IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)
e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 8, Issue 10 Ver. II (Oct. 2015), PP 81-88
www.iosrjournals.org

A Design Of A Data Warehouse And Use Of Data Mining Techniques For Analysis Of Risk Factors Affecting Agriculture In India

Nirav Desai, GeetikaKalra, Amit Khandelwal, Shashank Mishra, TejaswiniApte
Symbiosis Institute of Computer Studies and Research, Pune

Abstract:*In this paper, data made available in public domain by the Government of India, under the Open Government Data Initiative is analyzed using data warehouse design and data analysis techniques. A data model is developed to analyze risks in agriculture. Regression analysis and K-Means Clustering are used as analysis techniques to derive insights from the available data and identify potential risk factors.*

Keywords:*data warehouse, data analysis, regression analysis, K means clustering, agriculture in India, irrigation, rainfall, crop yields, Online Analytical Processing, data mining.*

I. Introduction

In 2014, there were 12,360 farmer suicides reported in India. That is slightly more than the number of farmer suicides reported in the previous years. Nearly half of these suicides were in Maharashtra and Telangana alone. Bankruptcy is the main reason driving these farmer suicides.The government recently unveiled a plan to enroll farmers in a comprehensive agriculture and health insurance scheme known as Unified Package Insurance Scheme (BharatiyaKrishiBimaYojana)[1]. It will combine nine features with a mandatory crop insurance[2]. In this paper, we attempt to analyze various factors affecting agricultural yield in India and develop low risk alternatives for high risk farming practices.

Specifically, in this paper we will attempt to develop a data warehouse with a star schema dimensional model for the agriculture industry in India. We will next identify the data sources that give this information and load the data warehouse. We will develop OLAP (Online Analytical Processing) Cubes for the analysis of the data loaded into the data warehouse. Regression analysis and K-Means Clustering will be used to derive insights from the available data.

Dimensional Modelling

Dimensional Modelling approach requires identification of different factors that affect agriculture such as rain, irrigation, fertilizers, pesticides and crop warehouses and store specific value for yield for each combination of these factors. A time dimension is also normally introduced for number of measurements taken over a period of time. A star schema based dimensional model for agriculture industry is shown below:

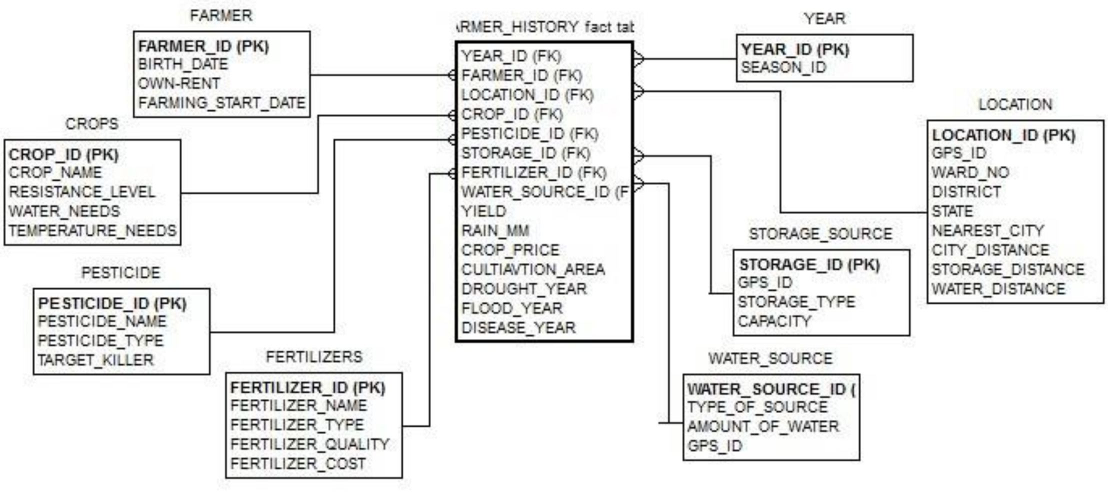


Figure 1: Dimensional Model for the agriculture industry

The resistance of crops, availability of water, climate (based on location), use of pesticides and fertilizers and availability of storage warehouses are factors identified as affecting the risk of agriculture in India. External factors such as annual rainfall and impact of floods and droughts are included in the fact table. Scientific farming practices such as crop rotation affect agricultural yields and could be modelled using the farmer experience model parameter.The farmer history fact table contains measured values for rainfall, yield and total crop area for each of these combinations of dimension parameters.

