on each other, researcher concentrates on certain components like soil fertility and rainfall. The data which he collected of soil is not suited for the analysis because one of the factor like neutral is one correct. So again he recollects the data and see that it is ok for the model. Now researcher concentrates on PH. There is another factor namune (samples) i.e. number of suitable samples which PH factor is in between the fertility index, which is in .5 to .7. Again researcher analyze the data in model and check the result.

year(rice)	Taluka	PR INDEX	RAINFALL	PH
2005-06	Ajara	1989	2137	208.156
2006-07	Ajara	1852	1966	358.048
2007-08	Ajara	2029	1844	333.112
2008-09	Ajara	2155	1268.6	561
2009-10	Ajara	1815	1370	643
2010-11	Ajara	2508	1405.3	357

x1=	2058	x2=	1665.15	x3=	410.0527
σ_1 =	230.57754	σ_2 =	330.9149	σ_3 =	146.7711
r12=	-0.400584	r23=	-0.82185	r13=	-0.19283

eq of x1 on x2 and x3 is $x_1 = 4350.92 - 1.2x_2 - 0.67x_3$
 The estimated yield is $x_1 = 1475.945$ when $x_2 = 1500$, $x_3 = 250$

Taluka	PR INDEX	RAINFALL	PH
Ajara	1989	2137	265
Ajara	1852	1966	309
Ajara	2029	1844	465
Ajara	2155	1268.6	561
Ajara	1815	1370	643
Ajara	2508	1405.3	357

x1		x2		x3	
1989		2137		208.156	
1852		1966		358.048	
2029		1844		333.112	
2155		1268.6		561	
1815		1370		643	
2508		1405.3		357	
x1=	2058	x2=	1665.15	x3=	410.0527
σ 1=	230.5775358	σ 2=	330.9149	σ 3=	146.7711
r12=	-0.400583895	r23=	-0.82185	r13=	-0.19283

eq of x1 on x2 and x3 is $x_1 = 5022.85 - 1.85x_2 - 2.48x_3$
 The estimated yield is $x_1 = 3621.42$ when $x_2 = 287$, $x_3 = 351$

year(rice)	Taluka	PR INDEX	RAINFALL	PH
2005-06	Ajara	1989	1599	265
2006-07	Ajara	1852	1966	309
2007-08	Ajara	2029	3093	465
2008-09	Ajara	2155	1922	561
2009-10	Ajara	1815	1370	643
2010-11	Ajara	2508	1405	357

x1	x2	x3
1989	1599	265
1852	1966	309
2029	3093	465
2155	1922	561
1815	1370	643
2508	1405	357

x1=	2058	x2=	1892.5	x3=	433.33333
σ_1 =	230.5775358	σ_2 =	583.70619	σ_3 =	135.95383
r12=	5.871435815	r23=	2.85989443	r13=	1.2415199

eq of x1 on x2 and x3 is $x1 = 2306.38 - 0.0567x2 - 0.3252x3$

The estimated yield is $x1 = 2123.8546$

After completing these stages researcher process the data in data mining model which he constructed using Matlab.

MATLAB is fundamentally a matrix programming language. A matrix is a two-dimensional array of real or complex numbers. Linear algebra defines many matrix operations that are directly supported by MATLAB. Linear algebra includes matrix arithmetic, linear equations, eigen values, singular values, and matrix factorizations. MATLAB provides many functions for performing mathematical operations and analyzing data. Creation and modification of matrices and vectors are straightforward tasks.

MATLAB has several auxiliary libraries called Toolbox, which are collections of m-files. That have been developed for particular applications. These include the Statistics toolbox, the Optimization toolbox, and the Financial toolbox among others. In our particular case, we will do a brief description of the Statistics a Optimization toolboxes and we will use same commands of the Financial toolbox.

Statistics toolbox (Applications). The Statistics toolbox, created in the version 5.3 and continuously updated in newer versions, is a collection of statistical tools built on the MATLAB numeric computing environment. The toolbox supports a wide range of common statistical tasks, from random number generation. The Statistics toolbox provides functions for describing the features of a data sample. These descriptive statistics include measures of location and spread, percentile estimates and functions for dealing with data having missing values. The following table shows the most important ones with a brief description of their use.

MATLAB is a high-level language that includes matrix-based data structures, its own internal data types, an extensive catalog of functions, an environment in which to develop our own functions and scripts.

