ISSN No. 0976-5697
Volume 6, No. 3, May 2015 (Special Issue)
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Prediction of Paddy Production based on Rainfall and Ground Water Level

Surbhi Sarda,
IV/IV B. Tech Student of Department of Information Technology
PVP Siddhartha Institute of Technology
Vijayawada, India
sardasurbhi@gmail.com

N.Deepu
IV/IV B. Tech Student of Department of Information Technology
PVP Siddhartha Institute of Technology
Vijayawada, India

P.Ajay Kumar
IV/IV B. Tech Student of Department of Information Technology
PVP Siddhartha Institute of Technology
Vijayawada, India

S.Venu Babu
IV/IV B. Tech Student of Department of Information Technology
PVP Siddhartha Institute of Technology
Vijayawada, India

Dr. J. Rajendra Prasad
HOD & Professor
Department of Information Technology,
PVP Siddhartha Institute of Technology,
Vijayawada, India,
rp.rajendra@rediffmail.com

D. Ratnam
Assistant Professor
Department of Information Technology
PVP Siddhartha Institute of Technology
Vijayawada, India,
ratnam.dodda@gmail.com

Abstract: Paddy is the staple crop of Andhra Pradesh due to which forecasting its future productions is one the most necessary application among the current issues. Ground water levels are one of the major water resources to the Krishna district, Andhra Pradesh where most of the cultivation of the crops such as paddy, ground nuts and many more rely on the ground water. The ground water fluctuations also make the prediction of paddy production quite cumbersome. Thus to determine the paddy production and to analyze the fluctuations of the ground water level as it affects the paddy production, we make use of the classification algorithm i.e. linear regression. In the current scenario the huge amount of fluctuations in the ground water level is affecting the economy at a higher level. It affects the paddy productions, economy, and the low lying people. The future predictions of the ground water level will be of great use for advancement in the growth of the crops and increase in the paddy production. Finally the main motive of our model is to predict the future paddy productions based on the ground water level and the rainfall data.

Keywords: ground water, rainfall, production, fluctuations, linear regression.

I. INTRODUCTION

In a state like Andhra Pradesh ground water levels are of prominent importance. Most of the crops grown in the state rely on the ground water as their primary source. Recently, the economic development, urbanization process and over-exploitation, etc. have led to severely decreasing groundwater resources in some areas. Therefore, prediction of groundwater level is an important requirement for groundwater planning and use in all areas. In most cases the surface water is not easily accessible, in such cases ground water is the best alternative. In some of the metropolitan cities ground water is even considered as the major source of drinking water [2]. Monitoring and management of groundwater resources is achieved by monitoring the ground water levels in the well through measurement and using computer models to forecast the water levels in the wells [5]. Our study is to determine the ground water level in the Krishna district of Andhra Pradesh using Simple Linear Regression. Linear Regression has great ability to make predictions from your data; either future predictions or indications of past behavior. The input variables selected for the model are the past five year's data on rainfall and the ground water table level.

II. DATA ANALYSIS

The simple linear regression is used in order to determine the forecasting of the future paddy yield using ground water level and rainfall. In this procedure using simple linear regression as our primary algorithm we compute the Pearson's correlation coefficient [4]. The correlation coefficient gives us the underlying meaning between the two factors production and the ground water levels. It defines the extent of dependency between the two factors. The coefficient shows how close two variables lie along a line. If the coefficient is equal to 1 or -1, all the points lie along a line. If the correlation coefficient is equal to zero, there is no linear relation between x and y. However, this does not necessarily mean that there is no relation at all between the two variables. There could e.g. be a non-linear relation. A positive relationship means that the two variables move into the same direction. A higher value of x corresponds to higher values of y, and vice versa. A negative relationship means that the two variables move into the opposite directions. A lower value of x corresponds to higher values of y, and vice versa.

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

III. REGRESSION

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. There are two types of regressions on the list such as simple linear regression and multiple linear regressions. However we only concentrate on the simple linear regression as we need find the dependency between only two variables [6].

Linear regression:

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the parameters (but need not be linear in the independent variables). For example, in simple linear regression for modeling n data points there is one independent variable: x_i, and two parameters, β_0 and β_1:

Straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$

Ei is an error term and the subscript i indexes a particular observation.

Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The residual, $e_i = y_i - \widehat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\widehat{y}_i$, and the true value of the dependent variable, y_i. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSE [7], also sometimes denoted RSS:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimator $\widehat{\beta}_0, \widehat{\beta}_1$.

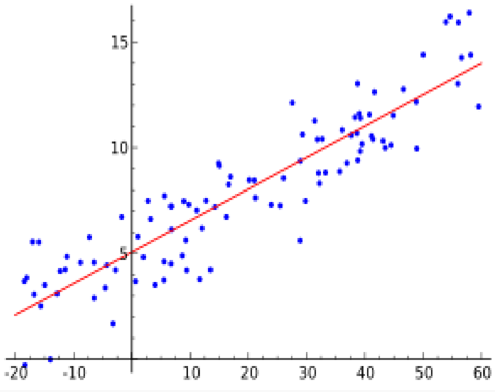

Illustration of linear regression on a data set.

In the case of simple regression, the formulas for the least squares estimates are

$$\widehat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where $\bar{x}$ is the mean (average) of the x values and $\bar{y}$ is the mean of the y values.

Under the assumption that the population error term has a constant variance, the estimate of t

hat variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{n-2}.$$

This is called the mean square error (MSE) [2] of the regression. The denominator is the sample size reduced by the number of model parameters estimated from the same data, (*n-p*) for *p* regressors or (*n-p*-1) if an intercept is used. In this case, *p*=1 so the denominator is *n*-2.

The standard errors of the parameter estimates are given by

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}.$$

III.WEKA TOOL

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. In this model we make use of weka tool for forecasting the future water table levels. The computed datasets are justified by comparing them with the manually calculated values.

In the weka tool we use various parameters for justifying the dependency between the rainfall and ground water table which we have taken as our input variables. They are:

i.CORRELATION COEFFICIENT:

A correlation coefficient shows the degree of linear dependence of x and y. In other words, the coefficient shows how close two variables lie along a line. If the coefficient is equal to 1 or -1, all the points lie along a line. If the correlation coefficient is equal to zero, there is no linear relation between x and y

ii.MEAN ABSOLUTE ERROR:

The **mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n} |e_i|.$$

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

As the name suggests, the mean absolute error is an average of the absolute errors $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors.

The mean absolute error is a common measure of forecast error in time series analysis, where the terms "mean absolute deviation" is sometimes used in confusion with the more standard definition of mean absolute deviation. The same confusion exists more generally. For each instance in the test set, Weka obtains a distribution (for each class label a value from 0 to 1, i.e., 0-100%). This distribution is matched against the expected distribution (the expected class label has 1 in that array, the others 0). For each class label the following is calculated:
AbsErrPerLabel = abs (actual - predicted)/# of class labels.
The absolute error per Instance is than the sum of these:
AbsErrPerInstance = Sum (AbsPerLabel)
(Note: The instance weight is taken into account as well. But this is normally just 1.)
The mean absolute error is the sum over all the instances and their AbsErrPerInstance divided by the number of instances in the test set with an actual class label (that should normally be all of them).
MeanAbsErr = Sum (AbsErrPerInstance) / # inst. with class label.

iii.ROOT MEAN SQUARED ERROR:

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between value (Sample values) predicted by a model or an estimator and the values actually observed. In weka the root mean squared errors is computed by squaring the difference to the actual value and the predicting value for all class labels and divide it by the number of class labels, these are summed (= "SqrErr") for each instance; this is summed over all instances (= "SumSqrErr")
For example consider an instance with five instances, then
Class1

predicted	actual	diff^2/2
0.66100	1.00000	0.05746
0.98500	1.00000	0.00011
0.98500	1.00000	0.00011
0.98500	1.00000	0.00011
0.98500	1.00000	0.00011

SumSqrErr=0.11582
our RMSE = 0.152197897[=sqrt (SumSqrErr/5)]
Weka output= 0.152

iv.RELATIVE ABSOLUTE ERROR:

The **relative absolute error** is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error. Thus, the relative absolute error takes the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|\bar{a} - a_1| + \ldots + |\bar{a} - a_n|}$$

v.ROOT RELATIVE ABSOLUTE ERROR:

RRSE is computed by dividing the RMSE by the RMSE obtained by just predicting the mean of target values (and then multiplying by 100). Therefore, smaller values are better and values > 100% indicate a scheme is doing worse than just predicting the mean. RAE is computed in a similar manner.