Data Sources:

Historical data on agricultural yield for different crops in India at a district level resolution for each season of the year has been made available by the Government of India on the Open Government Data platform data.gov.in. Data on precipitation and irrigation in these districts over the years has also been made publicly available by the Indian Metrology Department on their own website as well as the Open Government Data platform.Data on irrigation is available for the year 2001-2002 for all districts in India. Irrigation data includes surface water and ground water components. For analysis of the total sum insured by the crop insurance policy, Government of India has released the Minimum Support Price for all major crops in India from the years 2001 to 2010. Based on the available data, a reduced dimensional model is prepared as shown below:

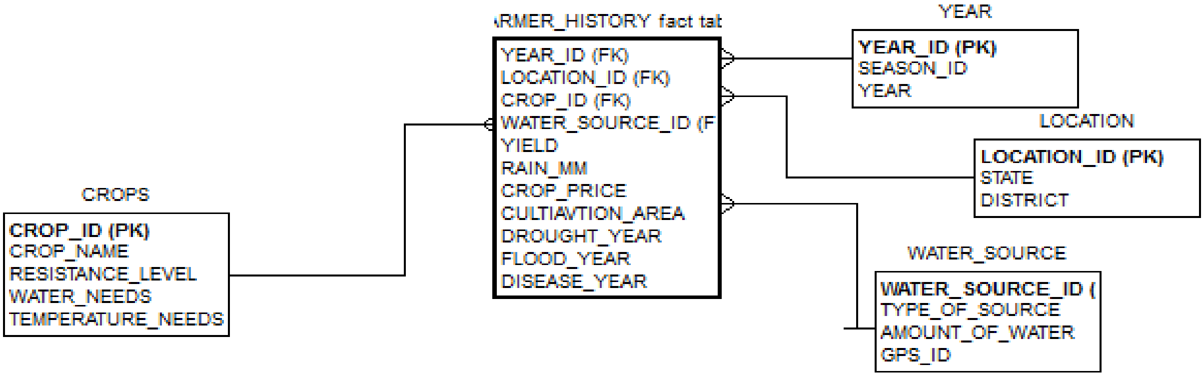


Figure 2: A reduced dimensional model for the agriculture industry in India which can incorporate data made available on data.gov.in

Data Integration

The data made available in Excel spreadsheets and .csv files from data.gov.in is downloaded and integrated using the CLOVER ETL tool so as to match all the key fields on crop names, district names and year values. The workflow for data integration using CLOVER ETL is shown in the figure below.

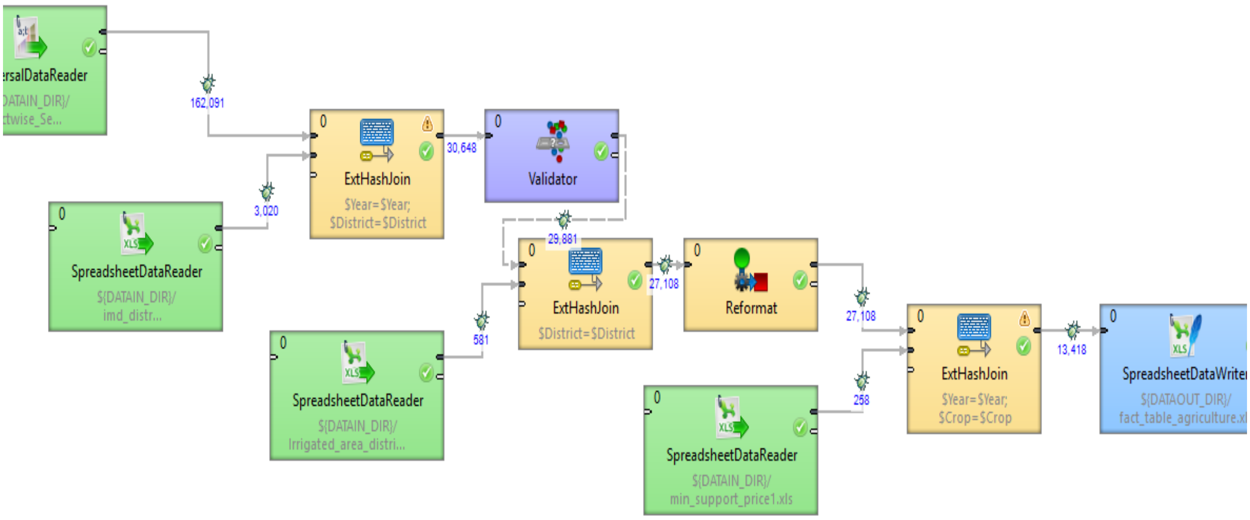


Figure 3: Data integration work flow in CLOVER ETL, used for integrating data sets available from different sources.

