IMPACT OF RAINFALL ON COTTON CROP YIELD IN SABARKANTHA DISTRICT USING DATA MINING TECHNIQUES

***PATEL, AMIKSHA A.[1] AND KATHIRIYA, DHAVAL R.[2]**

**CENTER FOR AGRICULTURAL INFORMATION
AND COMMUNICATION TECHNOLOGY
SARDARKRUSHINAGAR DANTIWADA AGRICULTURAL UNIVERSITY
SARDARKRUSHINAGAR – 385 506, GUJARAT, INDIA**

***E-MAIL: amiksha_patel@yahoo.com; dit@aau.in**

1.Assistant Professor (Computer Science), Centre for Agricultural Information & Computer Technology, Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar – 385506 Dist. Banaskantha, Gujarat.
2.Director, Information Technology, Anand Agricultural University, Anand, Gujarat

ABSTRACT

In the present study, the methodology research and a case study using both qualitative and quantitative methods was employed for the analysis of rainfall and cotton yield data in an agricultural context. In Gujarat, Sabarkantha region is predominantly growing the cotton as rainfed crop. The year to year fluctuation in the crop yields are mainly attributed to the variation in rainfall and its distribution. In order to study interrelationships between rainfall distribution and cotton yield in Sabarkantha district of Gujarat state, correlation and regression analysis techniques were employed. The district wise average yield data of cotton and daily rainfall data were used. Secondary annual rainfall profiles for a selected study area within the Sabarkantha Agricultural region of Gujarat state were used to identify areas of high crop production.

KEY WORDS: Data mining, Correlation Co-efficient, Cotton, Rainfall

INTRODUCTION

The aim of this work was to find a relationship between rainfall and crop production and in order to justify possibly for predicting the crop production yield at certain places within the agricultural region with a rainfall data. Cotton occupies a very large area under rainfed condition. By years variation in the crop yields are mainly depending on the variation in rainfall and its distribution. In all purpose, it could be inferred that the quantum of rainfall during different stage of the cotton had significant influence on cotton productivity.

Cotton is one of the most important fiber and cash crop of India and plays a dominant role in the industrial and agricultural economy. Gujarat is the second largest cotton producing state of India. Bharuch, Sabarkantha, Surendemagar, Vadodra and Ahmedabad districts are the major cotton producing districts in Gujarat (Anonymous, 2015). The average yield is 1.8 quintals / hectare, which is almost the same as the national average. With 80-100 cm annual rainfall, Gujarat provides favourable conditions for cotton cultivation. The year to year fluctuation in the crop yields are mainly attributed due to the variation in rainfall and its distribution.

The work related to present study was carried out by several scientists in the past. Boken (2000) forecasted the spring wheat yield using time series analysis for

Saskatchewan, Canada, using yield data for three years. The effects of rainfall pattern on the productivity of groundnut were studied by several scientists (Suryanarayana *et al*., 1982 and Sahu *et al*., 2004). Padhan (2012) studied the application of ARIMA Model for forecasting agricultural productivity in India using more than 50 years data. Patel and Vaishnav (2003) studied the effect of rainfall on groundnut yield under dry framing situation of Gujarat. Khatri and Patel (1990) tried to locate critical phases in groundnut crop by selecting rainfall variables through stepwise regression analysis technique. The effects of rainfall distribution on the yield of groundnut during its growth period were studied by Singh and Singh (1994) in Rajkot district of Gujarat. Suresh and Krishna Priya (2011) tried to forecast the sugarcane yield of Tamil Nadu using ARIMA models using data of 62 years. Forecasting of groundnut yield using rainfall variables for Saurashtra region of Gujarat was carried out by Parmar *et al*. (2004).

RESEARCH METHODOLOGY

Sabarkantha district is selected for the present study. The data of average yield over a period of 18 years *i.e*. from 1998-2015 and the corresponding daily rainfall data were collected from the Department of Meteorology, S. D. Agricultural University, Dantiwada. Two broad approaches and correlation-regression analysis techniques were applied to study interrelationships between rainfall distribution and cotton yield. In case of gross rainfall approach, the gross rainfall received during the crop period meteorological standard weeks (MSW) was considered. For monthly rainfall approach, monthly total rainfall received during the months of June, July, August and September were worked out for each year. Usually monsoon gets withdrawn by the end of September and hence, October was not considered. Further, in data mining activities to use the training set to determine the best-fit algorithm using a simple model of crop yield as a function of the average annual rainfall. All of the classification algorithms within WEKA were tested in this research and a short-list of five algorithms Gaussian Processes (GP), Multilayer Perceptron (MLP), Kstar, Sequential Minimal Optimisation (SMO) and Additive Regression (AR) are selected.