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(\bar{a} - a_1)^2 + \ldots + (\bar{a} - a_n)^2}$$

The following screen shots are the sequence in the order we have obtained the output:

1. Initially we import the data file to the weka tool containing the data regarding rainfall and the ground water table level. Select the “open file” option to import.

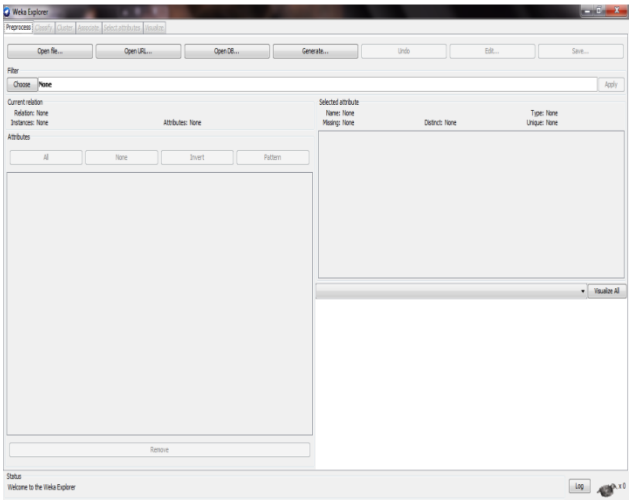

Figure 1. Initial page

2. The file is imported from the appropriate folder.

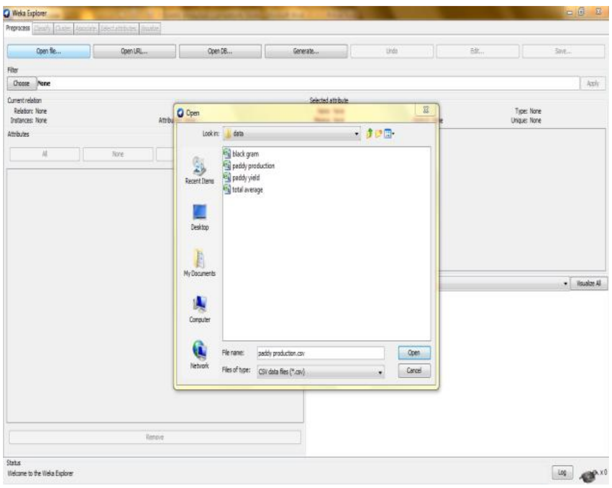

Figure 2. importing file

3. Selection of the data file

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

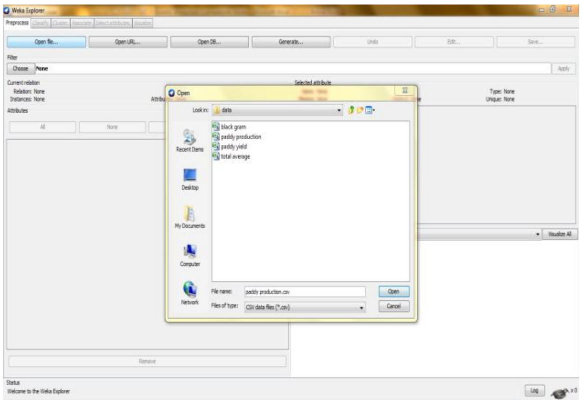

Figure 3. Selection of the file

4. All the labels of the imported file are selected.

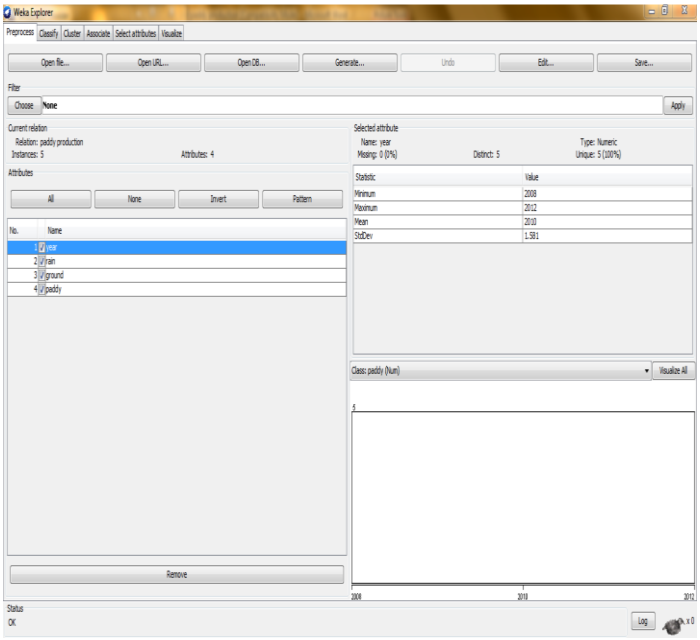

Figure 4. Selection of the all the labels

5. We make use of classification technique in order to obtain the appropriate results for determining the future ground water levels