There are 162,091 records on crop yield history for years from 1951 to 2010. Normalized rainfall data is available only from 2004 to 2010. Hence, the match produces only 30,648 records. Irrigation data is not available for all the districts and the match reduces the number of records to 27,138. This data is sufficient for risk analysis. For total sum insured analysis, we do a match with the MSP prices from 2004 to 2010. Since MSP prices for all crops are not available, we are left with 13,418 records for this. The end result of this work flow is a fact table with measure values and referential integrity with dimension table values that can be loaded into Microsoft SQL Server 2012[3].

The OLAP cubes are developed on this tool and the structured data in the dimensional model and OLAP cubes is imported into an Excel spreadsheet for analysis[4]. The analysis carried out using Microsoft Excel OLAP plugin with pivot charts and Orange Canvas data mining tool is presented next.

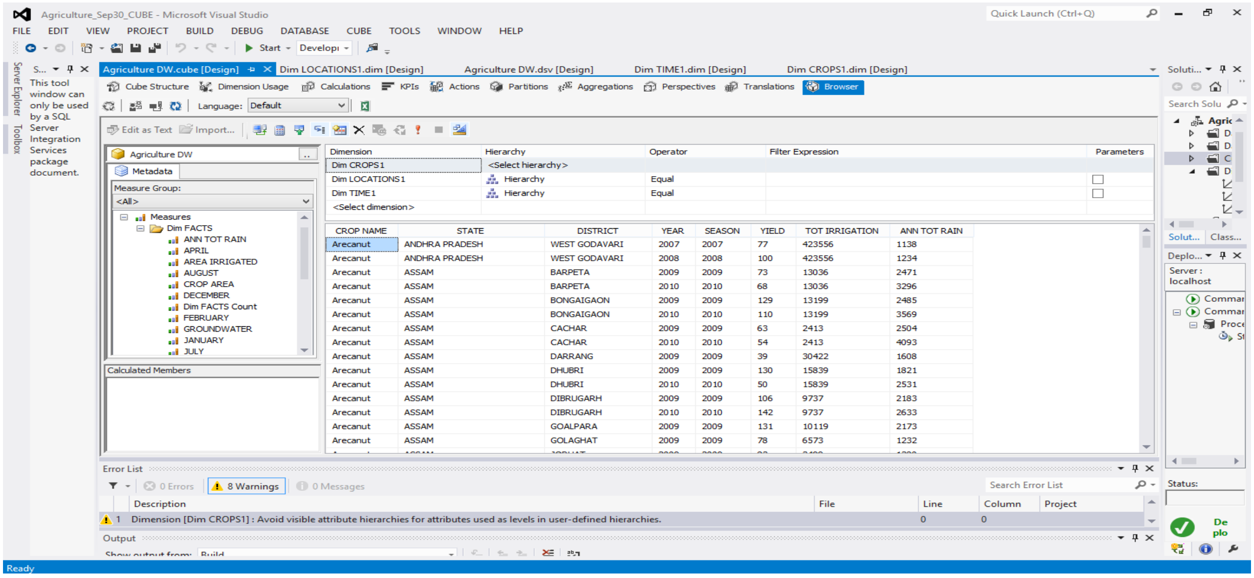
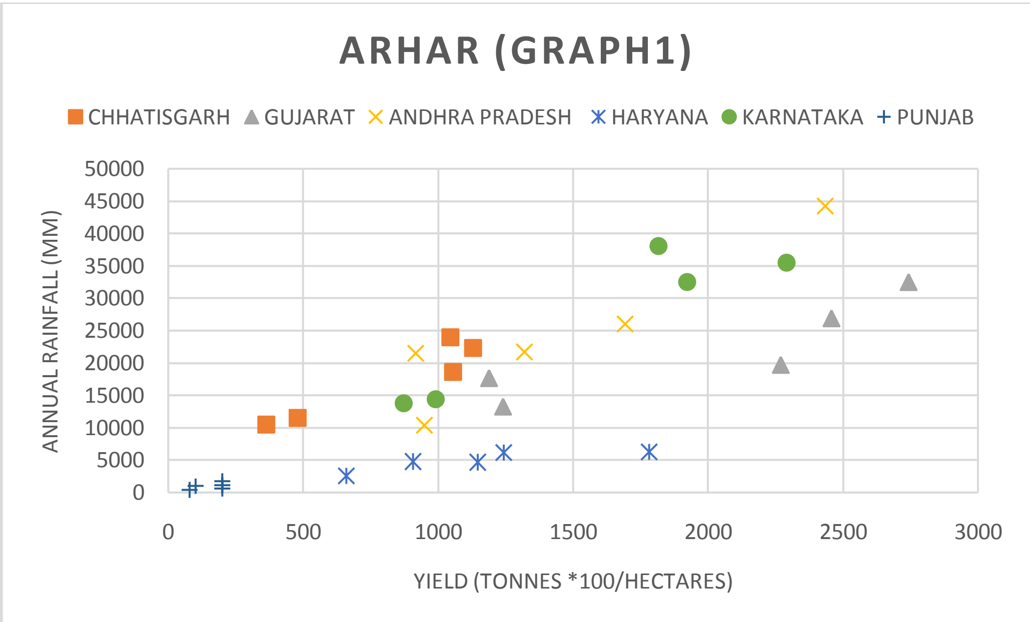
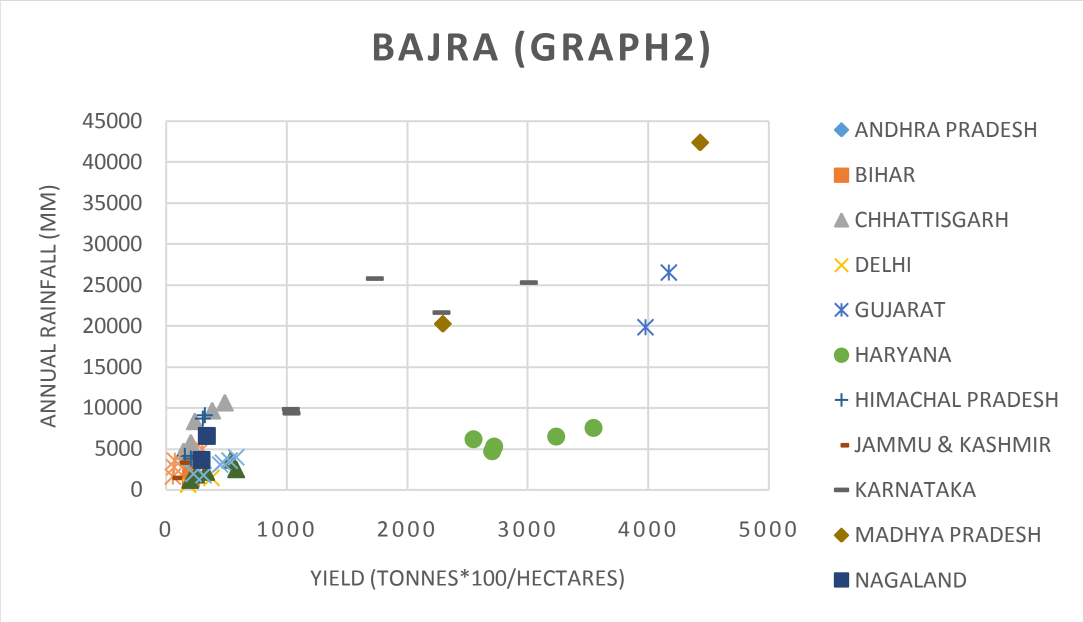


