Indian Journal of Science and Technology, *Vol 9(38), 10.17485/ijst/2016/v9i38/91714, October 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Prediction of Crop Yield using Regression Analysis

V. Sellam* **and E. Poovammal**

Department of Computer Science and Engineering, SRM University, Kattankulathur – 603203, Tamil Nadu, India;
sellamveera@gmail.com, poovammals@gmail.com

Abstract

Yield prediction benefits the farmers in reducing their losses and to get best prices for their crops. The objective of this work is to analyze the environmental parameters like Area under Cultivation (AUC), Annual Rainfall (AR) and Food Price Index (FPI) that influences the yield of crop and to establish a relationship among these parameters. In this research, Regression Analysis (RA) is used to analyze the environmental factors and their infliction on crop yield. RA is a multivariate analysis technique which analyzes the factors groups them into explanatory and response variables and helps to obtain a decision. A sample of environmental factors like AR, AUC, FPI are considered for a period of 10 years from 1990-2000. Linear Regression (LR) is used to establish relationship between explanatory variables (AR, AUC, FPI) and the crop yield as response variable. R^2 value clearly shows that yield is mainly dependent on AR. AUC and FPI are the other two factors influencing the crop yield. This research can be extended by considering other factors like Minimum Support Price (MSP), Cost Price Index (CPI), Wholesale Price Index (WPI) etc. and their relationship with crop yield.

Keywords: Regression Analysis, Yield of the Crop

1. Introduction

Agriculture has always been one of the vital occupations that serve mankind, both in terms of livelihood and employment. Due to the substantial increase in the population, the nutritional status of the poor is growing bad, which must be improved. The major effect of population increase has been prominently shown on the environment, the damage of which is increasing rapidly, which ultimately hinders agricultural production. Studies show that the modern techniques used in agriculture have not been environment-friendly, though they are technologically advanced than the primitive techniques.

The past achievements in the field of agriculture clearly depict the power and ability of man being able to meet the agricultural demand in spite of the population growth. Hence, a balanced relationship between the nature's major creations i.e. human beings and their environment has to be maintained in order to lead a sustainable life.

Table 1 clearly presents the contribution of agriculture to the national income and its share in export for a period of 50 years. An examination of Table 1 makes clear that the share of agriculture in the national income and in the total export is declining consistently. Now agriculture contributes only about one-third to the national income as against 54% in 1950-51. Similarly, the share of agricultural goods in export has declined from 52.5% in 1950-51 to only 16.5% in 1990-91. To increase the agricultural contribution to the national income; the production of crops should be increased.

Yield prediction is one of the most critical issues faced in the agricultural sector. Farmer's lack of knowledge about harvest glut, uncertainties in the weather conditions and seasonal rainfall policies, depletion of nutrition level of soils, fertilizer availability and cost, pest control, post–harvest loss and other factors leads to decrease in the production of the crops.

Regression Analysis can be defined as a structured approach which stresses on the analysis of data for the research purpose on decision making and problem solving. There are problems/situations that require simultaneous analysis of multiple variables or objects for efficient decision making. We consider various factors like Area under Cultivation (AUC), Annual Rainfall (AR) and Food Price Index (FPI) that contributes to the

*Author for correspondence

Table 1. India: Position of agriculture in national income and total export (1950-91)

Year	Contribution of agriculture to national income	Share of agriculture to total exports of India
1950-51	54	52.5
1960-61	49	44.0
1970-71	47	37.5
1980-81	36	25.5
1990-91	31	16.5

Source: Food and Agricultural Organization (2000: Press Note)

yield of crop. In this work, Regression Analysis is used to establish the relationship among these 3 factors and to identify their influence on crop yield.

Regression Analysis is a commonly used technique in the research where relationship among the three considered variables (AUC, AR, FPI) has to be established and to identify their effects on crop yield. Crop yield is considered as a dependent variable and AUC, AR, FPI are considered as independent variables. Regression Analysis is used to find the relative strength between a dependent variable and an independent variable i.e. impact of AUC on Yield, AR on yield and FPI on yield. The crop considered for analysis is rice because it is the most common crop cultivated in many areas of India.

2. Agriculture in India

The history of agriculture in India was documented in 1100 BC during the Rig Veda

period. [ref: https://en.wikipedia.org/wiki/Agriculture_in_India] [1]Discusses about pattern classification techniques like k-means, SVM etc., for various applications in agriculture and involvement of total workforce in agriculture. [2]Discusses about identifying potential cropping zones using Relative Spread Index (RSI) and Relative Yield Index (RYI) for increasing the yield of a crop.

Recent studies show that India's agricultural production can be increased by improving the grain storage infrastructure and farm productivity, not only to feed its growing population but also export them globally.

During the monsoon season of the year 2011, Indian agriculture achieved an all-time record of producing 85.9 million tons of wheat, an increase of 6.4% from the previous year.