RESEARCH INVESTIGATION

The findings of the present study as well as relevant discussion have been presented under following heads:

Gross rainfall approach

The outcome presented in Table 1 showed that, the correlation co-efficient (r) between cotton yield and rainfall founds a positive and significant. Thus, the results showed that cotton productivity is directly connected with the gross rainfall received during its growing season in Sabarkantha district. The regression co-efficient along with the consistent co-efficient of determination is presented in Table 2. The regression co-efficient corresponding to gross rainfall is positive and significant.

Monthly rainfall approach

The correlation co-efficient (r) between cotton yield and monthly total rainfall is presented in Table 3. The results indicated that correlation co-efficient (r) was positive and significant to the cotton yield in July month for Sabarkantha district.

The multiple regression equation was fitted for Sabarkantha district is presented in Table 4. The result revealed that partial regression co-efficient is positive and significant in July month. The R^2 value is 0.39. Thus, the results indicated that splitting of total rainfall of the season into four monthly variables (Table 4) improved R^2 values when compared with R^2 values presented in Table 2.

Data mining analysis of rainfall and cotton crop yield

The next step in the individual scrutiny of the application is the use of regression in order to determine if the relationship established through correlation could be supported by a mechanism of predicting the cotton crop yield through the rainfall. This was carried out using the classification technique of data mining (DM) in the Waikato Environment for Knowledge Analysis (WEKA) software. The gross data for average annual rainfall and cotton crop yield for the Sabarkantha district were used for this activity. The gross cotton crop yield and rainfall data set was split up into a training set (2001, 2002, 2004 and 2006 data) and a test set (2003 and 2005 data). The exploratory part of the data mining activity was to use the training set to determine the best-fit algorithm using a simple model of crop yield as a function of the average annual rainfall. All of the classification algorithms within WEKA were tested in this step and a short-list of five algorithms was selected. These algorithms were Gaussian Processes (GP), Multilayer Perceptron (MLP), Kstar, Sequential Minimal Optimisation (SMO) and Additive Regression (AR). All these algorithms use regression for predicting continuous values in response to input values. Gaussian Processes (GP) is a form of regression where the distribution is over mean and covariance functions without hyper parameter tuning for the classifier function (Rasmussen, 2004); Multilayer Perceptron (MLP) is a feed forward multi-layer artificial neural network(ANN) function approximate classifier that uses the supervised learning technique of back propagation to classify instances (Nazzal et al., 2008); Sequential Minimal Optimisation (SMO) uses the support vector machine for its regression by quadratically scaling the number of training patterns (Cao *et al.*, 2006); the lazy Kstar algorithm is an instance based classifier that classifies a test instance based on its similarity to the training instance; and Additive Regression (AR) is a meta classifier that seeks to enhance the performance of the regression based classifier (Witten *et al.*, 2011). Each of the algorithms trialled had different characteristics, correlation co-efficient and Root Mean Square Errors (RMSE) as shown in Table 5. Good predictions were considered to have a percentage error of less than 20 per cent, average predictions a percentage error of 21-40 per cent and weak predictions a percentage error of over 40 per cent. SMO and AR were ruled out in the first instance due to the high RMSEs for the cross validation results. The GP algorithm had the lowest RMSE for the predictions on the test data and a correlation co-efficient of 0.99. Based on these results, together with a good cross validation result, the GP algorithm was selected and run for the prediction phase of the data mining activity. The results indicated that the link of rainfall and the cotton yield predictions fell down respectively as the rainfall heavily increased.

CONCLUSION

The comparative study of multiple prediction models for relation showing of rainfall on cotton crop yield using a large dataset along with a cross-validation provided us with an insight into the relative prediction ability of different data mining methods. Using sensitivity analysis on covariance functions without hyper parameter tuning for the classifier function models provided us with the prioritized importance of the predictive factors used in the study.

The result also showed that, the correlation co-efficient (r) between cotton yield and rainfall founds an optimistic and significant. Thus, the results showed that cotton productivity is directly associated with the gross rainfall received during its growing season in Sabarkantha district. Thus, it is clear from the discussion that the rainfall for the cotton crop had considerable impact on cotton productivity, however, more amount of rainfall or its distribution alone could not determine the cotton yield. As a consequence, it was concluded that rainfall may, therefore, not be such a decisive factor for cotton yield.

The use of the data mining classification function of Gaussian Processes (GP) showed that the correlation between the random average annual rainfall and cotton yield was strongly positive one and that as a result generally, cotton yield in the Sabarkantha agricultural region can be expected to increase with an consistency in rainfall, but there could be an increasing under-estimation error in predicting the cotton yields. The uncertainty of the prediction was thought to be related to the influence of other factors such as the total seasonal rainfall as opposed to the rainfall for the months separately, as well as to the sparseness related to yield measurement of the data set. It is believed that the performance of the WEKA algorithms could be enhanced by increasing the sample size of the crop yields from 3 year selection to the 10 year selection for the years 2001-2010. Further investigation may also be needed in order to investigate the other elements of climate such as temperature as well as soil moisture index which integrates the effect of temperature, soil physical parameters along with the rainfall distribution could determined the cotton crop productivity to a large extent.