Figure 4: Screenshot of OLAP cubes developed using Microsoft SQL SERVER 2012 for the agriculture data warehouse.

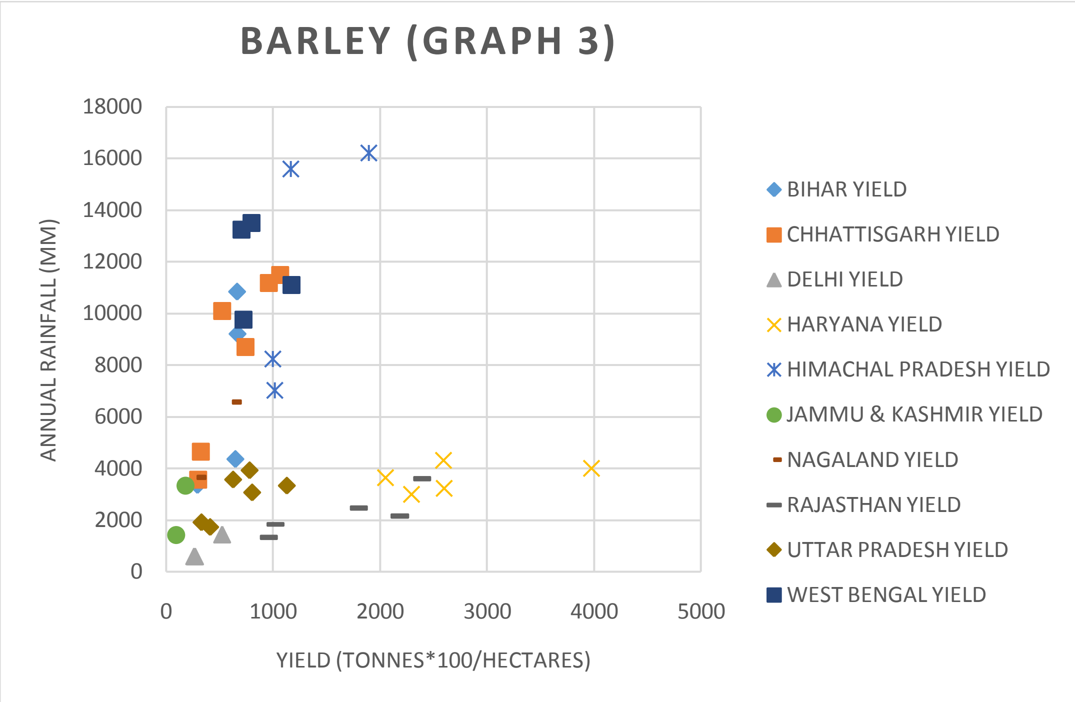
Data Analysis:

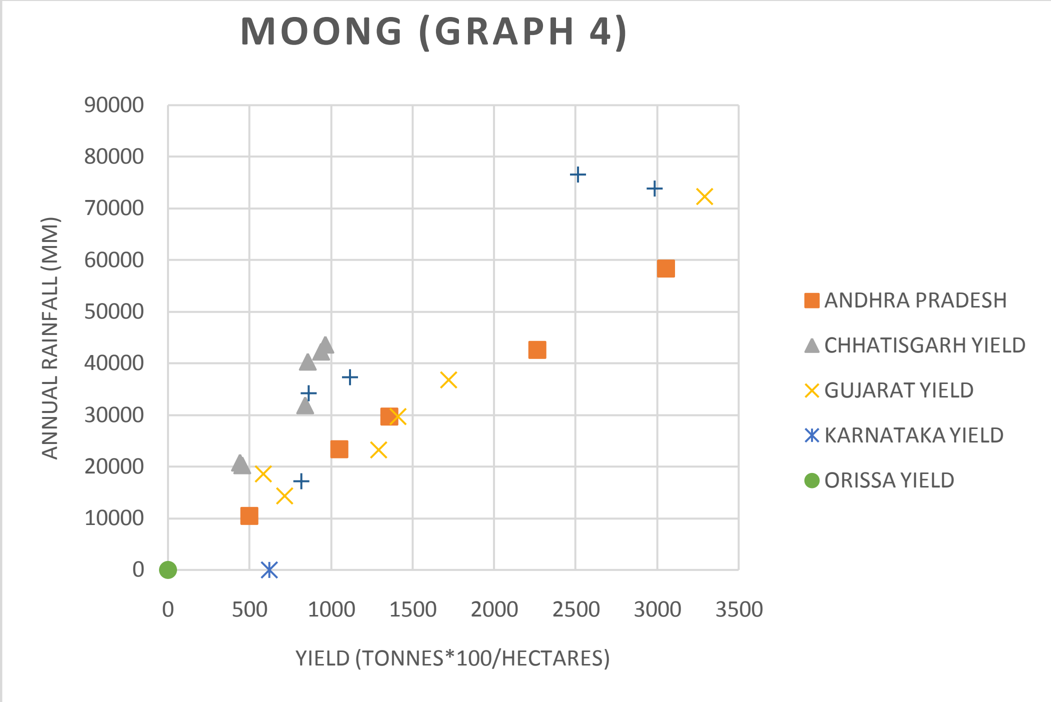
The data mining techniques of regression analysis and K-Means clustering are used to develop insights into the available data.For regression analysis, we filtered the data by crop name and created a scatter plot of agricultural yield vs rainfall for various states in India for the years 2004 to 2010. These plots are presented here below. The graph title shows the crop name and legend has been included to demarcate the yield in different states.