During the same period, rice production showed an increase of 7%, hitting a new record of producing 95.3 million tons.

India exported $39 billion worth of agricultural products in 2013, making it the seventh largest agricultural exporter worldwide and the sixth largest net exporter. According the survey conducted in 2013, India stands second in farm output. Also, Agriculture, forestry and fisheries made up 13.7% of the total GDP (Gross Domestic product). A brief survey of various data mining techniques and its applications on agriculture was also summarized in[5].

Because of the broad based and fast economic growth in India, agriculture's contribution to the total GDP is declining. But, agricultural sector plays a significant role in the socio-economic growth of India and is still the broadest economic sector as in[6,7]. India exported $39 billion worth of agricultural products in 2013, making it the seventh largest agricultural exporter worldwide and the sixth largest net exporter.

India is now one among the world's largest suppliers of rice, wheat, cotton and sugar. It has exported over 2 million metric tons of wheat and 2.1 million metric tons of rice to most of the Asian and African countries. In the production of dry fruits, agro-based textile raw materials, roots and tuber crops, pulses, farmed fish, egg, coconut, sugarcane and numerous vegetables, India stands second or third in the world. In 2010, India stood one among the top 5 world's largest producers of over 80% agricultural products like coffee and cotton.

3. Regression Analysis

Regression analysis is used to analyze and determine the relationship between response variable and explanatory variable. The variables considered for analysis in this research work are Annual Rainfall (AR), Area under Cultivation (AUC), Food Price Index (FPI). Crop yield is a dependent variable which depends on all these ecological factors.

4. Linear Regression

As narrated in[10], Linear regression is discussed as a technique that is used to analyze a response variable Y which changes with the value of the intervention variable X. An approach of predicting the value of a response variable from a given value of the explanatory variable is

also referred to as prediction. The least-square fit, which is capable of fitting both linear as well as polynomial relationships, is the most commonly used linear regression. The approach of applying model estimate to values outside the domain of the original data is known as extrapolation. A linear regression model is computed to analyze the relationship between AR, AUC, FPI and Yield.

Stage 1: Compute a Linear Regression Model

A linear regression model is computed that evaluates the relationship among the variables. The conditions associated with model include, 1. Linearity, 2. Nearly normal residuals and 3. Constant variability.

Linearity explains that there should be a linear relationship between the response variable and the explanatory variable of the model, since we are using a linear model for the prediction. Nearly normal residuals states that the residual should be distributed centered around 0. There occur many instances during which there may be unusual observations that do not follow the general trend of the data. This condition can be easily checked for using a histogram or a normal probability plot of the residuals. If the histogram happens to be symmetric then we can interpret that the residuals are normally distributed. In case of the plot of residuals, if the plots are found closer to the normality, then the condition of symmetry is satisfied.

Stage 2: Compute the Residual Values

Residuals can be basically defined as leftovers from the computed model fit. The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Each data point has one residual.

Residual = Observed value – Predicted value $e = y - \hat{y}$

Stage 3: Compute the Residual Sum of Squares and Obtain the R^2

The residual sum of squares is defined as the sum of the squares of deviation of the predicted from the observed values of the data. In other words it can be referred as discrepancy between the data and an estimation model. The square of the correlation coefficient is calculated as R^2. The strength of any linear model is generally evaluated using R^2. The value of R^2 tells us the percentage of variability in the response variable. The value of the R^2 is always between 0 and 1 that corresponds to the variability of the response variable that is explained by the model. The R^2 value is calculated, considering one variable as the response variable and another as the explanatory variable, thereby establishing a successive linear relationship between the variables.

Stage 4: Implementation

The data in the Table 2 is used to establish a relationship among the variables AUC and Yield, AR and Yield and FPI and Yield. A decade's data is used for analyzing the relationship and its effects on yield.

So AUC, AR, FPI is considered as explanatory Variables and Yield of the Crop is considered as response variable. The process of linear regression is applied three times for rice crop to establish a relationship between the three variables (AR, AUC, FPI and Yield).

Table 2. Rice production in India

Year	Annual Rainfall (AR)	Area(Million Hectare)	Food Price Index (FPI)	Production (Million Tons)	Yield (Kg./ Hectare)	Area Under Irrigation(%)
2000-01	1120.2	44.71	92.4	84.98	1900.7	53.6
2001-02	981.4	44.9	101.0	93.34	2079	53.2
2002-03	1278	41.18	96.2	71.82	1744	50.2
2003-04	1085.9	42.59	98.1	88.53	2077	52.6
2004-05	1185.4	41.91	105.0	83.13	1984	54.7
2005-06	1133	43.66	106.8	91.79	2102	56
2006-07	1180.2	43.81	112.7	93.36	2131	56.7
2007-08	1075	43.91	134.6	96.69	2202	56.9
2008-09	972.8	45.54	155.7	99.18	2178	NA
2009-10	1212.3	41.85	132.8	89.13	2129.7	NA
2010-11	1213	36.95	150.7	80.41	2177	NA

Source: Department of Statistics and Agriculture, National Informatic Centre

MATLAB environment is used to implement linear regression for the data analyzed over a period of 10 years. If the value of 'rsq (R^2)' obtained is greater than 0.5 then the relation between the response variable and the explanatory variable is quite high.