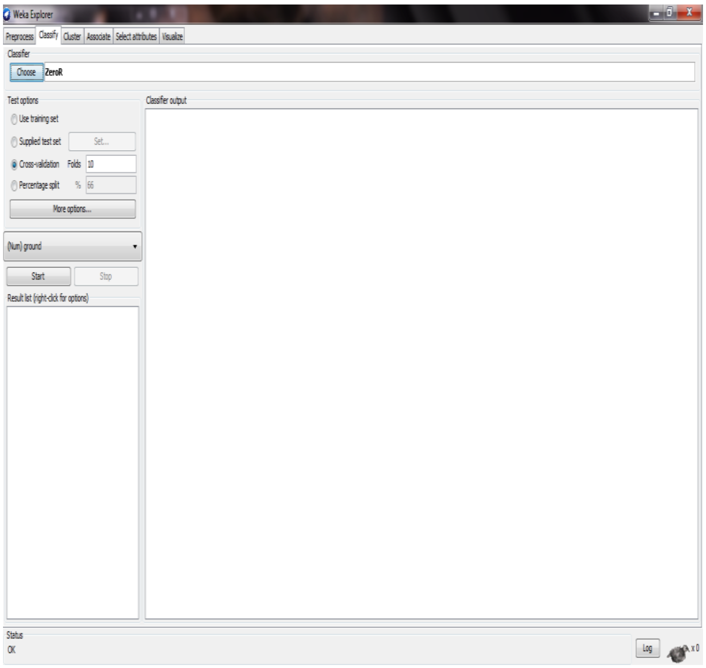

Figure 5. Choose classification technique

6. The simple linear regression algorithm is used to give the dependency between the rainfall and ground water level. In the testing options we select the use training set option. The random seed for XVAL/ % spilt. In our model we consider two splits.