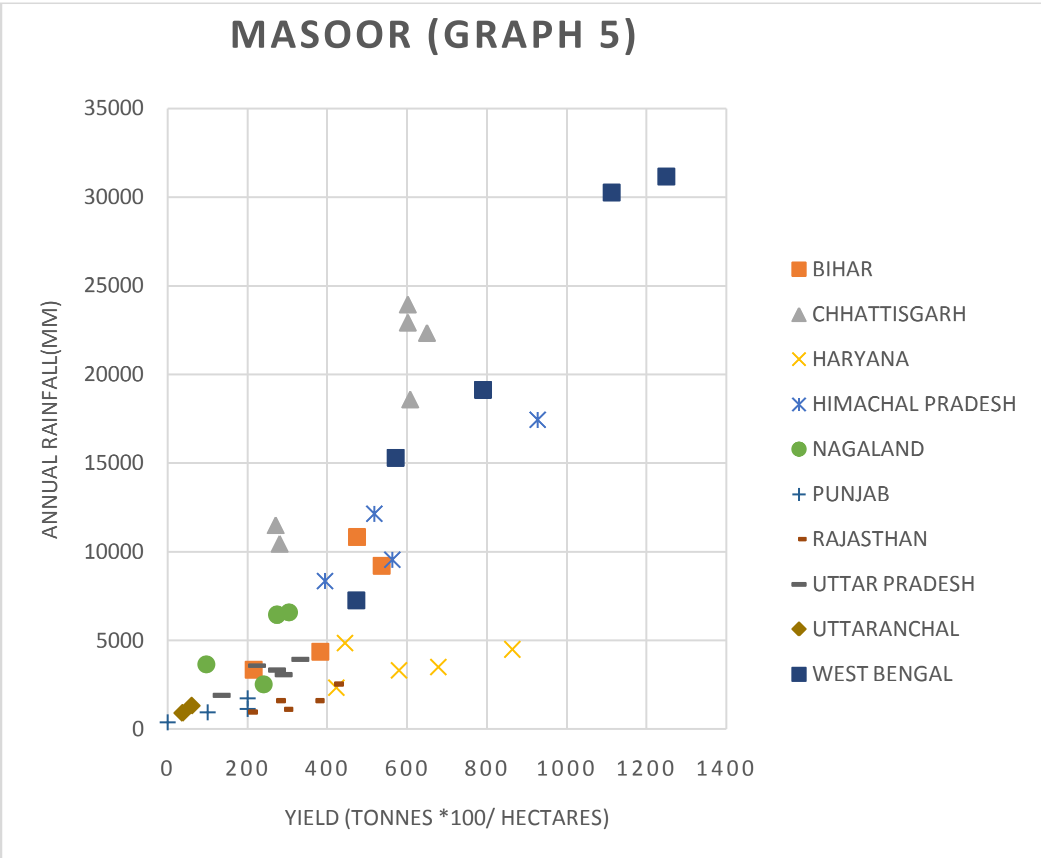


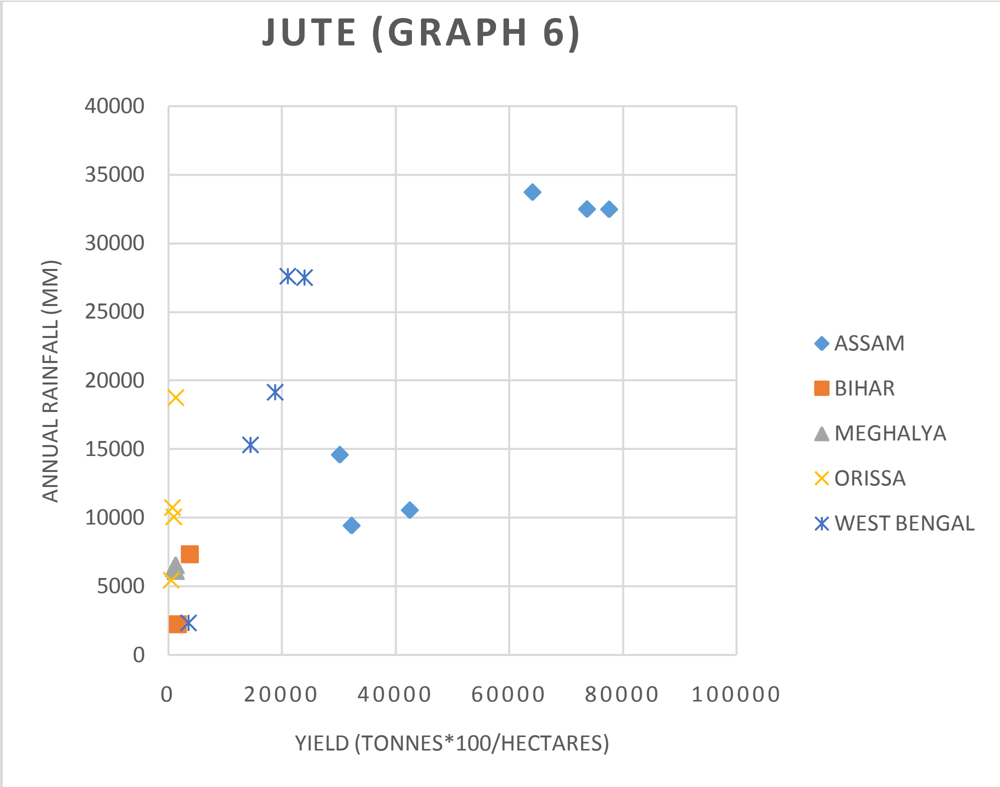
Graphs 1 and 2 depict the scatter plot of annual rainfall (mm) with agricultural yield (tonnes*100/hectares) measured over the years from 2004 to 2010 for arhar and bajra. The state names have been demarcated in the legend.