The prediction of Yield from AUC is quite high because the value of 'rsq' is 0.7242 so it is clearly inferred that as the cultivation area increases the yield of crop increases as in Figure 1. In Figure 2, the results of R^2 clearly indicates that the crop's yield is highly dependent on the Annual Rainfall (AR). Similarly it is found that yield plays a good role as a response variable for the explanatory variable FPI as in Figure 3. Thus yield of a crop is dependent on three important factors like AUC, AR and FPI.

Thus from all of the individual linear relationships obtained, we can construct a relationship from AR to FPI by combining the individual relations obtained. A relationship is obtained by using Regression Analysis from AR to FPI and thereby to yield is given in Figure 4.

Thus the obtained relation in Figure 4 shows that AR influences AUC, AUC in turn influences Food Price Index and FPI of a specific crop subsequently impacts the yield of crop.

```
>> p=polyfit(var1,var2,1);
>> yfit=polyval(p,var1);
>> yresid=var2-yfit;;
>> SSresid=sum(yresid.^2);
>> SStotal = (length(var2)-1) * var(var2);
>> rsq = 1 - SSresid/SStotal

rsq =

    0.7232
```

Figure 1. Yield prediction from AUC for rice

```
>> p=polyfit(var1,var2,1);
>> yfit=polyval(p,var1);
>> yresid=var2-yfit;;
>> SSresid=sum(yresid.^2);
>> SStotal = (length(var2)-1) * var(var2);
>> rsq = 1 - SSresid/SStotal

rsq =

    0.7242
```

Figure 2. Yield prediction from AR for rice.

```
>> p=polyfit(var1,var2,1);
>> yfit=polyval(p,var1);
>> yresid=var2-yfit;;
>> SSresid=sum(yresid.^2);
>> SStotal = (length(var2)-1) * var(var2);
>> rsq = 1 - SSresid/SStotal

rsq =

    0.6636
```

Figure 3.: Yield prediction from FPI for rice

Figure 4. Relationship between MSP to FPI.

5. Conclusion

As shown in Table 2 there is a steady growth or slight difference in Annual Rainfall, Area under Cultivation and Food Price Index which clearly has an effect on production of the crops. The influenced value $R^2 = 0.7$ is obtained by implementing the Regression Analysis for the data in Table 2. This R^2 value clearly states that AR, AUC and FPI have an average of 70% influence in the crop yield.

Thus in this work, Regression Analysis is used to establish a relationship among a set of variables AR, AUC and FPI and their effects on yield of rice crop. This work can be extended by considering more factors like Minimum Support Price (MSP), Weather Conditions, Soil Parameters etc. that affects the yield of a crop and by using various data mining, statistical techniques to analyse the factors influencing the yield.

7. References

1. Mucherino A, Papajorgji P, Pardalos PM. A survey of data mining techniques applied to agriculture. Springer–Verlag; 2009 Jun.
2. Kokilavani S, Geethalakshmi V. Identification of efficient cropping zone for rice, maize and groundnut in Tamil Nadu. Indian Journal of Science and Technology. 2013 Oct; 6(10).
3. Chinchuluun A, Xanthopoulos P. Data mining techniques in agricultural and environmental sciences. 26 International Journal of Agricultural and Environmental Information Systems; 2010 Jan-Jun; 1(1):26–40.
4. Suraparaju V, Misra B, Singh CD. Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics (ICCCI–2014); Coimbatore, India. 2014 Jan 3–5.

5. Patel H, Patel D. A brief survey of data mining techniques applied to agricultural data. International Journal of Computer Applications. 2014 Jun; 95(9):6–8.
6. Kumar DA, Kannathasan N. A survey on data mining and pattern recognition techniques for soil data mining. IJCSI International Journal of Computer Science Issues. 2011 May; 8(3):422–8.
7. Kalpana, Shanthi, Arumugam. A survey on data mining techniques in agriculture. International Journal of Advances in Computer Science and Technology. 2014 Aug; 3(8). 2320–2602.
8. Mankar AB, Burange MS. Data mining - An evolutionary view of agriculture. International Journal of Application or Innovation in Engineering and Management.2014 Mar; 3(3):102–5.
9. Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering.2013 Sep; 2(9):3477–80.
10. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate data analysis. 6th ed. Pearson Education Inc; 2006.