REFERENCES

Anonymous (2015). District wise area, production and yield per hectare of important food and non-food crops in Gujarat, Directorate of Agriculture, Gujarat, Krishi Bhavan, Gandhinagar, Gujarat, India.

Boken, V. K. (2000). Forecasting spring wheat yield using time series analysis: A case study for the Canadian prairies. *Agron. J.*, **92**(6): 1047-1053.

Cao, L. J.; Ong, C. J.; Zhang, J. Q.; Periyathamby, U.; Fu, X. J. And Lee, H. P. (2006). Parallel sequential minimal optimization for the training of support vector machines. *IEEE Transactions on Neural Networks*, **17**(4): 1039-1049.

Khatri, T. J. and Patel, R. M. (1990). Regression analysis for locating critical stages in groundnut crop. *J. Indian Soc. Agric. Stat.*, **42**(3): 334.

Nazzal, J. M.; El-emary, I. M.; Najim, S. A. and Ahliyya, A. (2008). Multilayer Perceptron Neural Network (MLPs) for analyzing the properties of Jordan oil shale. *World Appl. Sci. J.*, 5(5) 546-552.

Padhan, P. C. (2012). Application of ARIMA model for forecasting agricultural productivity in India. *J. Agric. Social Sci.*, **8**: 50-56..

Parmar, B. A.; Sahu, D. D.; Dixit, S. K. and Patoliya, B. M. (2004). Forecasting of groundnut yield using rainfall variables for Saurashtra region of Gujarat state. *J. Agrometeorol.*, **6**(1): 1-8.

Patel, J.S. and Vaishnav, M. R. (2003). Evaluation of different approaches to study the effect of rainfall on groundnut in dry farming area of Gujarat. *J. Agrometeorl.*, 5(1): 76-83.

Rasmussen, C. E. (2004). *Gaussian Processes in Machine Learning. In: Advanced lecture on Machine Learning.* (Eds.). Bousquet, O.; Luxburg, U. And Ratsch, G. *Springer, New York.* pp. 63-71.

Sahu, D. D.; Golakiya, B. A. and Patoliya, B. M. (2004). Impact of rainfall on the yield of rainfed groundnut. *J. Agrometeorl.*, **6**(2): 249-253.

Singh, J. and Singh, B. H. (1994). Effect of rainfall distribution on the yield of groundnut during its growth period. *J. Indian Soc. Ag. Stat.*, **46**(1): 99.

Suresh, K. K. and Krishna Priya, S. R. (2011). Forecasting sugarcane yield of Tamil Nadu using ARIMA models. *Sugar Tech.*, **13**(1): 23-26

Suryanarayana, G.; Rajashekara, B. G.; Jagannatha, M. K. and Venkataramu, M. M. (1982). Effect of rainfall and its pattern on the yield of groundnut. *Mysore. J. Agric. Sci.*, **16**(2): 128-133.

Witten, I. H.; Eibe, F. And mark, A. H. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 6th edition. Burlington, M. A. : Morgan Kaufman, 2011.

Table 1: Correlation co-efficient between cotton yield and gross rainfall

District	Correlation Co-efficient (r)
Sabarkantha	0.54*

***** Indicate significance of value at P=0.05***

Table 2: Models of cotton yield using gross rainfall approach

Variables	Sabarkantha
Constant	53.00
Regression Co-efficient.	
Gross rainfall	1.15*
R^2	0.36

***** Indicate significance of value at P=0.05***

Table 3: Correlation co-efficient between cotton yield and monthly total rainfall

District	Correlation co-efficient (r)			
	June	July	August	September
Sabarkantha	0.34	0.39*	0.34	0.23

***** Indicate significance of value at P=0.05***

Table 4: Models for cotton yield using monthly total rainfall

Variables	Sabarkantha
Constant	12.97
June	1.42
Regression Co-efficient.	
July	1.82*
August	1.17
September	0.46
R^2	0.39

***** Indicate significance of value at P=0.05***

Table 5: WEKA algorithms results from training data set

WEKA Algorithm	Correlation Co-efficient Training	RMSE Training Set	RMSE Cross Valid	RMSE Test Set
Gaussian Processes	0.987	3.539	4.912	3.217
MLP	0.993	0.654	4.685	5.856
SMOreg	0.987	1.071	5.213	6.699
Kstar	1.000	0.017	4.855	5.367
Additive Regression	0.998	0.117	6.478	3.646

[MS received : December 19, 2017] *[MS accepted : December 25, 2017]*