Graphs 3 and 4 depict the scatter plot of annual rainfall (mm) with agricultural yield (tonnes*100/hectares) measured over the years from 2004 to 2010 for barley and moong. The state names have been demarcated in the legend.



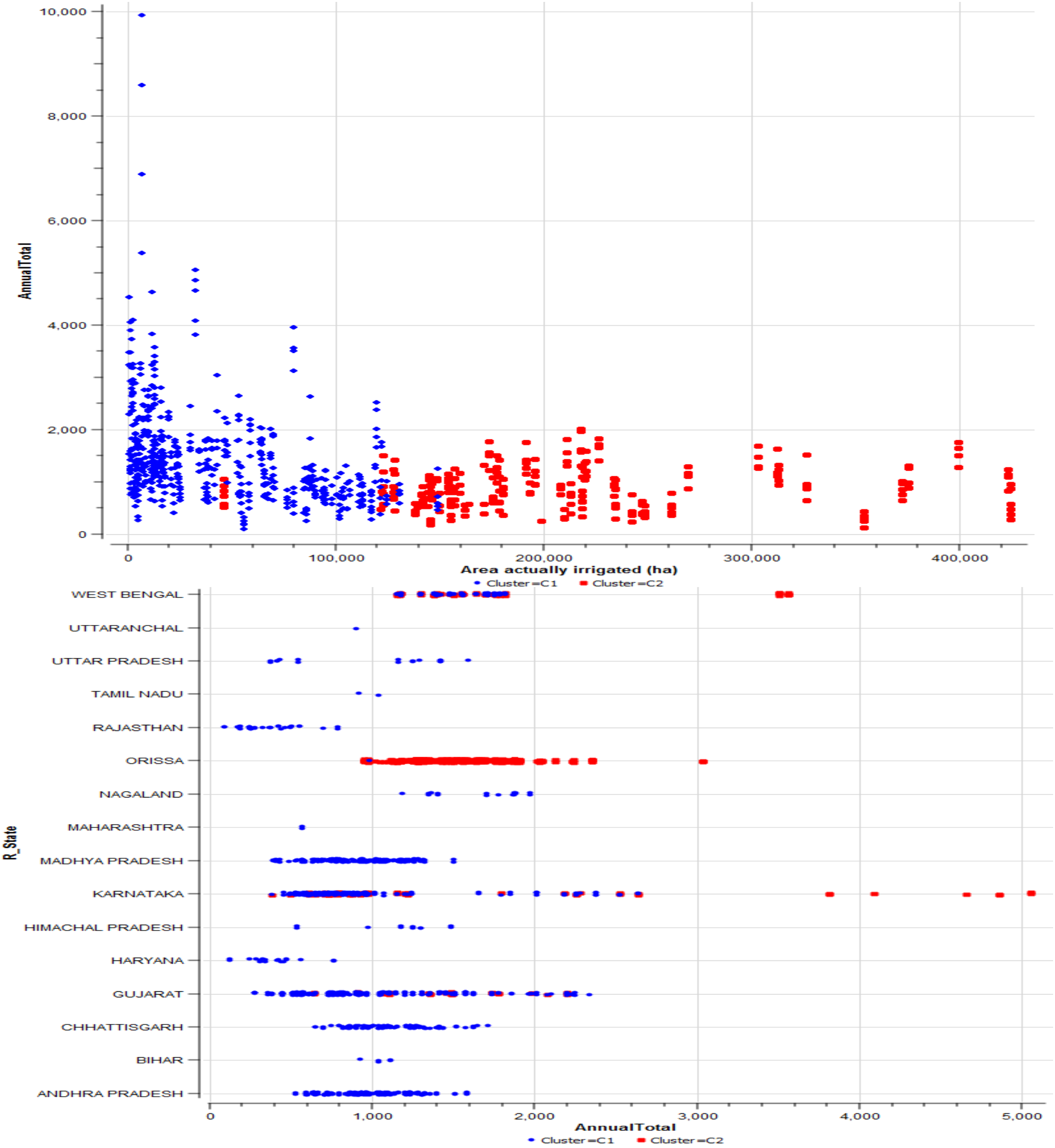


Graphs 5 and 6 depict the scatter plot of annual rainfall (mm) with agricultural yield (tonnes*100/hectares) measured over the years from 2004 to 2010 for masoor and jute. The state names have been demarcated in the legend.

The correlation in the yield and rainfall for arhar, bajra, barley, moong, masoor and jute depicted in the above graphs was calculated and tabulated in the table below.

CORRELATION OF YIELD WITH RAIN (2004-2010)	CROP					
STATE	ARHAR	BAJRA	BARLEY	JUTE	MOONG	MASOOR
ANDHRA PRADESH	0.930344449	0.917828			0.996063	
ASSAM	0.943522664			0.926094		0.958137
BIHAR		0.951714	0.687106	1		0.854112
CHHATISGARH	0.947285871	0.900424	0.889228		0.960851	0.944818
DELHI		0.999996	0.999451			
GUJARAT	0.889095396	0.822395			0.930552	
HARYANA	0.851207702	0.795557	0.425503			0.366054
HIMACHAL PRADESH		0.956926	0.744297			0.940919
JAMMU AND KASHMIR		1	1			
KARNATAKA	0.93569999	0.824118			0.951336	
MADHYA PRADESH		0.860612	0.996447			
NAGALAND		1	1			0.600443
ORISSA				0.932783	0.951243	
PUNJAB	0.53382004					0.902463
RAJASTHAN		0.828491	0.841923			0.865187
UTTAR PRADESH		0.963205	0.718154			0.848112
WEST BENGAL		0.842755	-0.21878	0.980644	0.954118	0.97789

Table 1: Calculated correlation values between yield and annual rainfall for a few important crops in India measured over the years 2004 to 2010. High correlation values indicate heavy dependence on rainfall for agricultural yield in these states