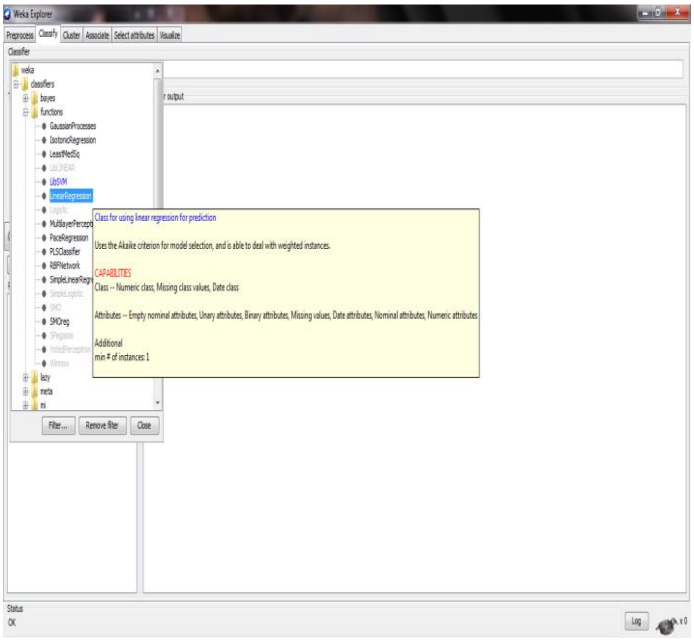

Figure 6. Use of linear regression

7. Final output of the model

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

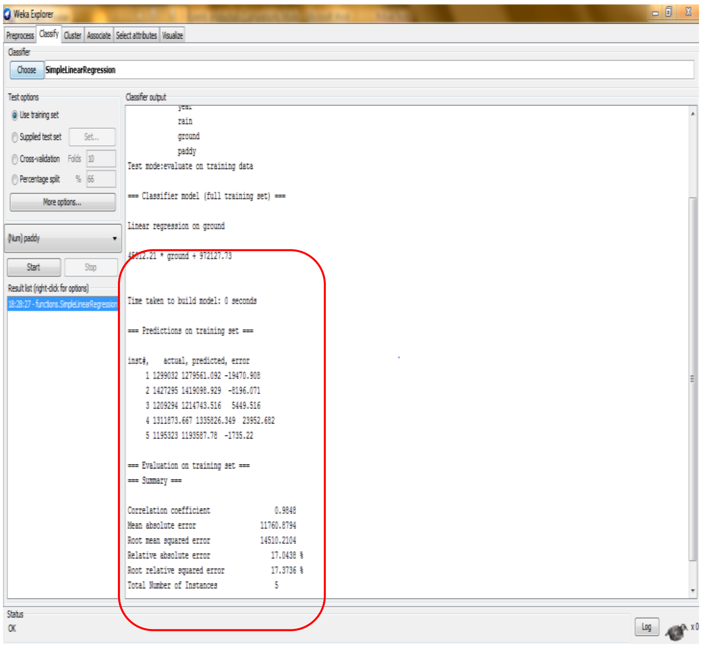

Fig 7: Output

From the analysis we may justify that there is a close relationship between the ground water level and the paddy productions. When there is increase in the ground water there is simultaneous increase in the paddy productions and vice versa.

IV. CONCLUSION

Our system is mainly focusing on the prediction of the paddy productions based on the ground water level and of those crops which rely on ground water as the main source of water. The prediction is done based on the dependency of the rainfall and the ground water level data. The regression method is very useful for predictive analysis of the data, among which simple linear regression is used to compute the dependency of the rainfall and the ground water level data, future predictions of the paddy productions and also fluctuations in the ground water level. This model will be of great use to the farmers, economists in various ways.

V. SUGGESTION

As the main concern of our project is on the prediction of the paddy productions. The crop production can be enhanced by efficient use of the ground water level. There are a few suggestions such as check dams, reservoirs and water management system s to improve the ground water.

VI. ACKNOWLEDGEMENT

The authors thank to the Deputy Director,Ground Water Department, Vijayawada for providing the opportunity for sharing information and receiving useful comments on Ground Water.

VII.REFERENCES

[1] V.K.Somvanshi, O.P.Pandey, P.K.Agrawal, N.V.Kalanker, M.Ravi Prakash and Ramesh Chand, Modelling and prediction of rainfall using artificial neural network and ARIMA techniques, J. Ind. Geophys. Union (April 2006) Vol.10, No.2, pp.141-151.

[2] Sanjeev Kumar, jay Indian, Zubair Khan, Neural Network Model for Prediction of ground water Level in Metropolitan considering Rainfall Runoff as a Parameter, International Journal of Soft Computing and Engineering(IJSCE) ISSN:2231-2307,Volume-3,Issue-3,July 2013

[3] Temporal Models for Groundwater Level Prediction in Regions of Maharashtra by Lalit Kumar, Department of Computer Science and Engineering, Indian Institute of Technology Bombay, June 2012

[4] Predicting groundwater levels using linear regression and neural networks by Sara Maatta

[5] Purna C.Nayak,Y.R.Satyajitrao , K. P. Sudheer, Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach, Water Resources Management (2006) 20: 77–90.

[6] D.Ratnam, Dr J Rajendra Prasad, Dr B V SubbaRao, Empirical Method Technique to Make Short Term Forecast of Rainfall for a Specific Region, International Journal of Advanced Research in Computer Science, Volume 5, No. 2, March 2014 (Special Issue)

[7] .Nabil Bessaih, Mohsin Qureshi, Fatima Salem Al-Jabri, Iman Rashid Al-Harmali and Zahra Ali Al Naamani, Groundwater Water Level Prediction in Wadi El Jezzy Catchment Using ANN, Proceedings of the World Congress on Engineering 2014 Vol I, WCE 2014, July 2 - 4, 2014, London, U.K.

[8] Data mining techniques for land use catchment planning by Neil Dunstan Gardiner, J., Thomson, k., Newson, M., Integrated watershed/river catchment planning and management: a comparison of selected Canadian and United Kingdom experiences, Journal of Environmental Planning and Management,37:1:53-67(1994).