Function Name	Description
Corr	Linear or Rank Correlation Coefficient
Corrcoef	Linear Correlation Coefficient with Confidence intervals
Cov	Covariance Matrix
Geomean	Geometric Mean
Grpstats	Summary Statistics by Group
Kurtosis	Kurtosis
Mad	Median Absolute Deviation
Median	50th Percentile of a Sample
Moment	Moments of a Sample
Nanstat	Descriptive Stastiscs ignoring NaNs (missing data)*
Prctile	Percentiles
Quantile	Quantiles
Range	Range
Skewness	Skewnes

MATLAB CODE FOR REGRESSION ANALYSIS

MATLAB provides a full programming language that enables us to write a series of MATLAB statements into a file and then execute them with a single command. We write our program in an ordinary text file, giving the file a name of file name .m. The term we use for filename becomes the new command that MATLAB associates with the program. The file extension of .m makes this a MATLAB m - file. For example, when we write a program in MATLAB, we save it to an m-file. There are two types of m-files: the scripts that simply execute a sequence of MATLAB statements, and the functions that also accept input arguments and produce outputs

1] CODE

```
conn = database('yield','sa','jadhav');
exec(conn,'use yield');
setdbprefs('DataReturnFormat','structure');
curs = exec(conn, 'SELECT productivity.pi AS X1, Rainfallph.fall AS X2, Rainfallph.ph AS X3 FROM
crop INNER JOIN productivity ON crop.cid = productivity.cid INNER JOIN Rainfallph ON
productivity.tid = Rainfallph.tid AND productivity.year = Rainfallph.year INNER JOIN taluka ON
Rainfallph.tid = taluka.tid where taluka.tid = 1 and crop.cid = 1');
a = fetch(curs);
X = [ a.Data.X1'
a.Data.X2'
a.Data.X3'];
X = X';
M=mean(X);
C=corrcoef(X);
a = size(X);
n=a(1,1);
s1 = std(X(:,1),1);
s2 = std(X(:,2),1);
s3 = std(X(:,3),1);
r12=C(1,2);
r13=C(1,3);
r23=C(2,3);
S=std(X,1);
R=(1-r23*r23)-r12*(r12-r23*r13)+r13*(r12*r23-r13);
R11=1-r23*r23;
R12=(-1)^(1+2)*(r12-r23*r13);
R13=(-1)^(1+3)*(r12*r23-r13);
R21=(-1)^(2+1)*(r12-r23*r13);
R22=1-r13*r13;
R23=(-1)^(2+3)*(r23-r13*r12);
R31=1-r23*r23;
R32=(-1)^(3+2)*(r23-r13*r23);
R33=1-r12*r12;
A=M(1,1)+R12*M(1,2)*s1/(s2*R11)+R13*M(1,3)*s1/(s3*R11);
B1=-R12*s1/(s2*R11);
B2=-R13*s1/(R11*s3);
R1_23=sqrt(1-R/R11);
R2_13=sqrt(1-R/R22);
R3_12=sqrt(1-R/R33);
r12_3=-R12/sqrt(R11*R22);
r13_2=-R13/sqrt(R11*R33);
r23_1=-R23/sqrt(R33*R22);
```

RECOMMENDATIONS

This integrated Crop model had four main uses: as a national, regional and local level conservation policy studies; in programme planning and evaluation; in project planning and design; as a research and teaching tool.

Importance:

- Stabilizing the price food in the market and continued supply of food grains in the rural markets are the best options for increasing household food security.
- The second priority option for mitigation household food insecurity was to improving PFDS by making it efficient and effective.
- Finally suggested to improving rural communication for better transportation of goods and that would certainly benefit the locality as well as community.

Further improvements will make the model a more useful tool for the agriculture sector study region. The model could be used to: estimate rain fed yield; investigate the reasons for low yields; assist as a tool to complement the soil and water management research; train agriculture extension personnel in the growth and management of the crop.

This study will help identify the types of public policies needed for the development of competitive and efficient agricultural markets that can contribute in reducing rural poverty and promoting agricultural economic growth in the country.

REFERENCES

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 nonsymetric disturbance" American Journal of Agriculture Economics 69. (Nov. 1987) : 796 .803.
8. Norwood B. Roberts, 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"
11. American Journal of Agriculture Economics 2003. Volume 85: 108
12. Georg Ruß Data Mining of Agricultural Yield Data: A Comparison of Regression Models, ICDM'09,. Leipzig, Germany, July 2009
13. 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
14. Rainfall variability analysis and its impact on crop productivity Indian
15. 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
16. "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).
 17. Gazetteer of Kolhapur District (2001)
 18. Aditya, Kaustav (2008). Forecasting of crop yield using discriminant function technique. M.Sc. thesis, PG School, IARI, New Delhi.
 19. Agrawal, Ranjana, Jain, R.C. and Singh, D.(1980). Forecasting of rice yield using climatic variables. Ind. J. Agric. Sci., 50 (9), 680-684.
 20. Agrawal, Ranjana and Jain, R.C. (1982). Composite model for forecasting rice yield. Ind. J. Agric.Sci., 52 (3), 189-194.
 21. Agrawal, Ranjana, Jain, R.C. and Jha, M.P. (1983). Joint effects of weather variables on rice yields. Mausam, 34 (2), 177-181.
 22. Agrawal, Ranjana, Jain, R.C. and Jha, M.P. (1986). Models for studying rice crop weather relationship. Mausam, 37 (1), 67-70.