The graph 5 (top) shows a scatter plot with k-means clustering algorithm [5] applied to wheat production in India. 2 optimized clusters have been identified. Higher irrigation levels lead to high yield clusters (red squares) as shown in the graph. The graph6 (bottom) shows ground nut clusters by state and annual rainfall. Certain mixed clusters in a state indicate a mixture of high and low yield farming practices. States like Orissa have high yield farming clusters for groundnut. Data collected from 2004 to 2010. Annual rainfall in mm and yield is in tonnes/hectares. Red clusters are high yield.

II. Conclusion

The analysis of the collected data reveal high correlations between rainfall and yield for most major crops and states in India (more than 85% on average). The heavy dependence on rainfall is a very high risk farming practice since the rain levels change every year and the yield cannot be fixed due to this. Highly irrigated clusters show significantly less dependence on rain and these are found in a small percentage of the total area under agriculture. More irrigation is clearly the means to risk reduction in majority of the farming clusters in India as per this data. The total area affected by droughts every year is around 6-7% of the total area under agriculture. The total liabilities for a crop insurance policy should also be somewhere in the range of 6-7% of the total sum insured.

References

[1]. Government set to expand the social security net with new launch of farm and health insurancehttp://www.policybazaar.com/health-insurance/general-info/news/government-to-expand-social-security-net-with-new-launch-of-farm-health-insurance/
[2]. Government set to unveil farm and health insurance for masses: http://indianexpress.com/article/business/business-others/govt-set-to-unveil-farm-health-insurance-for-masses/
[3]. Moving data from Excel to SQL SERVER, Andy Brown https://www.simple-talk.com/sql/ssis/moving-data-from-excel-to-sql-server---10-steps-to-follow/
[4]. Create a first data warehouse, Mubin Shaikh http://www.codeproject.com/Articles/652108/Create-First-Data-WareHouse
[5]. K-Means Clustering in Orange Canvas: Orange Canvas Manual http://user-manual.net/orange-canvas-tutorial-manual-pdf.html.